Clustering Chinese News Articles Using Latent Semantic Analysis and Latent Dirichlet Allocation

YANG XIA, Northeastern University, USA
ZEJIAN ZHANG, Northeastern University, USA

Mining from text data has been a hot topic for many years and most of the works and methodologies are based on English language. We as Chinese speakers are curious about the their effectiveness when it comes to a totally different language, i.e. Chinese. We picked topic modeling, which is a well-studied data mining problem, and tried to perform the same methods in Chinese news articles, with some language specific modifications. In this project, we used 10,000 news articles that covers 6 topics and tried to cluster them and measure the result by the purity of each cluster. We tried 3 ways to approach the problem, namely shingle analysis, LSA (Latent semantic analysis) and LDA (Latent Dirichlet allocation). We also use Jieba package to help us when parsing and cutting Chinese world. Based on the result that we've got, it's safe to say that statistical methods suitable for non-Latin languages like Chinese as long as we have a good vectorized matrix.

ACM Reference Format:
Yang Xia and Zejian Zhang. 2019. Clustering Chinese News Articles Using Latent Semantic Analysis and Latent Dirichlet Allocation. 1, 1 (January 2019), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Using computers to find out topics in books and articles has been one of the challenging topics in the field of computer science in recent decades. There has been many interesting approaches in order to solve this particular problem, by introducing many different ways to transform this problem into probabilistic models and applying mathematical methods to solve them. Though in the recent years neural network inspired methods like word2vec and BERT have achieved great success in the field of natural language processing, we focus on conventional approaches in this project.

We have learned algorithms like shingling, min-hashing, locality-sensitive hashing, spectral clustering, latent semantic analysis and latent Dirichlet allocation in the data mining course to deal with texts related problems, and we wished to use these algorithms to cluster news articles in Chinese into different categories without empirical information.

1.1 Motivation

We are interested with the idea of applying clustering approaches in data mining to Chinese news articles. First, if a clustering approach is proven to perform well on news articles, we can use it to analyze many different kinds of texts on the Internet since they share similar lengths of news articles. For example, we could adapt this approach to group articles autonomously for editors and content driven websites, help social network platforms to filter out hate speeches, dig out fake reviews on shopping websites etc. Second, the structures of texts in Chinese are very different from the ones in English, e.g. there are no blanks to separate words, same sequences of characters can have different word separations given the context and each might have contrasting meanings, generic rules for stop words are hard to define etc. As Chinese speakers, we would like to find out a clustering method which can handle with these difficulties.

1.2 Contributions

Report Contributions	
Section	Author
Abstract	Yang Xia
Introduction	Zejian Zhang
Related Works	Zejian Zhang
Background Information	Zejian Zhang
Proposed Approaches	Zejian Zhang
Experiments	Yang Xia
Conclusions and Future Work	Yang Xia
References	Zejian Zhang

2 RELATED WORKS

2.1 What's been done

There are many well-established methods in the field of topic modeling. We have learned Latent Semantic Analysis and Latent Dirichlet Allocation in the data mining course, both could be used for documents clustering after some adjustments and are easy to implement. As a generative probabilistic topic modeling approach, LDA provides a more comprehensive mathematical representation of text corpus than LSA, and we want to compare the results of these two methods on our dataset, i.e. Chinese news articles.

2.2 What needs to be done

Chinese language is a complex language and the structure of words and sentences differs significantly from Romance languages. Unlike Romance languages, Chinese words don't have blanks in between when they form a sentence, and generic word separation rules are hard to generate for computers since there are often ambiguousness in sentences without empirical knowledge in Chinese. As most researches are conducted on texts in Romance languages, these algorithms might not produce good results if we simply split Chinese texts into characters.

To solve this issue, we have adapted certain Natural Language Processing techniques in the preprocessing step, trying to obtain maximum accuracy in word separation. In addition, we maintained a stop words list to minimize the impacts of words with little semantic importance.

Authors' addresses: Yang Xia, Northeastern University, 75 st saint Alphonsus street, Boston, MA, 02120, USA, xia.yang@husky.neu.edu; Zejian Zhang, Northeastern University, 1575 Tremont Street, Boston, MA, 02120, USA, zhang.zeji@husky.neu.edu.

3 BACKGROUND INFORMATION

3.1 K-means

K-means is the algorithm we used to do clustering on the intermediate results in the two approaches we mentioned. We chose this method because of its efficiency since the original data has 100,000 dimensions.

ALGORITHM 1: K-means

Input: Input variables $X=\{x_i\}$ and number of clusters k
Output: Coordinates of k centroids and labels for each variable x_i
for *n different centroids initializations* **do**
  repeat
    Cluster each point with the center nearest to it.;
    Find the centroid of each cluster and replace the set of old centers with the centroids
  until *the centers converge*;
end
keep the set of centroids generating lowest sum of squared errors

3.2 Purity

The performances of two approaches on the test dataset are compared with external validation. We used the original labels of the input data to calculate the purity in each cluster we obtained after running the programs. In addition to purity, the variance in sizes of each cluster produced by two methods is also considered when evaluating the result.

ALGORITHM 2: Purity Computation

Input: Labels of articles assignments to clusters $L(a_i)$, mapping of article index to original catalog $c(index)$
Output: Purity of each cluster $P(c_i)$
D = empty dictionary of empty dictionaries, each sub-dictionary has 0 as the default value; P = empty dictionary;
for *each article a_i* **do**
  $D[L(a_i)][c(a_i.index)]$ += 1;
end
for *each cluster C_i in dictionary D* **do**
  $P[C_i] = \frac{\max_j D[C_i][j]}{\sum_j D[C_i][j]}$
end

3.3 Online Dirichlet Allocation Algorithm

In order to reduce the computation time of Latent Dirichlet Allocation, Hoffman, Bach, Blei[Matthew D. Hoffman 2010] came up with an online variational Bayesian algorithm for LDA, a.k.a. online LDA. This algorithm used online stochastic optimization and achieved great performance boost in producing good parameter estimates. In our project, we used this algorithm for approximate inference of the latent variables.

4 PROPOSED APPROACHES

We have worked with latent semantic analysis, latent Dirichlet allocation and spectral clustering using shingling, min-hashing and locality-sensitive hashing. However, the last one we mentioned in our proposal perform poorly on our dataset since each article is about 200 words, the performance is impacted greatly by the choice of shingling and the sequence of the words in the sentence. In this report, we focus on the first two approaches.

4.1 Latent Semantic Analysis

With the rise of large-scale data, statistical methods become popular when analyzing the relationships among terms and documents due to their efficiency and simplicity. Latent Semantic Analysis embraced this idea, it was introduced in Dumais, Furnas, Landauer, and Deerwester(1988)[Dumais 2004] to the field of information retrieval as a way to reduce dimensionality. We have seen wide applications of this method and it has been proven to be effective on certain datasets. LSA is an unsupervised learning technique and only takes the texts as input without any knowledge in human languages.

(1) Split the text corpus into words and remove the stop words.
(2) Compute the TF-IDF score for each term in documents and construct a document by term matrix.
(3) Run truncated SVD algorithm on the matrix and keep most variance in a lower dimensional space.
(4) Run K-means algorithm on both the original TF-IDF matrix and the lower dimensional matrix.
(5) Calculate the purity of each cluster.

4.2 Latent Dirichlet Allocation

With the further study on generative probabilistic models of documents, people discovered that LSA is unable to reason its model in a mathematical way. Probabilistic LSA is presented by Hoffman(1999)[Hofmann 1999], providing one way of probabilistic modeling of text. However, it fails to overcome overfitting and neglects the probabilistic modeling at the level of documents. Latent Dirichlet Allocation[David M. Blei 2003] claims to conquer these problems by providing a generative probabilistic modeling on text corpus. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

(1) Split the text corpus into words and remove the stop words, construct a bag-of-words matrix.
(2) Run Online LDA algorithm for latent variable inference for several different pairs of parameters α and β, select the result with lowest perplexity.
(3) Construct a document by topic matrix using the latent variables. Run K-means algorithm on the matrix.
(4) Calculate the purity of each cluster.

5 EXPERIMENT SECTION

In this section, we'll briefly talk about shingles matrix approach and analysis why it is not suitable in this case. And then we'll focus on LSA and LDA, list out the result that we've got, and try to explain the logic behind the it.

5.1 Experiment set up

The dataset that we used are scrawling from sina.com, you can think of it as Chinese version of BBC. We scrawled more than 10,000 news articles in 6 catalogs to obtain 1,600 to 1,800 articles for each catalog.[Education, Entertainment, Fashion, Health, Sports, Technology]. We did some preprocessing here:

(1) Remove the files that are less than 300 bytes.
(2) Hash the file name and sort them based on their original topics. e.g. file name 10001 suggests it's a sport news. But of course this information is unknown to the model.
(3) Use Jieba package to transform each article into a vector of bag of words.
(4) Filter out words with very little semantic importance using a Chinese stop words dataset with modifications.
(5) After we got the result from LDA and detected some words that does not have semantic meaning, added them to stop words lists and iterate.

We ran our codes on iMac (Retina 5K, 27-inch, Late 2015):

Processor	4.0GHz quad-core Intel Core i7
Memory	16GB (two 8GB) of 1867MHz DDR3 memory
Storage	1TB (7200-rpm) hard drive
Graphics	AMD Radeon R9 M380 graphics processor

In terms of implementations, we mainly use sklearn, the pros and cons will be discussed in the following section.

5.2 Shingle matrix and spectral clustering

The first thing that came to our mind is the shingles matrix. We tried this approach on a small dataset with around 1000 articles and construct Laplacian matrix using Jaccard similarity, and apply spectral clustering. However it does not perform well even in a small dataset. Our guess is that the exact order of words affects the result of this approach so much, we would have to tune the threshold above which 2 articles are considered in a same topic carefully. However, this is not feasible due to the fact that the shingles matrix is much larger (and sparser) than both TF-TDF matrix and word-count matrix, hence makes operating on it extremely slow on the original dataset.

5.3 Latent Semantic Analysis

After shingles we tried the LSA approach. We first generate a word count matrix where each row is an article and each column represent a word. Next we generated a TF-IDF matrix and try to capture the latent semantic from it. The matrix is approximate of size 10,000X100,000. On the right are the comparison of 2 result, the first one is the clustering result for the original matrix, and the second one is the result after we perform SVD on the original matrix and kept 85% of the variance. The new matrix is approximate of size 10,000X4,000. In fact, we also tried multiple combinations, including removing all non-Chinese character, normalized/ normalized TF-IDF, etc. But the results are almost the same. From the plots above, we can come up with the following results:

(1) Even if we increased the number of clusters, the top cluster still gets most of the items, which means LSA cannot separate the articles.

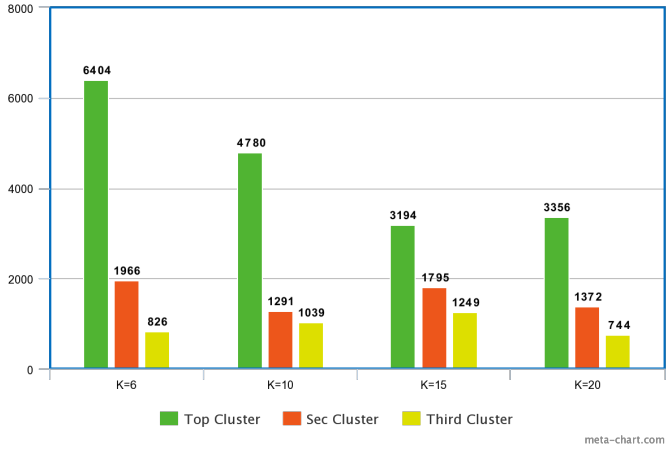

Fig. 1. Top3 cluster size without dimension reduction.

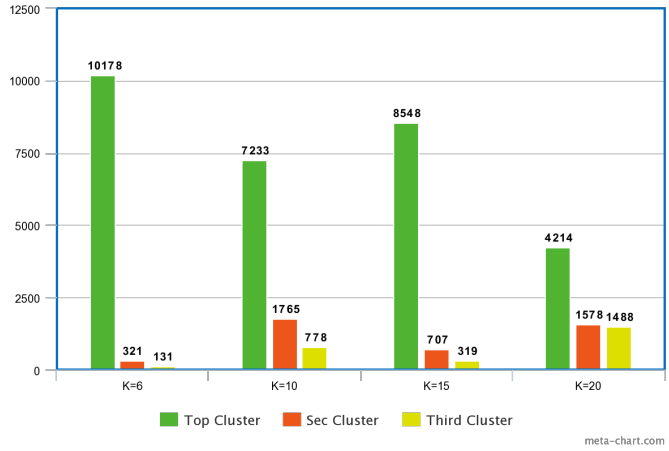

Fig. 2. Top3 cluster size after dimension reduction.

(2) K-means clustering performs better in original space than in lower dimensional space, which means the variance that SVD captured does not have positive impact when it comes to separate article based on their latent topics.
(3) Based on the fact that the outcomes of different pre-processing are almost the same, it's most likely LSA is not suitable in this case.

5.4 Latent Dirichlet Allocation

Since in our previous approach, most of the items are clustered together, we tried to impose sparsity prior and hoped the new model gives better clustering result. So LDA is the way we go.
In this experiment, we used online variational Bayes algorithm to do approximate inference. We first tried LDA with 6 topics. The results is shown in Fig. 4 shows the top words for each topic, a English translation is attached on the right. Seeing from the words,

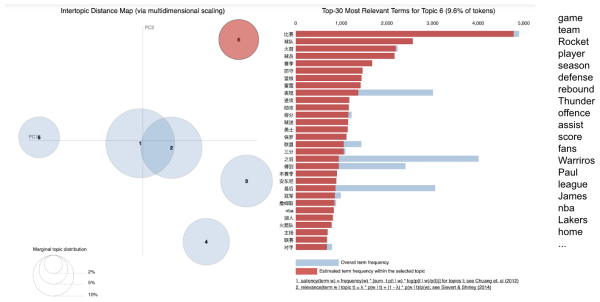

Fig. 3. Top words in LDA resulting topic6.

we can easily infer that this is sport topic. Other topics are just like sport, except for entertainment and fashion articles, which you can see from the figure the 2 circles are overlapping with each other. After checking the words, we found that these 2 topics have a lot of common words, so LDA cannot tell them apart. But apart from that, the result is significantly better than LSA, measured by both the number of items in each cluster and purity of each cluster. Results are shown below.

Country List		
Cluster	Number of items	Purity
cluster1	2728	50%
cluster2	581	97%
cluster3	2379	68%
cluster4	1732	98%
cluster5	1709	77%
cluster6	1513	84%

Later on, we tried running LDA with 15 topics hoping to get better clustering. The logic behind this is within each so called general topics, there could be some subtopics, e.g. within sport news, the subtopic could be basketball, football, and soccer. We made an assumption here: all the articles that belong to a general topic will have similar distribution over the subtopics, thus we can cluster them by applying K-means clustering on the distribution vectors.

After doing LDA, we'll get a 10,000X15 matrix where each row signifies the distribution of subtopic for each article. And then we applied K-means clustering. The results are shown below.

Country List		
Cluster	Number of items	Purity
cluster1	1661	99%
cluster2	1793	73%
cluster3	4142	41%
cluster4	1312	93%
cluster5	761	96%
cluster6	973	96%

As we can see, the later approach gives even better clusters. Most of the articles got clustered correctly. However, there's still one big cluster where 2 or 3 topics got messed together. We thought adding more topics to LDA model might help, but our computer is not capable of larger topics. We'll talk more about in next section.

6 CONCLUSIONS AND FUTURE WORKS

In this report, we tried 3 commonly used topic-modeling ways to cluster Chinese articles, namely Shingles approach, LSA and LDA. We found that Shingles did not work even in a small dataset, and LSA yields uneven cluster. LDA is able to separates the articles based on their latent topics and cluster based on the distribution of subtopics often gives better clusters. Throughout the process, we found that

(1) LDA can be used as a way of dimensionality reduction. In order to cluster articles into k groups, we can use LDA to embed the input data into n topic space, and then apply methods like K-means to cluster the articles.
(2) Further more, based on the result that we had, it's safe to say that statistical methods are also suitable for non-Romance languages like Chinese as long as we have a good vectorized matrix.

In terms of future work, we think that we can do better by trying the following approach:

(1) We should use the implementations that can distribute load and run in paralleled, like Spark or Tensor-flow, instead of Sklearn. This will give us more computing power so that we can try out more.
(2) In LSA, we can try some other clustering algorithms, especially K-medoids. Also try NMF.
(3) Besides TF-IDF, shingles, and word-count, we can vectorized the articles with word-to-vec, which takes in account the sequence of words as well as their semantic meanings.

REFERENCES

Michael I. Jordan David M. Blei, Andrew Y. Ng. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (Jan. 2003), 993–1022.

Susan T. Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology* 38, 1 (2004), 188–230. https://doi.org/10.1002/aris.1440380105

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (1999), 289–296.

Francis Bach Matthew D. Hoffman, David M. Blei. 2010. Online Learning for Latent Dirichlet Allocation. *NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems* 1 (2010), 856–864.