Overview of LSA and PLSA and their effectiveness in small corpus

Kerui Zhu

University of Illinois at Urbana-Champaign keruiz2@illinois.edu

Abstract. Latent Semantic Analysis (LSA)[2] and Probabilistic Latent Semantic Analysis (PLSA)[3] are two commonly used unsupervised topic models based on term-document information. Both models can well make use of the term occurrence in each document to learn to distinguish documents into different topics. The performance benefits from the large corpus, where great amount of redundancy in terms can provide statistically distinguishable features for the model to learn. However, in real life, human can easily separate a small set of documents into several clusters. This small corpus setting rises an interesting question: can the topic model captures important features in different topics by seeing only a few number of documents? In this article, we provide a general overview to the mechanism of the LSA and the PLSA model and reviews of recent works that analyzed the performance with different size of corpus.

1 Introduction

Topic modeling is a task that takes in a set of documents and returns a set of topic, where each topic is represented by a set of words that contain essential meaning about the topic. For example, topic "computer science" may be represented by terms "programming", "coding", "Java", etc. Each document may contain information of more than one topic and different topics may share some common terms. Therefore, topic model should try to find the features that can imply the topic of the document in an unsupervised manner.

To solve this problem, many useful models have been invented. The two commonly used models are LSA[2] and PLSA[3]. This two models try to dig the features from the term document occurrence table. Since the number of documents and the number of different terms may influence the information richness of the term document occurrence table, it is natural for us to pull out the question: how will the size of the corpus influence the performance of these two topic models.

In real life, performing topic learning over thousands of documents may not always be the case. Sometimes we just want to split some documents into a few topics. However, due to the ambiguity and variety nature of natural language, documents from different topics may share some meaningful terms and documents from the same topic may also have little overlap in terms. This brings challenge for the topic models to correctly assign topics to documents especially when the corpus is small.

2 Background

In this section, we will introduce the basic mechanism of the LSA and PLSA.

2.1 LSA Model

For the LSA model, we first collect the term occurrence table across all the documents. The term occurrence table has size M $\times$ N, where M is the number of document and N is the number of terms. Then we perform Singular Value Decomposition (SVD) to the occurrence table. The SVD analysis will decompose the table as the multiplication of three matrices. The left matrix has the size of M $\times$ M, where each row represents information for a document. The right matrix has the size of N $\times$ N, where each column represents information for a term. The middle matrix has the size of M $\times$ N and is a diagonal matrix. The singular value on the diagonal line can tell the importance of each feature in the document representation and the term representation. We select the top k singular value and the corresponding features in both left and right matrices. After that, each document and term representation has k features. Since both the document representation and term representation are in the same dimension space, we can compare their similarity using cosine similarity between vectors. That means we can do clusters between document and document, term and term, and document and term. In this way, we can treat each cluster as a topic, where documents in the cluster are considered as under the same topic and the terms in the cluster are considered as the topic terms.

2.2 PLSA Model

For the PLSA model, we assume that each document may have information from more than one topic. Each word in the document may be generated by one of the topics and each topic has a probability to be chosen. The probability of all topics in a document is fixed but unknown at the beginning. The algorithm first collects a term document table as the LSA algorithm. Then we assume that there are l topics, Z = $\{z_1, z_2, ..., z_l\}$, in total among all the documents. For each word w in document d, we have

$$P(d, w) = P(d) * P(w|d), P(w|d) = \sum_{z \in Z} P(w|z) * P(z|d)$$

Combining the two equations we have

$$P(d, w) = \sum_{z \in Z} P(w|z) * P(d|z) * P(z)$$

Since the value of $P(w|z), P(d|z) and P(z)$ are unknown, we need to follow the steps of EM algorithm to calculate them. Detailed steps can be referred at .

3 Evaluation on small corpus

[1] did an experience to compare the performance of LSA model and LDA model trained by two corpus of different size. Although it did not compare between LSA and PLSA directly, we can consider LDA as an improved version of PLSA because the difference is that in LDA we assume that the probability distribution of topic z over document d and the probability distribution of word w over topic z are not fixed but following some distribution determined by Dirichlet parameters. The datasets used were TASA and COCA. The TASA dataset contains 44k documents and 5 million terms and the COCA dataset contains 57k documents and 41 million terms. The LSA algorithm was applied to the two datasets with the same number of features. The LDA algorithm was also applied to the two datasets with similar hyper-parameter. The result was evaluated by comparing the similarity between average word vector of top 3 associated words in Nelson norm dataset and the most associated words found by the algorithm. Result showed that for both LSA and LDA algorithm, the performance on COCA dataset was better than the performance on TASA dataset. And also, the performance of LDA was better than the performance of LSA on the same dataset.

4 Conclusion

Based on the result from [1], we can infer that the performance of both LSA and PLSA would be better if trained on a larger corpus. This makes sense because larger corpus can provide more information about the term-topic relation and the result would be more robust. However, we think that there is still a gap between this experience setting and our ideal setting. Although the experience used two datasets of different size, the size of the smaller one is still too large for our assumption. We think more experience could be done on corpus with only several hundred of documents and test the performance on documents that contain words that only appear a few times in the training dataset.

References

1. Crossley, S., Dascalu, M., McNamara, D.: How important is size? an investigation of corpus size and meaning in both latent semantic analysis and latent dirichlet allocation (2017), `https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15441/14942`
2. Dumais, S.T.: Latent semantic analysis. Annual review of information science and technology **38**(1), 188–230 (2004)
3. Hofmann, T.: Probabilistic latent semantic analysis. arXiv preprint arXiv:1301.6705 (2013)