Document Clustering vs Topic Models: A Case Study

Meng Yuan
University of Melbourne
Melbourne, Victoria, Australia
myuan3@student.unimelb.edu.au

Pauline Lin
University of Melbourne
Melbourne, Victoria, Australia
pauline.lin@unimelb.edu.au

Justin Zobel
University of Melbourne
Melbourne, Victoria, Australia
jzobel@unimelb.edu.au

ABSTRACT

Document collections can be characterised in a variety of ways. Two key approaches are clustering, which partitions collections into subcollections with the expectation that the contents will be thematically linked, and topic models, which describe the contents in terms of weighted lists of words that are expected to represent different themes. In this paper, we report experiments on the observed relationship between clusters and topic models in a preliminary study of a large text collection. Both produce results that appear cohesive in their own right, but surprisingly – given the very different ways in which they are formed – the descriptions of the collections that they generate are strongly similar. This unexpected mutual reinforcement creates confidence in both approaches as tools for annotating and describing the contents of document collections.

CCS CONCEPTS

• **Information systems** → Document collection models; *Document topic models*; *Content analysis and feature selection.*

KEYWORDS

document clustering, topic models, collection representation

ACM Reference Format:
Meng Yuan, Pauline Lin, and Justin Zobel. 2021. Document Clustering vs Topic Models: A Case Study. In *Australasian Document Computing Symposium (ADCS '21), December 9, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3503516.3503527

1 INTRODUCTION

A challenge in information retrieval (IR) is for users to understand the scope of the collection to which they are posing queries. Two broad automatic approaches to characterisation of collections can be used to assist with this task. One of these is clustering, in which the collection is partitioned into subcollections with the goal of the subcollections being distinct from each other and more narrowly themed than is the collection as a whole. The other is topic modelling, in which each topic is a mapping from the set of terms (words) in the collections to a set of weights; the highly weighted words for a topic are anticipated to represent a semantic theme.

ADCS '21, December 9, 2021, Virtual Event, Australia

ACM ISBN 978-1-4503-9599-1/21/12…$15.00
https://doi.org/10.1145/3503516.3503527

In this paper, as a case study of the potential value of clustering and topic modelling as descriptive tools we examine the relationship between them on a curated document collection. Both of these approaches are well known and have been used in IR for a variety of tasks. Clustering has been proposed as a mechanism for supporting retrieval directly via iterative collection narrowing and indirectly as an enhancement to retrieval models [19, 28, 29]. Applications of topic modelling include document representation and collection summarization; it has not been widely used for collection partitioning, but note that we do not explore that problem here. Topic models have however been used to improve the performance of clustering in a variety of ways [9, 22, 33].

Our goal is to compare clustering and topic modelling as description tools, using a newswire collection as a case study. There are three main contributions in this work. First, we compare two simple cluster labelling methods, which are intended to create compact thematic representations of their contents. Second, on this collection, and subsets of it, we show that there is significant alignment between cluster membership and the major topic labels of documents. We regard this is as a highly surprising result, as the methods of creating these representations is very different. As we show, the alignment can be observed in two different ways: through the dominance of specific topics within individual clusters, and through inspection of the significant words in clusters and topics. That is, in a complementary way we have compared clustering to topic modelling by mapping them in both directions.

Third, we use the topic–cluster alignment to demonstrate that, in this case study at least, the words in documents that are close to the centroids of clusters are a poor representative of the cluster as a whole. We hypothesise that this arises because even the nearest document to a centroid may be relatively distant, and may be close to the centroid on dimensions that are untypical of the cluster as a whole; this kind of counter-intuitive behaviour easily arises in high-dimensional data.

The results of our case study show that the clustering and topic-modelling methods can be mutually confirming and appear to be generating semantically meaningful representations, while also suggesting that there are limits to the assertions that can be made about how distinct different topics, and different clusters, are from each other. We believe that these results establish the need for a more systematic investigation of the topic–cluster relationship and that, if confirmed, it can provide the basis of rich mechanisms for exploration of document collections.

2 BACKGROUND

Our work explores topic modelling and clustering for information retrieval. We first describe these approaches, then review work in which they have been integrated.

In this paper we denote a document collection as D. The vocabulary of collection D is denoted as V, which is the set of all words appearing in the collection. A document from collection D is denoted by $d \in D$ and is regarded as a set of words w.

Clustering. Clustering has a long history in IR, and up to the 1980s was regarded as a plausible search mechanism, under which, it was supposed, users would iteratively navigate to ever smaller clusters rather than issue a query [14, 30]. While such approaches did not continue to be explored, as collections grew in size and index-based search technology matured, there has been ongoing interest in clustering in its own right and as a support for a range of search-related tasks [2, 13, 15–17].

A range of mechanisms have been proposed for clustering. One approach is bottom-up methods, in particular agglomerative or hierarchical clustering [10, 11, 14, 34]. The other is top-down methods, in particular K-means [4, 8, 23]. In practice, agglomerative methods have not been found to scale to realistic collection sizes, and reported experiments on large collections all seem to use K-means.

We therefore use we use K-means for the work in this paper. Although the robustness of K-means has been questioned, for example with regard to stability and the potential to fall into local minima rather than produce globally optimal outcomes, in a separate project (manuscript in submission) we have observed reasonably stable performance. In brief, in K-means clustering an expected number of clusters, k, is chosen. Initially k documents are chosen as seeds, and each document in the collection is allocated to the cluster for which the seed is the nearest neighbour. At each subsequent iteration, the centroid of each cluster is computed; this is the average of the allocated documents. The documents are then allocated again to the cluster with the nearest centroid. This repeats until an iteration limit is reached or the process converges, that is, the clusters are unchanged between iterations. The value of k can be optimised by exploring behaviour as k is varied, for example with the Elbow method [27], but in this work we use fixed k based on results in the separate project mentioned above.

The concept of 'nearest neighbour' requires a document representation. In the results reported here, we use a TFIDF calculation of word weights so that documents are represented as a vector whose dimensionality is $|V|$. A straightforward Cosine formulation is used to determine similarity.

To describe clusters we need to label them [24]. Three elementary methods are high-weight terms in the centroids, mutual information (the terms that best discriminate from other clusters), and information drawn from document titles. A wide range of other methods have been described, including use of WordNet [26], choice of terms from concept graphs for clusters [1, 5, 12], and use of neural networks [20]. Here, we only consider simple methods.

Topic modelling. Latent Dirichlet allocation (LDA) is a probabilistic Bayesian approach to topic modelling proposed by Blei et al. [3]. In contrast to K-means clustering, documents are assumed to be comprised of multiple topics. Intuitively, topics are distinguished by the different degrees to which they are associated with words in the vocabulary; a word that is prominent in one topic may be insignificant in another. The approach assumes that a collection covers a known number m of topics and that each document is a blend of a subset of these topics. It also assumes that there is an underlying data generation process in which, first, the topics of a document follows a multinomial distribution, controlled by a common Dirichlet prior for the collection; and second, for each of the topics, the words are generated from another multinomial distribution, controlled by a Dirchlet prior that is shared among topics. The two distributions are assumed to be independent.

LDA learns the distributions governing this generative process for a given corpus as follows. Initially there is a random assignment of topics to each document and of the words in each document to topics. It then iteratively fits a word in each document to a topic by updating the following probabilities.

- The topic-word distribution, $p(w|t)$, as the number of documents that assigns word w to topic t divided by the number of documents containing w.
- The document-topic distribution, $p(t|d)$, as the number of words in d that are assigned to topic t divided by the total number of words in d.

This continues until the process converges, or a limit on the number of iterations is reached. The updated values allow estimation of the probability of a word being assigned to a specific topic in a given document, that is, $p(w|t, d) = p(w|t) \times p(t|d)$. Word w in document d is then re-assigned to the topic (amongst the set of topics) with the highest probability, that is, the topic given by $\text{argmax}_i\, p(w|t_i, d)$.

To allow training of an LDA model, documents need to be transformed into vectors as for K-means. We use the common vectorisation method TFIDF, whose dimensionality is $|V|$, to train the LDA models for our experiments.

Words with high probabilities can be regarded as the keywords, or the signature, of a topic; by construction, different topics will tend to have very different keywords. It has been shown that lists of keywords generated by topic modelling are effective as topic descriptors that are understood by humans [7, 18]. There are several approaches to term selection, such as ranking words in a topic by pairwise mutual information (PMI) and selecting the highest-scored for presentation [25], or use of graph-based topic descriptors [6], which however are not directly suitable for human consumption.

In this initial work, we take a simple 'top N' approach to selecting lists of keywords, but note that there are richer methods that are argued to give results that are better appreciated by humans. Our primary goal here is to observe alignment with clustering rather than to optimise for interpretability.

Combined methods. Lu et al. [22] explored an approach to integration of topic modelling with clustering. They compared the performance of two topic modelling methods, pLSA and LDA, in the context of document clustering, considering two ways in which topic modelling and clustering can interact. The first is to represent documents with topic distributions and then use the topic distributions as vectors, replacing the original document vectors. The second is to treat topics as clusters and pick the topic with the highest probability as the predicted cluster for each document. However, Lu et al. did not make use of clustering algorithms, but instead treated a given partitioning as a collection of clusters; there was no exploration of the relationship between the approaches.

Xie and Xing [33] proposed a framework in which topic modelling and clustering were integrated, simultaneously learning the global topics of a collection and local topics in clusters. In addition

to the LDA model, in which the document collection is represented as a mixture of topics [3], documents are partitioned into groups where each group (that is, cluster) has its own topic model with multiple local topics. This work shows that clusters can be described with the top-ranked words from local topics, in addition to the previously explored approach of describing the whole collection with global topics. In similar work, Curiskis et al. [9] examine performance of several clustering and LDA methods with ground truth on social network text; approaches considered include hierarchical clustering, which we chose to not examine because it does not scale to collections of realistic size.

Contrasting these approaches, in the work of Lu et al. [22] clusters are identified by most significant topic and therefore the number of clusters is naturally equivalent to the number of topics. In the work of Xie and Xing [33], clusters are considered as mixtures of multiple local topics and global topics are mixtures of clusters. Clustering has also been used in a range of ways in IR that are arguably similar. For example, Liu and Croft [19] use clusters to refine similarity measures based on language models, and Kulkarni and Callan [15] use clusters in a similar way.

Lossio-Ventura et al. [21] compare a rich variety of topic modelling methods to clustering, though not in ways that are directly comparable to our study. They report comparison of methods using internal clustering metrics in a unified way and compare correlation with external ground truth, but it is not clear whether the measured values, which are low, indicate significant alignment.

3 CLUSTER-BASED CONTENT DESCRIPTORS

An initial contribution of this paper is comparison of two simple methods for cluster labelling, which we now explain. These are intended to give a human-interpretable description of the semantics of the topic. In contrast to some of the work discussed above, in our research clusters are represented with keywords and are determined independently of consideration of topic models.

Keywords from central documents (central keywords). A standard visual explanation of clusters is a presentation of them as multiple clouds of dots in a two-dimensional space. The clouds overlap but, appealingly, are sufficiently distinct that they can be perceived as organic units – an effect that can be enhanced through use of different colours. A feature in some of these visualisations is that the dot-clouds are more crowded at the centre, just as a galaxy is dense in the middle and sparse at the edges.

The intuition the dot-cloud metaphor suggests is that there are some archetypal items (documents in this study) at the centre of each cluster that can stand as ideotypes for the whole. In the first of our two approaches to generating representative word lists, we assume that the documents that are closest to the centroid are a good proxy for the whole, so that word lists generated from those documents will be informative about the cluster.

In detail, the Cosine distance between each document in the cluster and the centroid is calculated. We select the top 10 documents closest to the centroid and pool them, and select the top (highest-weighted) N words as the representative keywords.

Keywords by all documents of a cluster (cluster keywords). An alternative to using the documents that are close to the centre is to use them all. Intuitively, this is an appealing approach if the documents are reasonably homogeneous but – considering the dot-cloud metaphor again – the documents that are remote from the centroid may be on rare subjects and be distant from the other documents in the cluster. That is, they may dilute the topical integrity that might be observed closer to the cluster's centre. However, we thought this a useful alternative to explore.

Straightforwardly, we determine the vector for a cluster as the average of all the vectors in the cluster – that is, we use the centroid. The top (highest-weighted) N words from the centroid are selected as the representative keywords, that is, we use elementary cluster labelling as described above.

Collection	No. docs
WSJ	98,733
WSJ-LONG	58,120
WSJ-SHORT	40,613

Table 1: Sizes of collections used in experiments.

4 EXPERIMENTS

We now compare clusters and topic models, considering topic composition of clusters and then comparing their word representations.

The dataset we use is the TREC WSJ corpus [31]. Words in the corpus are stemmed; the 'words' shown in our results below are the stemmed versions, not the original words. Stopwords and words that appear in more than 50% of the documents are removed.

In the data pre-processing stage, we observed that about 40% of the documents are 100 words or shorter. Short documents may contain insufficient information for LDA to accurately learn topic and word distributions. Also, the short documents, when mixed with other, long documents, might become super clusters because their features have higher sparsity, as K-MEANS may perform poorly when documents are varying in size and density.

We therefore decided to explore the methods on subcorpuses as well as the whole collection, splitting WSJ into WSJ-SHORT, of documents containing 100 words or less, and WSJ-LONG, of documents with more than 100 words. The size of each collection is summarised in Table 1. Independent representations were trained for each of the collections.

In order to observe the alignment between clustering and topic models, we set a degree r for the two alternative representations of collections, to help ensure comparability. For example, if $r = 5$ then we will run K-MEANS clustering with $k = 5$ and LDA with 5 topics. In the results reported here, we set degree $r = 20$.

Cluster and topic overlap. In our first experiment, we examine the topic composition of clusters, aiming to match topics with clusters.

For each document, we generate the topics with the LDA model; the topic with highest weight is chosen as the *dominant* topic label of the document. Hence, a document has a cluster label and a dominant-topic label and in effect we are using topic modelling as a form of clustering.

Results are shown in Figures 1, 2, and 3 for WSJ, WSJ-LONG, and WSJ-SHORT respectively. Each horizontal bar is a cluster, and each coloured segment within a bar is a different topic. (Note that the

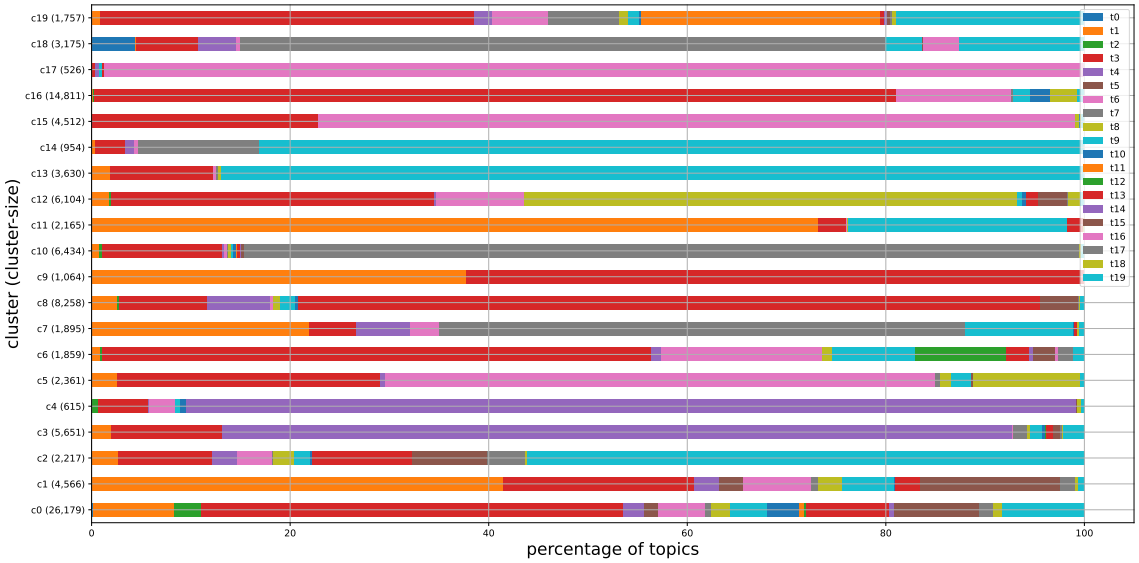

Figure 1: WSJ **cluster-topic distribution. Each topic has a colour. The sizes of the segments in each bar represent the percentage of documents in which the corresponding topic is the most significant topic. The size of each clusters is shown next to the cluster label on y-axis.**

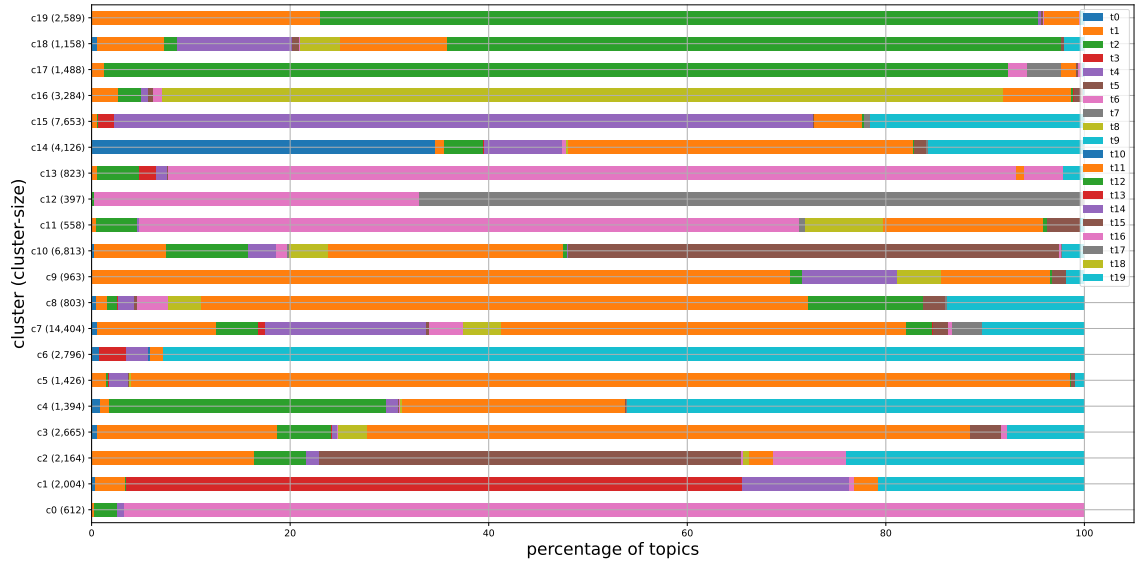

Figure 2: WSJ-LONG **cluster-topic distribution, as per Figure 1.**

colours in the different figures are unrelated to each other.) The number of documents in each cluster is shown in the label on the y-axis, while faint vertical lines mark pentiles from 0% to 100%.

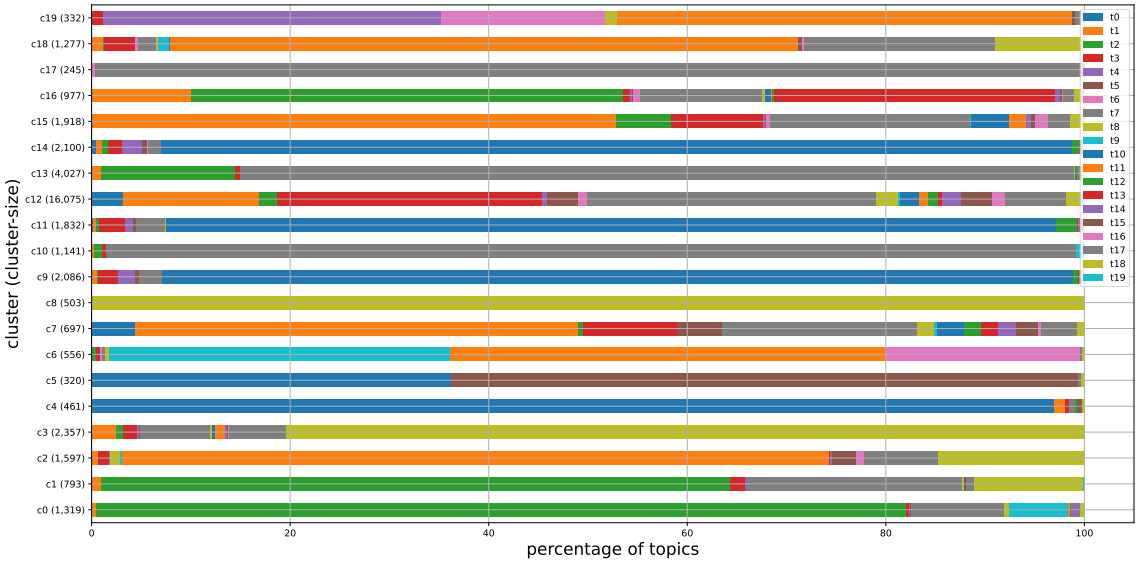

Figure 3: WSJ-SHORT **cluster-topic distribution, as per Figure 1.**

Thus, for example, considering the bar for cluster c_8 in Figure 1, it can be seen that there is representation of 11 topics across the 8,258 documents; further topics may be present but in such small numbers that they are insignificant (and thus invisible). A single topic, t_{13}, accounts for around 70% of the documents while the next biggest, topic t_3, accounts for around 10% of the documents.

A perfect alignment would be that each cluster only contains one topic label and that each topic only appears in one cluster. Visually, most horizontal bars are predominately filled by one or a small number of segments. For WSJ, for example, in 17 of the 20 clusters more than half the documents are from a single topic; for WSJ-LONG the result is 16 of 20 and for WSJ-SHORT it is 14 of 20. That is, for most clusters, the topic distribution is very skewed and gravitates towards one or few topics.

These results show that independently generated K-MEANS clusters and LDA topics are well aligned. We regard this result as highly surprising. The mechanisms by which the topics and clusters are arrived at is very different, and one is not a simple transform of the other: clustering involves the non-linear Cosine measure, while LDA allocates mixes of topics to documents and does not involve document comparisons.

Our expectation had been that the two approaches would produce essentially different results; indeed, in the early stages of this project our working hypothesis was that, at scale, clustering would become indistinct from a random partitioning. However, contradicting our scepticism, the strength of correlation shows that there are true underlying properties that these methods are exploiting.

A few clusters show a more even topic distribution of a good number of topics, in particular c_0 and c_{16} in WSJ, c_7 in WSJ-LONG, and c_{12} in WSJ-SHORT. A common characteristics of these clusters is that they are large compared to other clusters in the collection. These appear to be the dense clusters that K-MEANS does not break down, at least at this value of k.

Keywords comparison. In the results above, topic modelling was treated as a clustering mechanism. Here, we explore the complementary process. Our experiments examine use of clustering as a mechanism for generating topic descriptions, using the two methods described earlier for selecting keywords based on clusters, and compare the output to that of topic modelling.

For the three methods – topic models, cluster keywords, and central keywords – we show results on the three collections in Table 3 (presented at the end of the paper). For each collection, we show the ten cluster–topic pairs with the highest percentage match, defined as the percentage of the documents in a cluster that have that topic as their dominant topic. Each cluster and topic is shown only once; if a topic is well represented in two clusters, only the first is shown.

As the tables show, there are striking similarities between the top words found from the whole cluster and the topic model (the first and second columns), in all three collections. All are highly correlated and many are nearly identical. As discussed above, we had not anticipated this result.

The results also include a small number of exceptions, in particular c_5–t_{11} in WSJ-LONG and c_8–t_8 and c_{17}–t_{17} in WSJ-SHORT, where the words are different but there is evident topical similarity. Surprisingly, these matches have the highest match percentages in their respective collection, as shown in the figures above. These are also smaller clusters. Whether these are natural outliers, or indicative of some other underlying behaviour, is difficult to assess on this small volume of data.

Cluster terms	sale, unit, amp, inc, share, product, corp, oper, sell, plant
Topic terms	mr, busi, bush, one, time, get, peopl, go, like, work
Central terms	batteri, daewoo, tariff, ual, dispos, coal, md, mexico, ge, del

Table 2: A low-match example from the WSJ **collection. Here, 42.6% of the documents in cluster c_0 belong to topic t_3. There is no obvious topical match between the different 10-word representations.**

However, it is apparent in these tables that the 'central keywords' method is a complete failure. The parameter choice of 10 documents could be the explanation, but we judge it more likely that the hypothesis that there are documents that are representative of clusters may be misfounded. We discuss this further below.

Some cluster–topic matches have a low match percentage. For example, Table 2 shows the match between cluster c_0 and topic c_3 from WSJ, where 42.6% documents in cluster c_0 are labelled as topic t_3. There isn't a clear topical relationship amongst these words. Cluster c_0 is the largest in WSJ and this result suggests that the mix of topics could be an indicator that, in this case, the clustering has been unsuccessful. That is, a direction to explore is whether the topic model is diagnostic of clustering effectiveness.

We have presented results as a listing of words, conveying alignment from a perspective that allow direct human interpretation. A quantitative alternative to would be to compute the vector similarities between the full representations. We have not yet taken that step, but plan to do so in the next stage. In principle it would allow us to systematically explore the extent of the similarities, and, for example, to see the spread of similarities over a sequence of runs (each with different random seeds). Our focus in this stage was on the relative descriptive value of the different approaches, and our results clearly show the strong descriptive similarities.

Discussion. We have observed that the majority of clusters are dominated by a single topic and that cluster-based word lists often align closely with topic-based word lists. Our interpretation is that both are somewhat effective in gathering information by theme, and, moreover, it does appear that there are true underlying themes that can be meaningfully gathered.

However, clustering and topic modelling are not equivalent. Topic modelling is arguably richer, in that it provides a multifaceted description of each document; here we have focused on the dominant topic in each document, but other topics can be significant. The obverse of this observation is that some topics can assist in description of many documents but never be dominant, which is a common feature of topic models. As a consequence – for the purposes of comparing clustering and topic modelling, at least – the effective number of topics is reduced. Another perspective on this observation is that clustering can create subcategories of documents that share a dominant topic, while topic modelling supplements clustering by identifying subtopics.

Some clusters are a mix of topics, with no one topic well represented, and large clusters are less likely to have a strongly dominant topic than are small clusters. As noted above, this suggests that topic modelling can be used to test the cohesion of a cluster.

The 'central keywords' method for generating textual descriptions was not successful. In our view the likely explanation is that it is based on a false assumption, namely that clusters resemble a dot-cloud. The structure of high-dimensional spaces can be counter-intuitive [32]; for example, projections through a space constructed as 'distance to a specific object' can resemble donuts, in which all the items are in a ring roughly equidistant from the centre. Two objects can simultaneously be (relatively) close to a third object and remote from each other.

The fundamental confound is the cluster shape. Intuition (and dot-clouds) suggest that clusters consist of neighbours, but in a high-dimensional space it may be more meaningful to describe them as best fit. For example, a short document can lack the richer vocabulary of a longer document and as a consequence be a long way from the centroid, due to length normalisation, but is allocated to the cluster because of it is even further from other centroids. Other documents can get allocated to the same cluster due to proximity in other dimensions (sharing of other vocabulary) with the centroid. Documents with many rare words can end up close to the centroid but are therefore not similar to other words in the cluster. In general, where the feature dimensionality exceeds k there is latitude (degrees of freedom) for irregularity in shape that is not present at low dimensionality; this freedom grows as the number of dimensions is increased.

That is, the set of documents in a cluster that is close to its centroid may be somewhat arbitrary. In the experiments reported in Table 3, it can be seen that the cluster keywords largely correspond to topic keywords, while the central keywords do not correspond with cluster or topic keywords; or, indeed, with each other. This strongly suggests that the clusters have irregular shape – another feature that is not easy to grasp in two or three dimensions – so that the documents close to the centroid do not necessarily share features that are common in the cluster.

5 CONCLUSION

We have conducted a case study of the relationship between document clustering and topic modelling. In this preliminary investigation, undertaken to establish whether a deeper program of work was justified, our intention was to explore the contrasting descriptions they provide of document collections.

The expectation was that topic modelling and clustering would be very distinct from each other, given that they are constructed through different principles and they examine documents in very different ways.

However, our working hypothesis was not confirmed. Instead, on this collection topic models and clusters showed a remarkable degree of alignment. Most clusters are largely composed of documents with the same dominant topic, and topic keyword lists and cluster keyword lists are often near-identical. That is, both directions – transforming topics to clusters, and transforming clusters to word lists – yield results that are substantially aligned. To our knowledge, there has been no previous comparison of topic modelling and clustering in this mutual way.

We examined two elementary methods for cluster labelling, of which one – simply, the highest-weighted words in each centroid – appears to have been highly effective. Another outcome was that the results show that topics are useful for illuminating the structure of clusters, and vice versa, and suggest that there are limits to claims on how distinct different topics or clusters are from each other. Surprisingly, what is arguably the simplest cluster labelling method in the literature showed good similarity to topic modelling. This suggests, for example, that topic modelling might provide an automatic guide to the quality of cluster labelling.

The work reported here suggests several next steps. One is the obvious generalisation from a case study to a more comprehensive study, in which the alignments and levels of mutual support are explored in quantitative terms, a wider range of document types is considered, and so on. Another is to explore how these techniques can be used to inform each other and be mutually strengthened. A third is to examine methods for exploiting the representations they generate to support user navigation during search. However, our results have already established a link between the methods that had not previously been identified, a link that gives confidence in the robustness of both clustering and topic modelling as tools for analysis of document collections.

REFERENCES

[1] A. Aker, E. Kurtic, A. R. Balamurali, M. Paramita, E. Barker, M. Hepple, and R. Gaizauskas. 2016. A graph-based approach to topic clustering for online comments to news. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 15–29. https://doi.org/10.1007/978-3-319-30671-1_2

[2] K. K. Bharti and P. K. Singh. 2015. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications* 42, 6 (2015), 3105–3114. https://doi.org/10.1016/j.eswa.2014.11.038

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. https://doi.org/10.5555/944919.944937

[4] A. Broder, L. Garcia-Pueyo, V. Josifovski, S. Vassilvitskii, and S. Venkatesan. 2014. Scalable k-means by ranked retrieval. In *Proc. ACM Int. Conf. on Web Search and Data Mining* (New York, New York, USA). Association for Computing Machinery, New York, NY, USA, 233–242. https://doi.org/10.1145/2556195.2556260

[5] D. Carmel, H. Roitman, and N. Zwerdling. 2009. Enhancing cluster labeling using Wikipedia. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/1571941.1571967

[6] H. Chan and L. Akoglu. 2013. External evaluation of topic models: A graph mining approach. In *Proc. IEEE Int. Conf. on Data Mining*. 973–978. https://doi.org/10.1109/ICDM.2013.112

[7] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proc. NIPS Int. Conf. on Neural Information Processing Systems* (Vancouver, British Columbia, Canada). Curran Associates Inc., Red Hook, NY, USA, 288–296. https://doi.org/10.5555/2984093.2984126

[8] G. Cleuziou. 2008. An extended version of the k-means method for overlapping clustering. In *Int. Conf. on Pattern Recognition*. IEEE, Tampa, FL, USA, 1–4. https://doi.org/10.1109/ICPR.2008.4761079

[9] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management* 57, 2 (2020), 102034. https://doi.org/10.1016/j.ipm.2019.04.002

[10] B. C. Fung, K. Wang, and M. Ester. 2003. Hierarchical document clustering using frequent itemsets. In *Proc. SIAM Int. Conf. on Data Mining*. SIAM, SIAM, San Francisco, CA, 59–70. https://doi.org/10.1137/1.9781611972733.6

[11] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen. 2008. Enhancing text clustering by leveraging Wikipedia semantics. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*. 179–186. https://doi.org/10.1145/1390334.1390367

[12] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. 2013. Unsupervised graph-based topic labelling using Dbpedia. In *Proc. ACM Int. Conf. on Web Search and Data Mining* (Rome, Italy) *(WSDM '13)*. Association for Computing Machinery, New York, NY, USA, 465–474. https://doi.org/10.1145/2433396.2433454

[13] A. K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letter* 31, 8 (June 2010), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[14] N. Jardine and C. J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7, 5 (1971), 217–240. https://doi.org/10.1016/0020-0271(71)90051-9

[15] A. Kulkarni and J. Callan. 2010. Document allocation policies for selective searching of distributed indexes. In *Proc. CIKM Int. Conf. on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 449–458. https://doi.org/10.1145/1871437.1871497

[16] A. Kulkarni, A. S. Tigelaar, D. Hiemstra, and J. Callan. 2012. Shard ranking and cutoff estimation for topically partitioned collections. In *Proc. CIKM Int. Conf. on Information and Knowledge Management* (Maui, Hawaii, USA). Association for Computing Machinery, New York, NY, USA, 555–564. https://doi.org/10.1145/2396761.2396833

[17] K. Kummamuru, A. Dhawale, and R. Krishnapuram. 2003. Fuzzy co-clustering of documents and keywords. In *IEEE Int. Conf. on Fuzzy Systems*, Vol. 2. 772–777 vol.2. https://doi.org/10.1109/FUZZ.2003.1206527

[18] J. H. Lau, D. Newman, and T. Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proc. EACL Conf. of the European Chapter of the Association for Computational Linguistics*. EACL. https://doi.org/10.3115/v1/E14-1056

[19] X. Liu and W. B. Croft. 2004. Cluster-based retrieval using language models. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval* (Sheffield, United Kingdom). Association for Computing Machinery, New York, NY, USA, 186–193. https://doi.org/10.1145/1008992.1009026

[20] L. A. Lopes, V. P. Machado, and R. de A. L. Rabêlo. 2014. Automatic cluster labeling through artificial neural networks. In *International Joint Conference on Neural Networks (IJCNN)*. 762–769. https://doi.org/10.1109/IJCNN.2014.6889949

[21] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian. 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine* 117 (2021), 102096. https://doi.org/10.1016/j.artmed.2021.102096

[22] Y. Lu, Q. Mei, and C. Zhai. 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval* 14, 2 (2011), 178–203. https://doi.org/10.1007/S10791-010-9141-9

[23] E. L. Lydia, P. K. Kumar, K. Shankar, S. K. Lakshmanaprabu, R. M. Vidhyavathi, and A. Maseleno. 2018. Charismatic document clustering through novel k-means non-negative matrix factorization (KNMF) algorithm using key phrase extraction. *Int. Journal of Parallel Programming* 48, 3 (Jul 2018), 496–514. https://doi.org/10.1007/s10766-018-0591-9

[24] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*.

[25] D. Newman, S. Karimi, and L. Cavedon. 2009. External evaluation of topic models. In *Australasian Document Computing Symposium*. 11–18.

[26] H. Poostchi and M. Piccardi. 2018. Cluster labeling by word embeddings and WordNet's hypernymy. In *Proceedings of the Australasian Language Technology Association Workshop 2018*. Dunedin, New Zealand, 66–70.

[27] R. L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (1953), 267–276. https://doi.org/10.1007/BF02289263

[28] A. Tombros, R. Villa, and C. J. van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management* 38, 4 (2002), 559–582. https://doi.org/10.1016/S0306-4573(01)00048-6

[29] C. J. van Rijsbergen. 1979. *Information Retrieval, 2nd ed.* Butterworths.

[30] E. M. Voorhees. 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management* 22, 6 (Dec. 1986), 465–476. https://doi.org/10.1016/0306-4573(86)90097-X

[31] E. M. Voorhees and D. K. Harman (Eds.). 2005. *TREC experiment and evaluation in information retrieval*. The MIT Press.

[32] R. Weber, H. Schek, and S. Blott. 1998. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proc. VLDB Int. Conf. on Very Large Databases*, Vol. 98. 194–205. https://doi.org/10.5555/645924.671192

[33] P. Xie and E. P. Xing. 2013. Integrating document clustering and topic modeling. In *Proc. UI Conf. on Uncertainty in Artificial Intelligence*. Bellevue, Washington, USA, 694–703. https://doi.org/10.5555/3023638.3023709

[34] W. Xu, X. Liu, and Y. Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*. 267–273. https://doi.org/10.1145/860435.860485

Match	Cluster	Topic	Centre
c_3-t_4	quarter, cent, net, share, earn, loss, sale, revenu, profit, incom	quarter, cent, share, net, earn, loss, sale, profit, revenu, incom	laser, pnc, hansen, doctrin, billion, mile, mercer, quarter, daughter, court
c_4-t_{14}	contract, navi, aircraft, air, armi, forc, missil, receiv, equip, award	contract, navi, aircraft, air, armi, lockhe, receiv, forc, missil, corp	vista, louisiana, oil, bibl, pennzoil, opec, catalyst, g, canada, texaco
c_8-t_{13}	share, common, stock, outstand, offer, tender, sharehold, cent, exchang, file	share, common, offer, stock, outstand, sharehold, tender, holder, stake, file	taft, itt, home, arrest, haa, borden, adhes, mr, night, crew
c_{10}-t_{17}	presid, vice, name, chief, succeed, execut, offic, director, chairman, mr	presid, vice, name, chief, execut, succeed, director, offic, chairman, elect	burlington, quarter, ton, mcorp, complaint, billion, jefferi, newsprint, sec, ministri
c_{11}-t_1*	bond, yield, rate, treasuri, mortgag, issu, due, price, rat, point	bond, bank, yield, fund, loan, mortgag, rate, thrift, rat, treasuri	midland, index, hugh, com, tripl, bank, ashland, expir, loan, market
c_{14}-t_9	dollar, yen, currenc, mark, ounc, trader, trade, gold, comex, bank	index, stock, market, trade, dollar, trader, price, yen, rise, share	cbs, clark, ibm, tisch, media, execut, union, rubicam, aetna, intend
c_{15}-t_6*	soviet, mr, democrat, reagan, parti, iran, bush, polit, contra, arm	hous, soviet, govern, senat, democrat, polit, administr, drug, would, congress	templ, plymouth, pacif, light, auditorium, estat, packer, march, real, apart
c_{16}-t_3*	mr, peopl, one, drug, work, get, like, time, go, would	mr, busi, bush, one, time, get, peopl, go, like, work	mile, lynn, tyson, eleven, passeng, ec, cftc, warrant, accid, franchis
c_{17}-t_{16}	mile, passeng, load, traffic, revenu, factor, fli, airlin, seat, percentag	airlin, passeng, mile, traffic, carrier, air, load, northwest, revenu, flight	deaver, cabl, dixon, linen, telecom, polic, acreag, acr, welfar, extend
c_{18}-t_7*	statist, rise, earlier, ton, adjust, season, price, unemploy, consum, increas	billion, rise, statist, earlier, juli, june, export, adjust, month, surplus	centel, oil, wast, tax, packwood, parti, guest, suit, minimum, revenu

(a) WSJ

Match	Cluster	Topic	Central
c_0-t_{16}	dollar, yen, currenc, ounc, mark, comex, gold, trader, london, bank	dollar, yen, currenc, fed, mark, ounc, grumman, bank, trader, gold	fire, bayer, german, walker, sec, soviet, accid, student, economist, lawsuit
c_1-t_3	soviet, iran, gorbachev, missil, moscow, arm, militari, reagan, israel, iranian	soviet, isra, israel, militari, missil, iran, communist, east, gorbachev, bomber	guin, pan, caesar, jersey, asbesto, genet, bond, japan, bank, expansionari
c_5-t_{11}*	airlin, air, flight, pilot, aircraft, eastern, boe, carrier, pan, fare	bank, airlin, thrift, loan, insur, execut, firm, amp, offic, manag	disabl, patient, hoechst, upjohn, noranda, africa, herrington, deplet, aid, drug
c_6-t_{19}*	democrat, senat, hous, bush, reagan, bill, dukaki, republican, congress, sen	bush, tax, hous, senat, democrat, bill, polit, rep, congress, committe	ge, utah, drug, twa, chrysler, baker, kidder, renault, miller, union
c_9-t_1	comput, ibm, softwar, machin, appl, chip, digit, system, person, microsoft	ibm, comput, japan, japanes, bank, softwar, european, machin, market, franc	sec, beer, southland, manvill, vw, gm, nakason, german, japanes, tax
c_{12}-t_7	via, bond, due, yield, common, par, offer, amp, share, note	via, bond, due, yield, par, seri, offer, note, amp, pierc	gencorp, tpa, fleet, partnership, barbecu, aid, multifamili, israel, isra, children
c_{13}-t_6*	opec, oil, price, futur, barrel, soybean, contract, trader, cent, crude	oil, price, ton, gas, moodi, natwest, soybean, exxon, crude, ounc	taft, fire, symbol, helicopt, lewi, jefferi, merc, shop, mcdonnel, moodi
c_{15}-t_4	peopl, school, work, black, student, ms, get, women, polit, like	peopl, ad, ms, get, advertis, famili, film, women, show, like	oil, futur, dome, exchang, fujitsu, harper, cftc, price, fairchild, japanes
c_{16}-t_8	quarter, cent, net, share, earn, loss, sale, revenu, profit, earlier	quarter, cent, share, net, earn, loss, sale, profit, revenu, earlier	contra, prudenti, insur, amtrak, judg, giuliani, iran, intel, nec, reagan
c_{17}-t_2*	bond, yield, rate, treasuri, mortgag, issu, rat, price, day, point	bond, stock, yield, index, rate, market, investor, fund, mortgag, day	jefferi, faa, baker, condom, store, fork, bank, ladi, impeach, ir

(b) WSJ-LONG

Match	Cluster	Topic	Central
c_0-t_2*	debentur, due, bond, note, redeem, convert, redempt, offer, proceed, subordin	dividend, share, offer, debentur, secur, common, note, stock, payabl, via	mile, passeng, kaneb, loss, cablevis, tuesday, alaska, percentag, air, load
c_2-t_{11}*	rise, statist, earlier, price, year, sale, adjust, consum, season, month	billion, rise, franc, statist, earlier, year, month, price, adjust, juli	goldom, reno, tobacco, explor, reynold, singapor, packag, broadcast, offer, oil
c_3-t_{18}	quarter, million, net, cent, loss, share, earn, incom, profit, sale	quarter, million, cent, net, revenu, loss, rise, earn, year, sale	porsch, univers, debentur, mile, elder, aim, nec, share, chi, passeng
c_4-t_0	contract, navi, air, aircraft, forc, armi, missil, receiv, million, award	contract, navi, ton, aircraft, mine, air, forc, armi, award, million	directori, fairchild, abort, elliott, februari, crosbi, powel, singapor, execut, child
c_5-t_{15}	ton, week, steel, capabl, metric, output, iron, earlier, product, year	steel, ton, order, week, italian, capabl, aircraft, boe, dougla, produc	ex, develop, shell, zenith, budget, passeng, mile, million, merger, edward
c_8-t_8	mile, passeng, load, revenu, traffic, factor, fli, one, airlin, percentag	repurchas, model, bureau, car, australian, environment, truck, motor, inflat, plant	cyanamid, maryland, offic, hi, quebec, soviet, genet, institut, hydro, german
c_9-t_{10}*	chief, offic, execut, presid, mr, chairman, succeed, name, old, vice	presid, vice, name, chief, execut, director, mr, offic, chairman, succeed	bank, alcan, bill, mccarthi, breweri, walgreen, atkinson, mr, indonesia, via
c_{10}-t_7*	file, share, stake, group, common, invest, commiss, hold, secur, exchang	share, stock, common, offer, buy, compani, outstand, group, hold, sharehold	contract, ncr, januari, statist, februari, singapor, navi, first, mile, cosmet
c_{15}-t_1*	bank, loan, save, asset, feder, first, approv, million, hold, bancorp	bank, loan, asset, save, agreement, plant, unit, million, feder, complet	split, emerson, approv, loral, ameritrust, gibraltar, passeng, stock, royal, dividend
c_{17}-t_{17}	seat, quot, bid, exchang, membership, sell, current, ask, mercantil, sale	sale, pretax, yen, billion, mark, seat, bankruptci, profit, chapter, creditor	rock, ask, via, xtra, yanke, trailer, membership, share, educ, bid

(c) WSJ-SHORT

Table 3: WSJ **collections, showing top keywords for each method. Each table shows the 10 best topic–cluster matches in each collection, with c_i-t_j indicating that the best matching topic for cluster c_i is topic t_j. A star, as in c_i-t_j*, indicates that topic t_j is also the best aligned topic for other clusters that are not shown.**