Retrieval of Visually Similar Images for Handwritten Documents Through Agglomerative Hierarchical Clustering

Mohamed Boukfal

Abstract– In tasks of handwritten words recognition from a collection of manuscripts, a possible approach consists on grouping images using a measure of similarity in order to get a cluster distribution, aiming to have all the same words in the same cluster. Keeping in mind this idea, agglomerative hierarchical clustering (AHC) techniques are used to implement retrieval by example methods by reducing the bag of word images in a few cluster representatives. Various linkage criteria, different distance metrics and representative obtainment methods are evaluated. A simple dataset is used to create and validate the algorithm and a subset of word images collection is used to get the final results. Modeling with the manuscript image words shows that AHC considerably reduces the amount of operation (decrease of request time) in offline handwriting recognition, giving the same (and even better) results than tradition approaches.

Keywords– agglomerative hierarchical clustering, dendrogram, handwritten documents, image retrieval, query by example, word spotting.

INTRODUCTION

WEALTH of Catalan historic archives is pretty known between historians, especially between medievalists. Composed by a great quantity of conserved documents is considered an inexhaustible source to study the established humans in Catalonian territory, and even to fill some archivistics sources of surrounding countries. It is not the aim of this study to list and describe the documents of these archives, but some examples of them are the 6000 record volumes of Chancery of the counts-kings preserved in the old Royal Archive of Barcelona (now General Archive of the Crown of Aragon), or, among other documents, a IX-X century collection of original documents consisting of 650 pieces at "Arxiu Capitular osonenc" [1].

Work with handwritten documents, specially the historical ones, is a though labor, particularly when trying to transcribe or index them. Transcribing manuscript documents brings many benefits to society, it makes the content accessible, more readable (in some cases original documents are so damaged that is hard to identify the text even for a human eye), it permits to make document analysis (search, indexing, content statistics, etc) and , finally, compressible (at least more than an image file) applying traditional text processes. Current researches in historical documents are using manual methods to transcribe or index the target documents, what it would take not little time to accomplish it for human workers, this method is inefficient and expensive.

Automatic methods are desired to perform this tasks: the identified problem is described as "offline handwriting recognition". Many methods are created to treat this need from different approaches: some algorithms are dedicated to optical character recognition (OCR), where images of handwritten texts are converted to machine encoded texts; other algorithms are intended to identify the distinct words in a collection and index them, that is what is called word spotting. Word spotting is a particular case of image retrieval [2](we will focus on those methods).

Manmatha et al. coined the term of wordspotting [3] for splitting handwritten documents into word images, and using image matching, get the distances between them in order to group similar words in the same cluster (each cluster for correct word transcription). Then clusters are labeled manually to index the documents collection. Taking this idea, is desired to apply it in non static collections of manuscripts, for instance, if a new handwritten document is found and the existing clustering is pretty good, it would be faster to transcribe the new image words if the cluster's

• Contact E-mail: mohamed.boukfal@gmail.com
• Studied specialisation: Computació.
• Supervised by: Marçal Rossinyol (Computer Science)
• Academic year: 2016/17

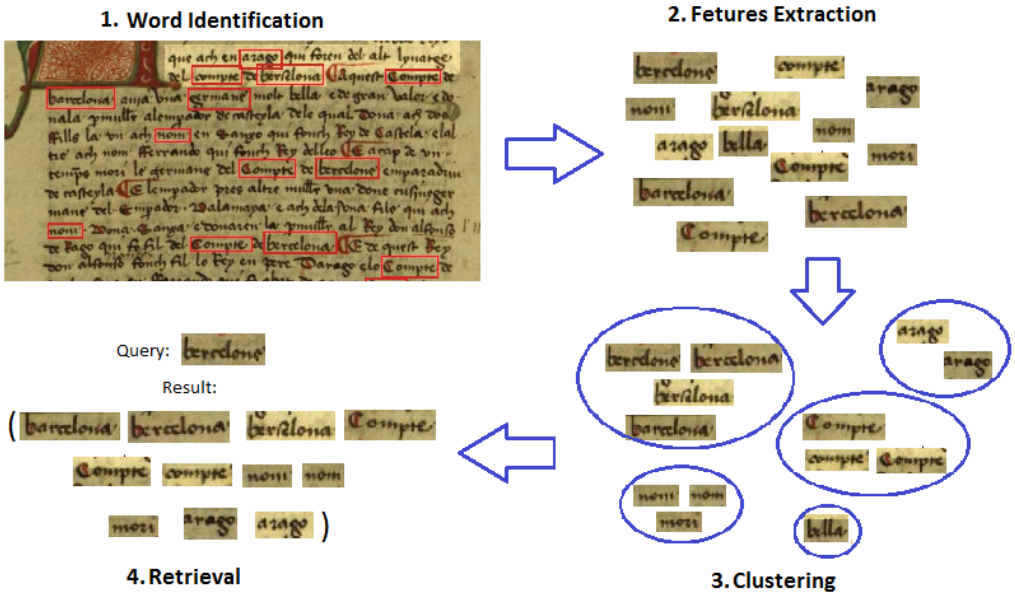

Fig. 1: Handwritten word images retrieval using AHC workflow.

label to which it belongs could be obtained. Or, what is the same, perform queries by example over the existing dataset.

In order to find have a good similarity measurement, it can be found many methods that given an image, or a collection of them, containing a handwritten word (or more than one), this image is treated to obtain a description of the words in it. In this particular case, full word description is chosen over character description due to the state in which can be found some historical manuscripts. Previous studies [4] have treated how to extract meaningful features from word images in order to make them identifiable (smaller distances between image instances of the same word, and larger for distinct words). In the current study we use this representations of handwritten word images as the input to our system.

Traditionally, to perform the query, distances (similarity criteria) between the example image and all dataset have to be computed and sorted, this is computationally expensive. The premise is that, given a clustered classification and extracting as many representatives as clusters (total of documents vocabulary), it's possible to reduce the computational cost by launching the query against those representatives and retrieving their represented samples ordered without decreasing the precision.

In order to keep track of the algorithm creation, and see if the hypothesis take sense, the experiment is divided in two parts, the first one with a small database where the premise is confirmed, and a second one, with handwritten word images dataset, from which are obtained the final results.

This paper is divided in 5 sections. Next section presents the followed method. In Section 3 the algorithm is introduced and the first results with the "control" dataset are exposed. Final database experiment and it's results are discussed in Section 4. And finally conclusions are presented.

METHODOLOGY

Given the particularity of this target, agglomerative hierarchical clustering (AHC) is considered [1][5], is a bottom-up approach where a dendrogram (tree representation of distances between the samples) is created. Cutting that dendrogram by a certain point a cluster split representation is obtained. Relation between clusters could be defined through different linkage methods and choosing between various metrics for distances, most common methods are evaluated and compared to get best image matching results.

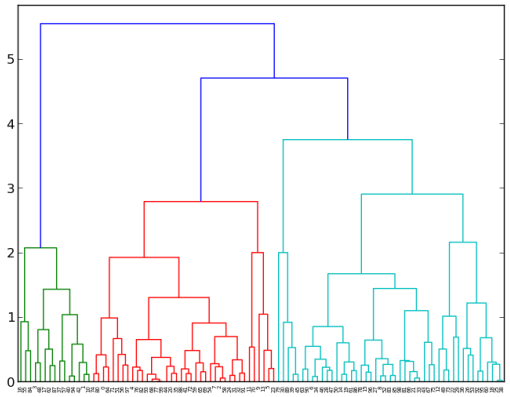

Fig. 2: Example of dendrogram.

Representative selection is done by taking in consideration three methods: computing the centroid through a median of each cluster, finding the geometrical median, and taking randomly a sample from each group (as a control measure). Then, an intracluster sort is performed around that representative. Methodology details are explained below:

AHC

Firs handwritten word images collections are treated to extract significant features [4], those are represented in vectorial form with the same length. So given a collection with n word images and m features, the matrix representation is described as it follows:

$$F = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{pmatrix}$$

where f_{ij} represents the feature j for the sample i.

A distance matrix D is computed using F, D notation is described as:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}$$

where d_{ij} represents the distance between sample i and sample j. Note that D is symmetrical ($d_{ij} = d_{ji} \forall i, j \leq n$) , $d_{ij} \geq 0, \forall i, j \leq n$ and $d_{ij} = 0, \forall i = j$.

Various distance metrics are considered:Euclidean, Cityblock (L1), Braycurtis, etc.

Using the distance matrix, the hierarchical agglomerative clustering is performed to get a 4 by $n-1$ matrix Z . At the$i$-th iteration, clusters with indices $Z[i,0]$ and $Z[i,1]$ are combined to form cluster $n+i$. A cluster with an index less than n corresponds to one of the n original observations. The distance between clusters $Z[i,0]$ and $Z[i,1]$ is given by $Z[i,2]$. The fourth value $Z[i,3]$ represents the number of original observations in the newly formed cluster.

The following linkage methods are used to compute the distance $d(s,t)$ between two clusters s and t. The algorithm begins with a forest of clusters that have yet to be used in the hierarchy being formed. When two clusters s and t from this forest are combined into a single cluster u, s and t are removed from the forest, and is added to the forest. When only one cluster remains in the forest, the algorithm stops, and this cluster becomes the root.

A distance matrix is maintained at each iteration. The $D[i,j]$ entry corresponds to the distance between cluster i and j in the original forest.

At each iteration, the algorithm must update the distance matrix to reflect the distance of the newly formed cluster u with the remaining clusters in the forest.

Various methods are taken in consideration for calculating the distance between the newly formed cluster u and each v.

- Single method assigns

  $$d(u,v) = min(dist(u[i], v[j]))$$

  for all points i in cluster u and j in cluster v. This is also known as the Nearest Point Algorithm.

- Complete method assigns

  $$d(u,v) = max(dist(u[i], v[j]))$$

  for all points i in cluster u and j in cluster v. This is also known by the Farthest Point Algorithm or Voor Hees Algorithm.

- Average method assigns

  $$d(u,v) = \sum_{ij} \frac{dist(u[i], v[j])}{(|u| * |v|)}$$

  for all points i and j where $|u|$ and $|v|$ are the cardinalities of clusters u and v, respectively. This is also called the UPGMA algorithm.

Representatives

Once Z is obtained (the dendrogram representation), a cut criteria is used to get the formed clusters. In this case, a max clusters (k) to be formed criteria is established. Cutting the dendrogram produces k clusters, each sample is labeled as the cluster number.

In mathematical notation we could consider $(l_1, l_2, ..., l_k)$ as the k cluster labels (or, directly, the k subsets).

Then for each label is computed a representative, three methods are taken:

- Mean representative :

  $$\overline{s_{l_i}} = \frac{\sum_j s_{l_i,j}}{|l_i|}$$

  for all points $s_{l_i,j}$ (points in cluster labeled as l_i) where $|l_i|$ is the cardinality of cluster l_i.

- Geometric Median representative:

  $$gm_{l_i} = argmin_{y \in l_i} \sum_j ||s_{l_i,j} - y||$$

  for all points $s_{l_i,j} \in l_i$. Note that gm_{l_i} is the point in l_i from where the sum of all distances to the s_{l_i} is minimum.

- Random representative:

  $$r_{l_i} = rand\{s_{l_i,j}\}$$

  a random point from l_i.

Using those representatives, a intra cluster re-sorting is done, to ensure that the most similar samples to the representative are retrieved first.

And finally, given a new sample p, k distances are computed (each for cluster), those are sorted and entirely ordered clusters are retrieved. Summing up, word images are ranked by similarity to a given example.

Input: **Matrix** F_{n*m} where each row is an array of m **features** of sample, metric, linkage method, max cluster numbers to be created and representative method selection.
Output: Array of representatives and resorted samples.

```
1. D = pairwise_distances(F, metric)
2. Z = linkage(D, method, metric)
3. clusters = fcluster(Z, num_clusters)
4. representatives = []
5. For cluster in clusters:
6.         representative = get_representative(
7.                 cluster, selection_method)
8.         representatives.add(representative)
9.         intra_sort(cluster, representative)
10.
```

Fig. 3: Main pseudocode algorith.

Accuracy measure

Mean average precision (mAP) is used to measure the accuracy of retrieval.For a set of queries, mAP is the mean of the average precision scores for each query:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

where Q is the number of queries and

$$AveP(q) = \frac{\sum_{k=1}^{n}(P(q_k) \times rel(q_k))}{|relevantdocuments|}$$

where $P(q_k)$ is the precision at cut-off k in the list and $rel(q_k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

Experiments and Results

Experiment I

This first retrieval by example experiment is evaluated in a modest and well known database: the iris dataset. Using a simple database permits to evaluate the behavior of the algorithm more reliably, and the execution is faster, which makes possible to execute it more times.

Iris dataset is usually used in studies of classification models, it's described as follows:

Classes	3
Samples per class	50
Total samples	150
Dimensionality	4
Features	real, positive

TABLE 1: Dataset characteristics.

Samples are randomly divided in two sets (procuring to have the same classes proportion in both sets) , the firs set with 120 samples (80As said in previous Section, multiple distance metrics (euclidean, correlation, canberra, ...) could be applied to get the similarity between two samples. In order to have the reference values of mAP [table 2] to evaluate the accuracy of the algorithm, mAP are calculated without clustering the training set. As it can be seen in table 2, the results vary according to the distance used. Given the complexity of the algorithm, it would be slow to test all of them in cluster creation, and considering that those results depend on the chosen dataset (may be considerably distinct from a database to another), metrics that give a $mAP \geq 0.7$ are taken to evaluate the system.

Metric	mAP
Canberra	0.9295
Cityblock(L1)	0.9068
Euclidean (L2)	0.9065
Braycurtis	0.9057
Cosine	0.9037
Chebyshev	0.8870
minkowski	0.8645
Sqeuclidean	0.8642
Seuclidean	0.7201

TABLE 2: MAP by metric without clustering.

Main algorithm [Fig. 3] is created following method described in Methodology Section. Many iterations have been executed to extract the mAP of all combination between distances, linkage methods, maximum number of cluster creation and representative choose criteria. Some results are exposed in figures 4, 5, 6.

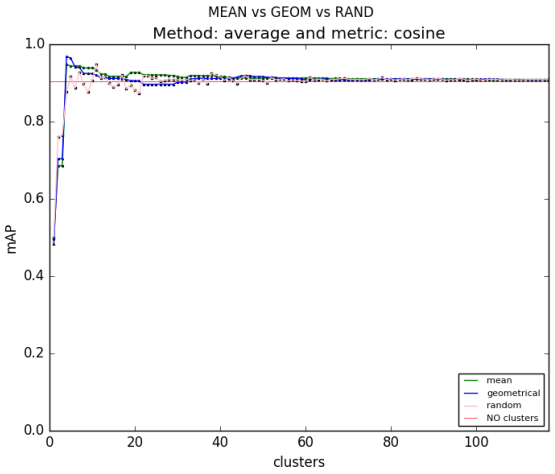

Fig. 4: Comparison between mean, geometrical and random representative selection methods.

As seen in in figures, with this dataset, encouraging results are obtained. MAP of the model grows rapidly as the number of clusters increases. Moreover, in most cases the obtained mAP is greater than expected results [table in Fig. 10]. Finally, as the number of clusters increases, mAP of retrievals converge to mAP calculated without clustering (as many clusters as train samples). For iris dataset, best results are obtained with a number of clusters near to number of classes with Cosine metric, other metrics need more clusters (max clusters ≥ 25).

The peak observed in the results could be explained by the use of AHC and representatives: using this method samples of the same cluster (and probably the same class) are

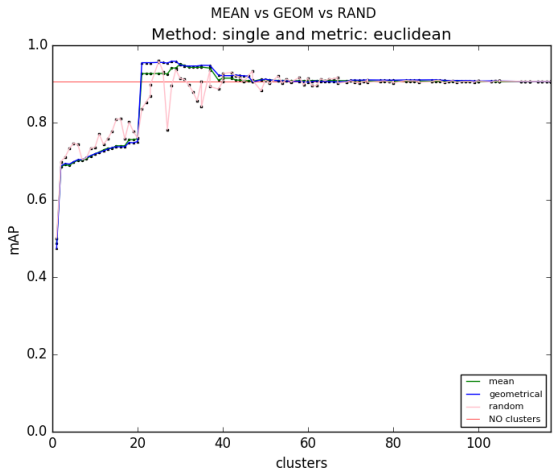

Fig. 5: Comparison between mean, geometrical and random representative selection methods.

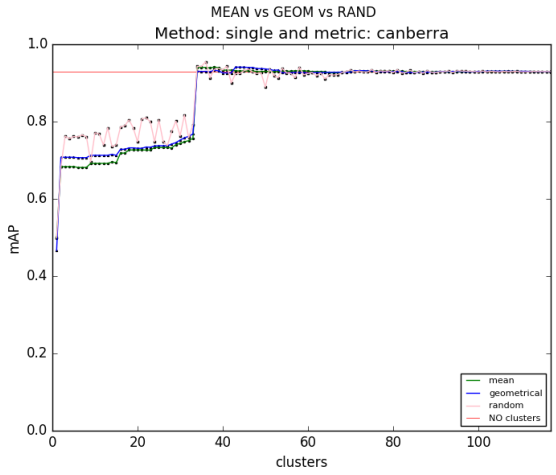

Fig. 6: Comparison between mean, geometrical and random representative selection methods.

ranked in earlier positions; against the method without clustering where samples are ordered regardless of the possible relationship between the data to be retrieved. For instance, if a query is placed in a frontier of a class, is more likely to get better results retrieving all the cluster than retrieving the closest samples.

Experiment II

Ones is confirmed the premise, same method is applied to a handwritten word images collection:

Total samples for train	3234
Total samples for validation	95
Dimensionality	9216
Features	reals
Language	English

TABLE 3: HANDWRITTEN IMAGES DATABASE DESCRIPTION.

In previous studies is seen that *braycurtis* metric gives the best results with a mAP near to 0.7460. For completion, and comparison purposes, the *euclidean* distance is considered (mAP = 0.6694).

With the growth of the new dataset is computationally more expensive to perform all combinations between the parameters of the algorithm, so, in order to limit them the best fitted parameters are taken:

- Distances: *euclidean* and *braycurtis*.
- Linkage methods: single, complete and average.
- Maximum number of clusters: from 1 to 3201 with step of 50.
- Representative choose criteria: Geometric median.

The results are exposed in graphics:

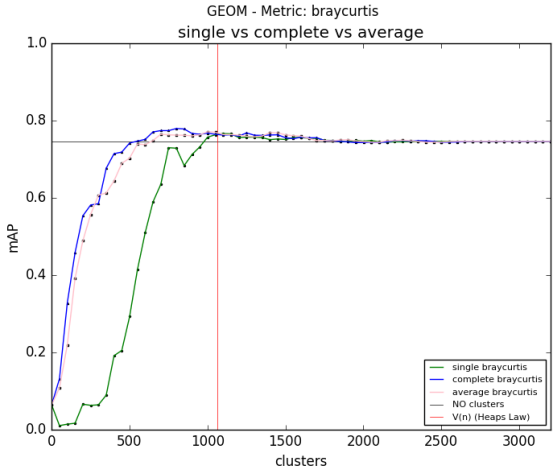

Fig. 7: Comparison between single, complete and average linkage methods for Braycurtis metric.

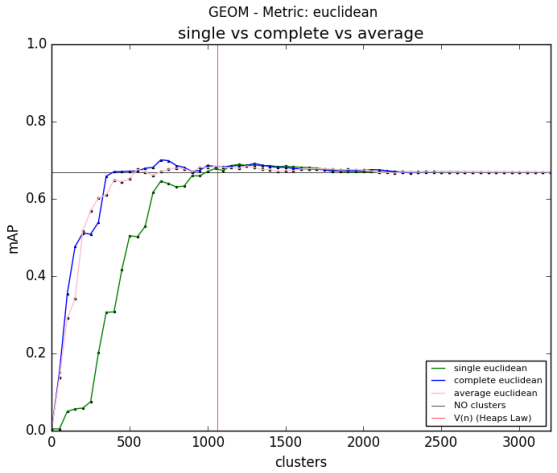

Fig. 8: Comparison between single, complete and average linkage methods for Euclidean metric.

Results show how the hypothesis is confirmed, with considerable less samples. With only the representative ones, its possible to get the same results as with the full collection, or even better if the number of clusters is well chosen. For this dataset a value of 801 clusters gives the maximum mAP, this is a 20.01% of total samples.

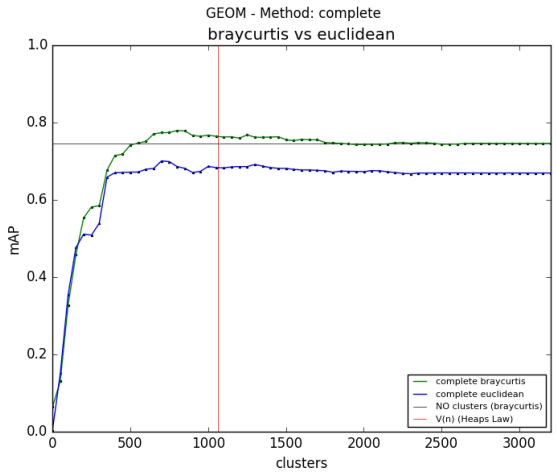

Fig. 9: Comparison between Braycurtis and Euclidean metrics.

In the best case a cluster per word is achieved, it could be said that clusters represents the vocabulary of the collection.

The proportion is pretty good, but as formulated in [6] the vocabulary of a text is given by

$$V(n) = K * n^B$$

where n is the total of words and K and B are determined by the language of the collection. Empirical results says that the bigger is the collection the hardest is to find new words, languages have a finite number of words in their vocabulary. So, for larger datasets this proportion could be considerably smaller.

This study [6] solves another important inconvenient: normally is not possible to know how many clusters have to be set, and given the relation that each cluster represents a word of vocabulary $V(n)$ gives that parameter.

With $K = 7.2416$ and $B = 0.6172$ taken from the same study [6] is obtained $V(3234) = 1061,8$ (32.81% of samples) which fits pretty well with the number of clusters where maximum mAP is obtained as seen in figures 7, 8 and 9.

We can see in table of Fig. 11 that between 651 and 1701 formed clusters, the results of clustering retrieval are better than traditional one for handwritten word images. The minimum number of representatives needed to get best results are 651 and the maximum mAP is obtained at 801 representatives (24.7%).

Conclusions

In the present paper, effects of many parameters when creating HAC are studied for a handwritten word images retrieval tasks. Given the particularity of a language vocabulary, is seen that explained algorithm can accomplish great improvements in retrieval tasks.

Provided method is highly scalable, large growth of manuscript collection does not necessarily imply a significant increase in cluster representatives. If efficiency is considered as the required representatives over the dataset cardinality, arguably efficiency is increased with big datasets.

Moreover, cluster retrieval reinforce similarity grouping tasks by ensuring that each cluster samples are retrieved together. The result of this is that better precision could be obtained with this algorithm than performing traditional methods.

May further research in handwritten word images feature extraction to fit the given method could bring even better results. Or another approach, for future investigation, could be to find some distance that fits better the given features.

Acknowledgements

I thank Marçal Rossinyol, this project's supervisor, for his assistance, guidence, advice and patience.

Referències

[1] Miquel dels Sants Gros, *Els arxius històrics catalans i la restauració de monuments*, Facultat de Teologia de Barcelona.

[2] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, *Introduction to Information Retrieval*,Cambridge University Press, 2008.

[3] R. Manmatha and W. B. Croft, *Word Spotting: Indexing Handwritten Archives. In: Intelligent Multimedia Information Retrieval Collection*, M. Maybury (ed.), AAAI/MIT Press 1997.

[4] D. Aldavert, M. Rusiñol, R. Toledo and J. Lladós, *A Study of Bag-of-Visual-Words Representations for Handwritten Keyword Spotting.*, International Journal on Document Analysis and Recognition. 18(3):223-234, September 2015.

[5] Murtagh F, *Complexities of Hierarchic Clustering Algorithms: the state of the art*, Computational Statistics Quarterly. 1, 1984, Pages 101–113.

[6] Heaps, H. S. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Orlando, FL, 1978.

Metric	**Linkage**	**Representative.**	**Clusters**	mAP	mAP_0
Cosine	Average	Geometric	4	0.97	0.90
Cosine	Average	Geometric	5	0.96	0.90
Euclidean	Single	Geometric	28	0.96	0.971
Euclidean	Single	Random	25	0.96	0.91
Canberra	Single	Random	36	0.96	0.93
Euclidean	Single	Mean	30	0.95	0.91

Fig. 10: Top mAP using clustering.

	Metric	**Linkage**	**Representative**	**Clusters**	mAP	mAP_0
1st	Braycurtis	Complete	Geometric	650	0.77	0.75
Max	Braycurtis	Complete	Geometric	800	0.78	0.75
Bests	Braycurtis	Complete	Geometric	650 - 1700	$>$0.75	0.75

Fig. 11: mAP results using clustering (metric: Braycurtis).