Clustering Documents in Topic Space: Worth the Effort?

A Fair Comparison with Established Vector-Space Model Techniques

Matthew Hawthorn
mah6wt@virginia.edu

Shannon Mitchell
som3dq@virginia.edu

Nikhil Mascarenas
nm4gt@virginia.edu

Keywords

document clustering; vector space models; topic models

1. BACKGROUND AND MOTIVATION

Document clustering is an important method used widely in information retrieval. It is also important in other fields where document labels are not available such as identifying new trends within research studies, forensic computing and innovations within patent applications. Hierarchical clustering methods can typically obtain better performance than simpler methods such as k-means, however hierarchical methods can be prohibitively expensive on large datasets in terms of computational speed. Documents are often represented with a vector space model using a term frequency-inverse document frequency (tf-idf) embedding, but the more recent approach of latent topic modelling, best exemplified by Latent Dirichlet Allocation (LDA)[1], can potentially embed documents in a much lower-dimensional space of latent 'topics', groupings of frequently co-occurring terms. This speeds up document similarity computations, resulting for instance in faster queries in document retrieval, or faster clustering once the model is computed. We suspect topic models can improve accuracy of similarity measures as well; especially when documents are short, two highly semantically similar documents might nonetheless share few terms (yielding a low cosine similarity), whereas they are more likely to overlap in topics (in the topic model sense). In this way we hypothesize that for many corpora, clustering on topic weights might yield better results than clustering on term frequencies.

2. PREVIOUS WORK

To our knowledge, relatively little work has been done on evaluating the utility of topic models for document clustering; the vast majority of studies of document clustering have investigated it in the context of vector-space models.

In a comprehensive and authoritative survey, Zhao and Karypis (2002)[6] compare 9 different agglomerative and 6 partitional hierarchical clustering techniques using 12 text datasets including the Reuters-21578 news corpus, representing documents in a traditional vector space model using tf-idf embedding. As partitional clustering algorithms such as k-means are often selected due to lower computational requirements, they wanted to confirm whether there was a tradeoff with model performance. The results surprisingly showed that more expensive agglomerative clustering algorithms were outperformed by a hierarchical divisive bisecting k-means algorithm, measuring by cluster agreement with labelled classes using an F-score criterion averaged over all levels of the resulting tree.

Xie and Xing (2013)[5] propose a new document clustering method integrating topic modelling and clustering into a single probabilistic graphical model with topics and clusters as latent variables. They jointly infer topics and clusters using this model, and then compare the accuracy of the resulting clusters against several other clustering algorithms. Among these other techniques are traditional k-means in normalized tf-idf space, k-means performed on LDA topic weight vectors, and "LDA+naive" which denotes the treatment of topics as clusters and assigning documents to their highest-weight topics. They find that their joint probabilistic model, the "Multi-grain Clustering Topic Model" (MGCTM), outperforms the other methods they tried, measuring by normalized mutual information (NMI) on the 20-Newsgroups and Reuters-21578 datasets, though only by a slight margin over the "LDA+naive" approach.

In contrast to the more complex joint inference approach of Xie and Xing, Ma et. al. (2014)[4] investigate the use of traditional clustering techniques on embeddings of documents in a standard LDA topic space. They employ a three-stage process: 1) identify an optimal number of topics according to a "topic significance degree" measure, 2) seed initial cluster centers using the k-means++ algorithm, 3) carry out k-means clustering in the usual manner, but using Jensen-Shannon divergence as a distance measure rather than euclidean distance. Compared against the same procedure without tuning the number of topics, they find some improvement on their two datasets, which include 20-Newsgroups.

3. RESEARCH PROBLEM

The added utility of topic models over simpler vector-space models for document classification and clustering in the general case is still largely open to question; papers we have found discussing one technique or the other use either different datasets or different evaluation metrics and are therefore not directly comparable. Xie and Xing make a direct comparison between their complex model and simpler approaches, but implement the simpler approaches naively, not giving a fair comparison. This was exacerbated by the fact that accuracy margins for their model over the simplest LDA approach were narrow (50.1% vs. 48.0% NMI and 56.6% vs. 54.9% accuracy on Reuters-21758). We propose to make a fuller comparison. What we want to know is: can we do much better clustering documents in topic space than we can by using the state of the art techniques in traditional tf-idf vector space? Is the computational cost worth it?

4. TECHNIQUES

We will explore document clustering techniques applied to topic embeddings, and compare them with clusterings using tf-idf embeddings. With large data sets in mind, we rule out agglomerative techniques as too computationally expensive, and we would like to use techniques which have already been efficiently implemented for large data. For instance, Apache Mahout already has good implementations of Latent Dirichlet Allocation and k-means, whereas the complex joint inference task of Xie and Xing would require a good deal of development to be deployed at scale, and may or may not prove to be scaleable. This leaves us with partitional techniques on LDA topic embeddings. Specifically we will try the bisectional k-means found by Zhao and Karypis to be optimal for tf-idf vectors. The hierarchical partitional technique gives us an opportunity to optimize the cluster count via cluster coherence measures as well, rather than artificially evaluating on the known number of categories.

The choice of metric is important in clustering, especially in high-dimensional spaces. Euclidean distance, for instance (which Xie & Xing use in k-means on LDA topic space for comparison) is an inappropriate metric on the simplex where topic distributions reside. Lebanon (2006)[3] found that learning an optimal metric on the simplex from term distributions resulted in nearly half the error rate of tf-idf cosine in a binary text classification task. We believe that for probability distributions, information-theoretic distance metrics have better theoretical support than usual metrics such as L2 norm (euclidean distance) or L1 norm. Thus, in agreement with Ma et. al., we will implement k-means using the Jensen-Shannon divergence (JS). We believe this is well-founded since the JS divergence is the square of a metric, and just as with squared euclidean distance, the sum of JS divergences from a centroid is minimized at the mean distribution, as demonstrated by Fuglede & Topsoe (2004)[2]. A comparison of our proposal to previous work is summarized in Appendix A.

5. EVALUATION

We will compare our clusterings of documents in topic space against the best hierarchical algorithm from Zhao and Karypis in tf-idf space on the following datasets:

- 20-Newsgroups
- Reuters-21578

We will evaluate our clusterings using the following metrics:

- F-score as implemented by Zhao and Karypis for hierarchical clusterings
- F-score as implemented by Ma et. al. for flat (k-means) clusterings.
- Normalized mutual information, which Xie and Xing use to evaluate their clusterings.

In this way, we will achieve a fair comparison of different document clustering techniques while bringing to bear the best available techniques in both the vector space and topic model representations. We will also get a fair comparison with the more complex joint inference technique of Xie and Xing.

6. EXPECTED OUTCOME

We expect to find a clear winner, and to finally be able to answer the question of which document embedding can be expected to perform best for clustering tasks on datasets which are comparable to our test sets. If it is the vector-space model, then there seems little utility in the topic model computation for the purposes of clustering. If it is the topic model representation, then it becomes a relevant question in an application context whether the computational cost is worth the improved accuracy. If our technique beats or matches the more complex and potentially less scalable joint inference technique of Xie and Xing, then there seems little utility in developing that technique for larger scale applications.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] B. Fuglede and F. Topsoe. Jensen-shannon divergence and hilbert space embedding. In *IEEE International Symposium on Information Theory*, pages 31–31, 2004.

[3] G. Lebanon. Metric learning for text documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):497–508, 2006.

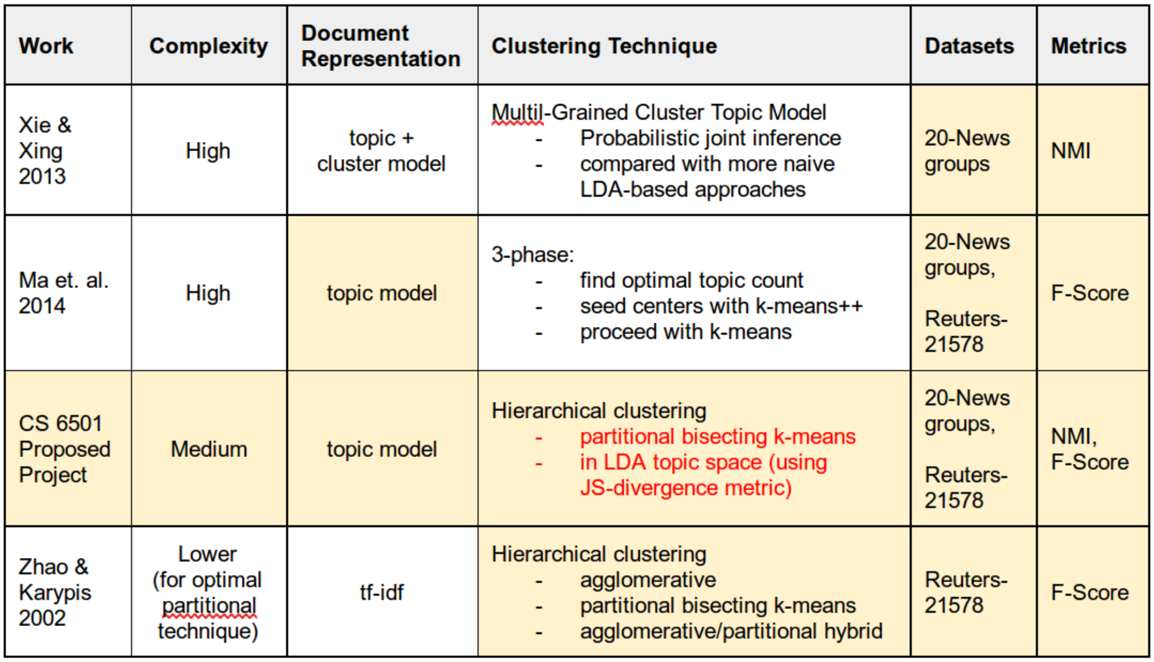
[4] Y. Ma, Y. Wang, and B. Jin. A three-phase approach to document clustering based on topic significance degree. *Expert Systems with Applications*, 41(18):8203–8210, 2014.

[5] P. Xie and E. P. Xing. Integrating document clustering and topic modeling. *arXiv preprint arXiv:1309.6874*, 2013.

[6] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524. ACM, 2002.

APPENDIX

A. COMPARISON WITH PRIOR WORK



Work	Complexity	Document Representation	Clustering Technique	Datasets	Metrics
Xie & Xing 2013	High	topic + cluster model	Multil-Grained Cluster Topic Model - Probabilistic joint inference - compared with more naive LDA-based approaches	20-News groups	NMI
Ma et. al. 2014	High	topic model	3-phase: - find optimal topic count - seed centers with k-means++ - proceed with k-means	20-News groups, Reuters-21578	F-Score
CS 6501 Proposed Project	Medium	topic model	Hierarchical clustering - partitional bisecting k-means - in LDA topic space (using JS-divergence metric)	20-News groups, Reuters-21578	NMI, F-Score
Zhao & Karypis 2002	Lower (for optimal partitional technique)	tf-idf	Hierarchical clustering - agglomerative - partitional bisecting k-means - agglomerative/partitional hybrid	Reuters-21578	F-Score

Figure 1: A comparison of our proposed project with prior research. Overlaps in document representation, clustering technique, datasets, and evaluation metrics are highlighted. Some information is omitted for clarity.