A Survey of Document Clustering Techniques & Comparison of LDA and moVMF

Yu Xiao

December 10, 2010

Abstract

This course project is mainly an overview of some widely used document clustering techniques. We begin with the basic vector space model, through its evolution, and extend to other more elaborate and statistically sound models. We compare two models in detail, the mixture of Von Mises-Fisher and Latent Dirichlet Allocation, since they have drawn wide attention in recent years due to their good performance over other models. Finally, we propose that more experiments need carrying out over multiple topic documents (or other objects).

Keywords: VSM, LSA, pLSA, K-means, Hierarchical Clustering, LDA, moVMF, Spherical Admixture Model.

1 Background

Nowadays the information on the internet is exploding exponentially through time, and approximately 80% are stored in the form of text. So text mining has been a very hot topic. One particular research area is document clustering, which is a major topic in the Information Retrieval community. And it has found broad applications in real world. Examples include search engines. Typically, a search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by Enterprise Search engines such as Northern Light and Vivisimo or open source software such as Carrot2. Also, Google is known to use clustering methods to match certain websites with a query, since a website can be viewed as a collection of topics (multi-topic document), and a query itself is a topic or a combination of several topics. This has been studied extensively by the Search Engine Optimization community, as to find a way to optimize a website, determine the optimal bids on certain keywords, and thus improve the ROI of online campaigns. Finally, with the rising of social network in recent years, such as Facebook and Twitter, more semantic data are available now which convey considerable amount of information. Take Twitter as an example. There are approximately 95M tweets per day, which is equivalent to 1100 tweets per second. Researchers from the Northeastern University College of Computer and Information Sciences and Harvard Medical School have developed an innovative way of tracking the nation's mood using tweets. And another two CMU researchers find Twitter posts in line with opinion polls. All these researches show the power of social computing in providing accurate assessments on many sorts of issues, at almost no cost and on a large scale. Likewise, document clustering techniques can be used to group tweets into relevant topics, in aid of the current mere 'Trends' function used by Twitter. For all these reasons, we find document clustering techniques valuable and therefore worth studying.

The remainder of this paper is organized as follows. In section 2 we introduce the basic vector space model and address its limitations. In section 3 we introduce some dimensionality reduction techniques. In section 4 we compare two attractive models, LDA and moVMF, over multiple-topic documents.

2 The Vector Space Model (VSM)

2.1 Document Preprocessing

Before we represent documents as tf-idf vectors, we need some preprocessing. There are commonly two steps:

- First, we need to remove stop words, such as 'a', 'any', 'what', 'I', etc, since they are frequent and carry no information. A stop words list can be found online.
- Second, we need to stem the word to its origin, which means we only consider the root form of words. For example, ran, running, runs are all stemmed to run, and happy, happiness, happily are all stemmed to happi. There are certain criteria, and the standard algorithm is the Porter's stemmer, which is also free online. A more elaborate way of stemming is by using the WordNet, which in addition to suffix-stripping also groups words into synsets, and leads to an ontology-based (instead of word-based) document clustering method. Related work can be found in [15].

2.2 tf-idf Matrix

The Vector Space Model is the basic model for document clustering, upon which many modified models are based. We briefly review a few essential topics to provide a sufficient background for understanding document clustering.

In this model, each document, d_j, is first represented as a term-frequency vector in the term-space:

$$\boldsymbol{d}_{j_{tf}} = (tf_{1j}, tf_{2j}, ..., tf_{Vj})' \qquad j = 1, 2, ..., D \tag{1}$$

where tf_{ij} is the frequency of the i^{th} term in document d_j, V is the total number of the selected vocabulary, and D is the total number of documents in the collection.

Next, we weight each term based on its inverse document frequency (IDF). The basic idea is that if a term appears frequently across all documents in a collection, its discriminating power should be discounted. So finally, we obtain a tf-idf vector for each document:

$$\boldsymbol{d}_j = (tf_{1j} \times idf_1, tf_{2j} \times idf_2, ..., tf_{Vj} \times idf_V)' \qquad j = 1, 2, ..., D \tag{2}$$

Put these tf-idf vectors together, we get a tf-idf matrix:

$$(tf - idf)_{ij} = tf_{ij} \times idf_i \qquad i = 1, 2, ..., V; \qquad j = 1, 2, ..., D \tag{3}$$

2.3 Similarity Measure and Clustering Algorithm

The most commonly chosen measure is the cosine similarity. As mentioned in [2], the choice of a similarity measure can be crucial to the performance of a clustering procedure. And the Euclidean or Mahalanobis distance behave poorly in some circumstances, while the cosine similarity captures the 'directional' characteristics which is intrinsic of the document tf-idf vectors. In addition, the cosine similarity is exactly the Pearson correlation, which gives it a sound statistical stance.

There are two typical categories of clustering algorithms, the partitioning and the agglomerative. K-means and the hierarchical clustering are the representatives of these two categories, respectively. There are many comparisons between K-means and hierarchical clustering. But our consideration is speed, since we are going to apply clustering algorithms on big social network data, which is always of GB or TB size. (The ultimate goal is to build a Map-Reduce procedure to parallelize these clustering algorithms, but this goal may not be reached in this project due to time limit). And the hierarchical clustering is extremely computational expansive as the size of data increases, since it needs to compute the $D \times D$ similarity matrix, and merges small clusters each time using certain link functions. In contrast, K-means is much faster. It is an iterative algorithm, which updates the cluster centroids (with normalization) each iteration and re-allocates each document to its nearest centroid. A comparison of K-means and hierarchical clustering algorithms can be found in [16].

2.4 The Baseline and its Problems

Document clustering by using the K-means algorithm with cosine similarity (*spkmeans*) to the full space VSM has been considered as a baseline when doing performance comparison. This algorithm is straightforward and easy to implement, but it also suffers from high computational cost, which makes it less appealing when dealing with a large collection of documents. The problem is the curse of dimensionality: V is large. Generally, there are more than thousands of words in a vocabulary, which makes the term space high dimensional. Hence, various dimensionality reduction techniques have been developed to make improvements above the baseline.

3 Dimensionality Reduction Techniques

As to our knowledge, there are two main categories of dimensionality reduction techniques. One is to first write down the full VSM matrix, and then try to reduce the dimension of the term space by numerical linear algebra methods. Another is to try using a different representation of document (not as a vector in a space) from the very beginning.

3.1 Feature Selection Method - LSA

The Latent Semantic Analysis method is based on the singular value decomposition (SVD) technique in numerical linear algebra. It can capture the most variance by combinations of original rows/columns, whose number is much less than the original matrix. In addition, the combinations of rows (term) always show certain semantic relations among these terms, and combinations of columns (document) indicates certain clusters. After SVD, the K-means algorithm is run on this reduced matrix, which is much faster than on the original full matrix. But the problem with this approach is that the complexity of SVD is $O(D^3)$, so as the number of documents increases, the computation of SVD will be very expansive, and therefore the LSA approach is not suitable for large datasets.

3.2 Alternative Document Representations

There are other document representations and similarity measures besides VSM and cosine similarity, such as Tensor Space Model (TSM) and a similarity based on shared nearest neighbors (SNN), etc. These alternatives may be effective in some special cases, but not in general.

One significant step in this area is the introduction of the concept of 'latent topics'. It is similar to the latent class label in the mixture of Gaussian models. This concept has led to a series of generative models, which specify the joint distribution of the data and the hidden parameters as well. Among these models are the Probabilistic Latent Semantic Analysis (pLSA), the Mixture of Dirichlet Compound Multinomial (DCM), the Latent Dirichlet Allocation (LDA). The pLSA has a severe overfitting problem, since the number of parameters estimated in the model is linear in D, the number of documents. On the other hand, LDA specifies three levels of parameters: corpus level, document level, and word level. And the total number of parameters is fixed: $K + K \times V$, where K is number of latent topics regarded as given. This multilayer model seems more complicated than others, but it turns out to be very powerful when modeling multiple-topic documents.

Another model that attracts my attention is the Mixture of von Mises-Fisher (moVMF) model proposed by Banerjee, et al [2]. The vMF distribution is one of the simplest parametric distributions for directional data, and has properties analogous to those of the multi-variate Gaussian distribution for data in R^d. It has a close relationship with K-means + cosine similarity (*spkmeans*), while its computation via an EM procedure can be severely reduced when adopting a hard decision (hard-moVMF). When adopting a soft decision (soft-moVMF), it shows an annealing characteristic, and performs much better than hard-moVMF.

4 Comparison of LDA and moVMF

4.1 Related Experiments

Due to the appealing properties of LDA and moVMF, comparisons have been done by many researchers. Related work can be found in [1], [14]. The conclusions from these studies are listed below:

- While LDA is good at finding word-level topics, vMF is more effective and efficient at finding document-level clusters.
- Generative models based on vMF distributions are a better match for text than multinomial models.

4.2 Further Analysis

It seems that moVMF performs better than LDA according to several indicators. But here we propose a question: what if each document consists of multiple topics, just as website mentioned at the very beginning, rather than of only one topic? In this circumstance, will LDA perform better than moVMF? Actually, one can use the soft-moVMF to address the multiple-topic issue. Given an instance, the soft-moVMF assigns posterior probabilities to each topic, and therefore performs a fuzzy clustering. Each document can be viewed as a weighted combination of topics. But with these probabilities, how can one further group them into clusters? Another question is that although people can represent a document in this way, but it is not intuitive, since the model assumes each document has only one topic as the initial setup. Furthermore, it is computational expansive to use the soft-moVMF, and thus making it less attractive.

By contrast, LDA is naturally created to model multiple-topic documents (but not restricted to this). The multiple-topic feature is rooted in the model itself through the word-level topic structure. As a result, LDA can find much more topics than moVMF. For example, with D documents, moVMF can not identify more than D topics, while such limitation does not apply to LDA. In today's internet world, this is a huge advantage, since the online information is topic-intensive. It should not be surprising that one single website could contain hundreds of topics. The power of LDA is that it can not only cluster these objects (such as documents), but also extract features (topics) from these objects. And these extracted features (topics) can be further pushed into other applications. For example, they can be used to match a query entered into a search engine. A search query is very short so it is not appropriate to view it as a document and then assign it to some predefined clusters, or view it as a cluster centroid and then find its nearest neighbors. By contrast, it is very easy to match it with a topic, since a topic is generally described by a few keywords. By using this method, one can easily match multiple-topic sources with a query by first identifying the most related topics, and then ranking these sources according to the matching score.

One thing we noticed during this project is that most of the well-know document data, such as classic4, Reuters-21578, and 20 Newsgroups, etc, are all single-topic documents. Many comparisons of LDA and moVMF have been done on these data, as in [1], [14]. This is unfair for LDA due to the reason mentioned above. So in order to make a more fair comparison, we need to use some multiple topic document data. One possible way of doing this is by mixing up these existent single topic documents.

4.3 Recent Work

Topic Models remain hot during these years, and lots of improvement has been done on LDA since the initial LDA model was proposed, such as the Hierarchical Topic Models [4], the Correlated Topic Models [3]. A recent paper introduced a kind of 'hybrid model', the Spherical Admixture Models [14]. It combines the 'directional' feature of the Von Mises-Fisher distribution which has been found to often model sparse data such as text more accurately than multinomial distribution [2] [1], and the 'multiple-topic' feature of LDA. In addition, it relaxes the the data distribution of Von Mises-Fisher from the positive quadrant to the whole unit hypersphere, modeling both word frequency and word presence/absence. As a result, this model is more complicated in sense of its nested structure and multiple layer of parameters. Hence its estimation methods need some attention.

References

[1] Arindam Banerjee and Sugato Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*. SIAM, 2007.

[2] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[3] D. Blei and J. Lafferty. Correlated Topic Models. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, 18:147, 2006.

[4] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] Inderjit Dhillon, Jacob Kogan, and Charles Nicholas. Feature selection and document clustering. In Michael W. Berry, editor, *Survey of Text Mining*, pages 73–100. Springer, 2003.

[7] Imola Fodor. A survey of dimension reduction techniques, 2002.

[8] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Mller, editors, *NIPS*, pages 914–920. The MIT Press, 1999.

[9] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[10] Thomas Hofmann. Probabilistic latent semantic indexing. volume 1999 of *SIGIR ForumSpecial issue*, pages 50–57, New York, NY, 1999. ACM.

[11] Guy Lebanon. Information geometry, the embedding principle, and document classification. In *in Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, 2005.

[12] Kristina Lerman. Document clustering in reduced dimension vector space. http://www.isi.edu/ lerman/papers/Lerman99.pdf, 1999.

[13] T.P. Minka and John Lafferty. Expectation-Propagation for the Generative Aspect Model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[14] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. Spherical topic models. In Johannes Frnkranz and Thorsten Joachims, editors, *ICML*, pages 903–910. Omnipress, 2010.

[15] Steffen Staab and Andreas Hotho. Ontology-based text document clustering. In *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'03 Conference held in Zakopane*, pages 451–452, 2003.

[16] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. Technical Report 00-034, University of Minnesota, 2000.

[17] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. *MIT Press*, pages 505–512, 2002.