Clustering

Katelyn Harlan
University of Otago
Dunedin, Otago, New Zealand
harka424@student.otago.ac.nz

ABSTRACT

Clustering is a complicated topic to tackle, and has many applications across many different fields of computer science. In this paper we will explore some different results obtained in regards to measuring the effectiveness of clustering - particularly in the field of information retrieval, as well as some more mathematical analysis of clustering itself. We will also pose some questions as to how we can further research in this field.

CCS CONCEPTS

• **Mathematics of computing** → *Cluster analysis*; *Graph theory*; • **Information systems** → **Clustering**; Document topic models; Content analysis and feature selection; • **Computing methodologies** → **Cluster analysis**; • **Applied computing** → *Document management and text processing*.

KEYWORDS

clustering, evaluation of clustering, clustering comparison, effectiveness measures, document clustering, topic models, collection representation

1 INTRODUCTION

Clustering involves partitioning some set, e.g. a document collection, into a set of clusters. In information retrieval, clustering can help to "guide retrieval"[3]. We should find that similar documents appear in the same clusters. If a document is relevant to some query, then clearly we would expect the other documents in said cluster to also be relevant to the given query. It is with this concept that we can see the possible benefits of employing clustering in this field.

But, a more pertinent question is: how can we measure the effectiveness of clustering? We need to have more robust measures for evaluating clusters. This will also be useful in determining how effective clustering really is, and how we can improve upon it. These ideas are some of the main motivations for the three research papers discussed in this text. In fact, it was in the pursuit of finding a reasonable way to measure the similarity of two clusterings that these papers were intially found.

In the first paper, Measurement of clustering effectiveness for document collections[3], we see an investigation into some different techniques for measuring the effectiveness of clustering. The second paper, Document Clustering vs Topic Models: A Case Study[2], explores two approaches to characterising document collections. In particular, they compare the performance of clustering and topic modelling as description tools on the WSJ collection as a case study. Finally, the third paper, Comparing Two Clusterings Using Matchings between Clusters of Clusters[1], is a mathematical exploration of clustering. Notably, we see the introduction of the D-family-matching problem on intersection graphs of two clusterings.

2 MEASUREMENT OF CLUSTERING EFFECTIVENESS FOR DOCUMENT COLLECTIONS

2.1 Research Question

We begin with an exploration into the effectiveness of clustering. Specifically, this paper poses the question: how might we go about measuring the effectiveness of clustering for document collections?

2.2 Main Contributions

There are two main contributions provided in this paper; they examine the use of extrinsic techniques in cluster evaluation, and they examine, in the context of information retrieval, the extent of effectiveness of the standard methods of clustering.

The use of extrinsic techniques is not foreign to information retrieval. The idea of using human judgements on whether or not certain documents are related to certain queries is well known. It is whether these techniques can be useful in the evaluation of clusters that is uncertain. If, through the use of such techniques, we know that the documents that are clustered together are, in fact, relevant to one another, then we can find some indication as to the quality of the clustering. Not only did this paper find that such extrinsic techniques were useful in cluster evaluation, but that the need for human judgement at all is unnecessary; the results obtained were sufficiently similar to those based on the documents from retrieval systems.

The indication that results from retrieval systems may be as useful as human judgement is promising - as human labelled data is much more difficult to obtain. In the paper, there are four measures introduced that will allow us to observe this discovery, where p is a proportion indicating the coverage of documents amongst the clusters.

$$R_c@p = \frac{\text{min number of clusters to cover } p\% \text{ of the relevant docs}}{\text{total number of clusters}}$$

$$R_v@p = \frac{\text{min total size of clusters to cover } p\% \text{ of the relevant docs}}{\text{total number of docs in those clusters}}$$

$$F_c@p = \frac{\text{min number of clusters to cover } p\% \text{ of the retrieved docs}}{\text{total number of clusters}}$$

$$F_v@p = \frac{\text{min total size of clusters to cover } p\% \text{ of the retrieved docs}}{\text{total number of docs in those clusters}}$$

For an easier comparison of these results, the gain is reported. That is, an estimate of the natural floor is used as a baseline, and the improvement/gain is then observed between a given clustering and the random partitioning.

The experiment discussed in this paper is, essentially, taking some clustering techniques of various quality, applying said techniques to various datasets, and observing the results of the measures. An example of such results is given in figure (1), in which we see the gain reported for the aforementioned measures on the four datasets

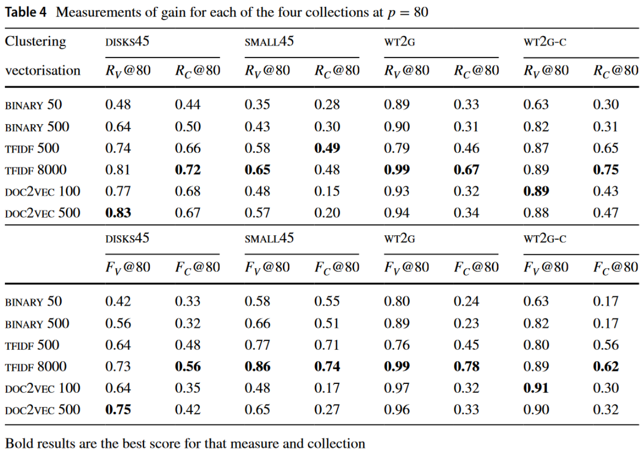

Table 4 Measurements of gain for each of the four collections at $p = 80$

Clustering	DISKS45		SMALL45		WT2G		WT2G-C	
vectorisation	R_V@80	R_C@80	R_V@80	R_C@80	R_V@80	R_C@80	R_V@80	R_C@80
BINARY 50	0.48	0.44	0.35	0.28	0.89	0.33	0.63	0.30
BINARY 500	0.64	0.50	0.43	0.30	0.90	0.31	0.82	0.31
TFIDF 500	0.74	0.66	0.58	**0.49**	0.79	0.46	0.87	0.65
TFIDF 8000	0.81	**0.72**	**0.65**	0.48	**0.99**	**0.67**	0.89	**0.75**
DOC2VEC 100	0.77	0.68	0.48	0.15	0.93	0.32	**0.89**	0.43
DOC2VEC 500	**0.83**	0.67	0.57	0.20	0.94	0.34	0.88	0.47
	DISKS45		SMALL45		WT2G		WT2G-C	
	F_V@80	F_C@80	F_V@80	F_C@80	F_V@80	F_C@80	F_V@80	F_C@80
BINARY 50	0.42	0.33	0.58	0.55	0.80	0.24	0.63	0.17
BINARY 500	0.56	0.32	0.66	0.51	0.89	0.23	0.82	0.17
TFIDF 500	0.64	0.48	0.77	0.71	0.76	0.45	0.80	0.56
TFIDF 8000	0.73	**0.56**	**0.86**	**0.74**	**0.99**	**0.78**	0.89	**0.62**
DOC2VEC 100	0.64	0.35	0.48	0.17	0.97	0.32	**0.91**	0.30
DOC2VEC 500	**0.75**	0.42	0.65	0.27	0.96	0.33	0.90	0.32

Bold results are the best score for that measure and collection

Figure 1: Gain values for $p = 80$ on four collections, via [3]

used in the paper. We notice that although there are some differences between the R and F measures (which the paper explains further), there are clear similarities in the results. The most prominent similarity being the general agreement in which vectorisation method provided the best score.

As for the effectiveness of standard clustering methods, it was found that the clusterings were generally meaningful; the clustering is "grouping documents together in a way that is consistent with their contents"[3]. However, it is not yet clear if this is entirely useful to information retrieval. We also note that it appears as though the quality of a clustering may decrease with an increase in the size of the collection. It may be worth looking into this further, especially given the scale often encountered in information retrieval.

3 DOCUMENT CLUSTERING VS TOPIC MODELS: A CASE STUDY

3.1 Topic Modelling

As a brief aside, we shall highlight topic modelling - in which we take each topic to be a mapping from the set of terms in our collection to a set of weights. The words which are highly weighted for a topic should be representative of some semantic theme.

3.2 Research Question

Here, we present a case study comparing clustering and topic modelling as description tools for the WSJ document collection. This comparison being the main driving force behind the paper.

3.3 Main Contributions

The main contributions for this paper are as follows: two simple cluster labelling methods are compared, significant alignment between the clusters and topic labels on the WSJ collection is found, and said topic-cluster alignment is used to show that words/documents close to the centroid are not good representations of the clusters.

For their first contribution, they highlight two basic methods for labelling clusters: the central keywords method, and the cluster keywords method. In central keywords, we select the N highest-weighted words from only the documents close to the centroids. For cluster keywords, we again select the N highest-weighted words - but from amongst all the documents in the cluster. It is clear that these methods are similar in nature and name, and it is unfortunate that more care was not taken to differentiate the two. Some areas of the text require further clarification, for example, in the conclusion there is a description of what appears to be the central keywords method as being "highly effective"[2], and yet elsewhere there are plentiful remarks that this method is a "complete failure"[2], and that the cluster keywords method was found to be more reliable. Nonetheless, it is easily observed in table (1) that the cluster keywords and topic labels share many more similarities than the central keywords. Further, it was found that there is significant alignment between the clusters and the topic labels, at least for the corpus in question. This paper found this result surprising, since clustering and topic modelling are rather different approaches. It would be very interesting to see these results replicated with other collections, or for this discovery to be more generalised.

Cluster terms	quarter, cent, net, share, earn, loss, sale, revenu, profit, incom
Topic terms	quarter, cent, share, net, earn, loss, sale, profit, revenu, incom
Centre terms	laser, pnc, hansen, doctrin, billion, mile, mercer, quarter, daughter, court

Table 1: An example of keywords and topic labels, via Table 3a [2]

The topic-cluster alignment was also used to demonstrate the unsuitability of choosing words/documents close to the centroid as a representative of the cluster as a whole. It seems perfectly logical to assume that an item close to a centroid would be a good representative of the cluster. However, likely due to the nature of working in higher dimensions, this was not found to be the case - or at least, it was not found to be true for the collection in question.

4 COMPARING TWO CLUSTERINGS USING MATCHINGS BETWEEN CLUSTERS OF CLUSTERS

4.1 Research Questions

In this paper, we take a step away from information retrieval and consider clustering in a more theoretical sense. The most prominent question lies in the exploration of the relationship between two clusterings, and how we might define this. But another question posed is in regards to the scale at which cluster might merge.

4.2 Main Contributions

There are a plethora of contributions provided in this paper, all very mathematical in nature. We shall briefly outline what these are, and then focus on the more interesting results. A new combinatorial optimisation problem, D-Family-Matching, is introduced - and then proven to be very difficult (NP-complete and APX-hard). The paper describes some efficient algorithms for general graphs, and then shows these algorithms are capable of identifying meta-clusters

(clusters of clusters) between a given clustering and an edited version. Finally, they provide some insights into the scale at which clusters combine.

The problem of D-Family-Matching introduced in this paper is very interesting. From two clusterings, F and F', we obtain the intersection graph G. We take a node of G to represent a cluster, with an edge between two such nodes indicating that the intersection between said clusters is nonempty - or rather, that there is at least one element shared between them. In fact, the weight of an edge is the number of elements common to both clusters. The idea of this problem is to find "an explicit many-to-many correspondence between groups of clusters of the two clusterings"[1]. The parameter D being an upper bound of the diameter of the graph G.

Let us now consider a reasonably straightforward example from the paper, using a 2D dataset of 40 points.

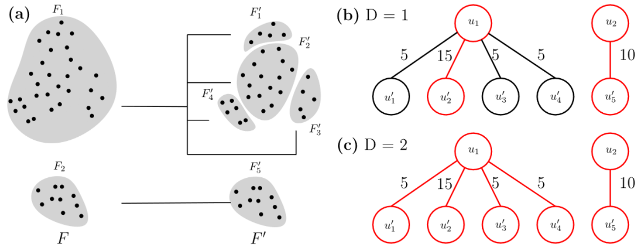

Figure 2: Example of D-Family-Matching, via [1]

In (a) we see the two different clusterings F and F' of the dataset, with the D-Family-Matching graphs for $D = 1$ and $D = 2$ showcased in (b) and (c) respectively. In the case of $D = 1$ (1-to-1), we only match the clusters with the highest weights - this is the method most existing methods use. With $D = 2$, we are now able to match F_1 to the meta-cluster $\{F'_1, F'_2, F'_3, F'_4\}$ - this is the 1-to-many case.

If we now look beyond the example in figure (2), we may wonder what $D > 2$ might represent. For such values, we observe the many-to-many cases. These deal with the case where we have different clusters from each clustering corresponding to each other without a "good" matching. We can think of D as being indicative of the complexity of the involved meta-clusters.

This is all very exciting, but may it be unclear as to how this relates to the other papers discussed, or to information retrieval at all. In fact this approach to clustering, and the introduction of the variable D, may bring us closer to figuring out how to appropriately choose the number of clusters. Further exploration of their works, perhaps with a particular focus on information retrieval, would prove very interesting.

5 RELATIONS BETWEEN THE PAPERS

It is immediately obvious that all three papers chosen are primarily focus on clustering; it is the main link between them. Further than this though, it is the exploration of evaluation that truly ties these papers together.

In [3] and [1], we see an exploration into the evaluation of clustering. The former primarily focuses on the effectiveness of certain techniques for cluster evaluation, whereas the latter delves more into the relationship between two clusterings. In both papers, it is clear that there is an investigation into the evaluation of clustering.

As for [2], there are more similarities held to [3] than to [1]. Ignoring the obvious connection of the papers sharing the same group of authors, we see that they share the focus of clustering in the context of information retrieval - particularly looking at the clustering of documents. To motivate advancements of one field, e.g. information retrieval, it is important to consider related advancements that originate elsewhere. It is with this idea that we may consider [1] as being a suitable selection.

The link between [2] and [1], then, is very much the weakest. It is with the inclusion of the other paper that we are truly able see to much of a link at all - besides, of course, the central theme of clustering.

6 IS THERE A UNIQUE IDEAL CLUSTERING?

If we revisit figure (2), we observe that for the same set of data we have two different clusterings. This inspires the question: is there a unique ideal clustering? Further, is there some ideal number of clusters? Or perhaps an ideal partitioning of a given set? There are many more such questions that come to mind.

For individual collections, there must be some unique clustering that best suits the desired task. How might we find this, and what would we consider to be "best"? In a more general sense, is it possible to find an ideal number of clusters?

Using the results regarding cluster comparison and evaluation, we should be able to determine if a given clustering of an specific collection is the "best" such clustering. The knowledge of the collection will allow us to have some notion of "best" for the given context. However, the more difficult question remains: how can we reliably find this clustering?

In fact, [3] and [1] both touch on this question. With [1] in particular assessing their parameter D and its implications for choosing the "correct" number of clusters. Although we neglected to touch on the concept in this text, the notion of cluster stability is also investigated by both of these papers. This could also be beneficial in the answering of our question.

Perhaps, then, clustering could benefit from a proper investigation into the choice in the number of clusters. Such an investigation could begin with a thorough review of existing methods and techniques for partitioning some dataset into clusters. Then, we could examine the effects of cluster sizes, and the number of clusters, possibly even incorporating this notion of D. This exploration would be based in the context of information retrieval, but would clearly have implications for clustering as a whole.

7 CAN MORPHOLOGICAL TECHNIQUES IMPROVE CLUSTERING?

If we, again, assume that we can reliably evaluate our clusters and clusterings, might there be a way we can improve upon them? Specifically, could morphological techniques such as stemming improve the quality of the clustering? Further, how would this effect any meaning potentially held within the clustering?

To investigate this, this author proposes that we begin by exploring the effects of said morphological techniques on some clusterings of a well-known corpus. The results of this should give some indication as to whether such a proposal is at all useful. It will also allow us to more closely examine the nature of the clusters themselves, perhaps shedding some light on what meaning these clusters hold.

8 CONCLUSION

Clustering has much room for further investigation. At first glance it may seem like a simple problem (grouping things together), but as we have seen, even evaluating a set of clusters can be very complex. It is a very interesting field of study, and one that could benefit from further research.

In [3], we saw an examination of different techniques for measuring the effectiveness of clustering, as well as an exploration into the suitability of standard clustering methods for information retrieval. From this we have discovered several measures that can be used to evaluate a clustering (in the context of information retrieval). Another result of note was the indication that the quality of a clustering may decrease as the size of the collection increases.

[2] gave us a good comparison of document clustering and topic modelling. In particular, we got to explore clustering as a descriptive tool - and we saw the unsuitability of using items close to the centroid as a representative of a cluster.

Finally, in [1], we were introduced to the D-Family-Matching problem. This paper holds some very exciting results, and it would be very interesting to see them applied to information retrieval.

REFERENCES

[1] F. Cazals, D. Mazauric, R. Tetley, and R. Watrigant. 2019. Comparing Two Clusterings Using Matchings between Clusters of Clusters. *ACM J. Exp. Algorithmics* 24 (Oct. 2019), 1–41. https://doi.org/10.1145/3345951

[2] Meng Yuan, Pauline Lin, and Justin Zobel. 2022. Document Clustering vs Topic Models: A Case Study. In *Proceedings of the 25th Australasian Document Computing Symposium (ADCS '21)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3503516.3503527

[3] Meng Yuan, Justin Zobel, and Pauline Lin. 2022. Measurement of clustering effectiveness for document collections. *Information Retrieval Journal* 25 (Jan. 2022), 239–268. https://doi.org/10.1007/s10791-021-09401-8