Cluster Theme extraction using Explicit Semantic Analysis

Natural Language Processing Project Report

Parsodkar Adwait Pramod (CS19E001) and Akshay Gupta (CS19S011)

Indian Institute of Technology Madras, India
https://www.iitm.ac.in/

Abstract. Document clustering techniques help us efficiently organize documents and allow efficient retrieval. Though the traditional clustering techniques can create clusters, they lack the ability to explicitly present the 'theme' of the cluster. These methods, at times, simply discard the semantic information present in the documents and assign a document to a cluster based on the words present in it. We make an attempt to arrive at concepts that indicate the theme of clusters using the rich semantic knowledge captured in Explicit Semantic Analysis (ESA) [5].

Keywords: Explicit Semantic Analysis · Latent Dirichlet Allocation · K-Means Clustering

1 Introduction

Data Clustering is traditionally seen as a problem of assigning an object to one or more groups, such that objects belonging to the same cluster are similar in some way. However, the standard techniques do not make an attempt to express the theme of the clusters. By theme of a cluster, we mean a set of concepts to which the documents in the cluster are strongly associated. In this project, our contribution is twofold: First, we propose an unsupervised text clustering method in a semantically rich concept space. Second, we propose heuristics to arrive at the theme of the document clusters obtained. This task requires the vector representation to be rich and highly interpretable. Explicit Semantic Analysis (ESA) projects text into high dimensional Wikipedia[1] concept space, dimensions of which are interpretable to humans.

2 Literature Survey

Several attempts have been made to capture the theme of a collection of documents. This demands clustering 'similar' documents together. Introspective

[1] https://www.wikipedia.org/

methods such as Latent Dirichlet Allocation (LDA) [4] can be used to assign a document to a topic, where each topic is a distribution over words. Every document is assigned to one of these topics, such that the probability of generating the document from the topic is greater than the probability of generating it from any other topic. A more formal description to LDA is included in Section 3. The theme of a topic can be interpreted using the 'k' most probable words in the topic. LDA, however, does not take into account the semantic knowledge present in the documents and is sensitive to the word forms used in the documents. As reported in [1], LDA performs poorly in arriving at a set of words that describe the theme for a topic.

Traditional text clustering techniques demand large amounts of labeled data, which is expensive and time consuming. Methods such as ClassifyLDA [2] and Topic-Sprinkled LDA [3] reduce the Knowledge Acquisition Bottleneck by demanding supervision over a set of topics, which is much smaller than the complete collection of documents. The supervision in ClassifyLDA requires the expert to analyze some 'm' most probable words in each topic and assign the topic to an expert defined class. In Topic Sprinkled LDA, the expert is additionally required to provide a set of representative 'artificial words' for each class, which are appended to the documents belonging to that class.

WikiLDA, on the other hand, demands no supervision and uses the semantic knowledge present in Wikipedia. WikiLDA sprinkles relevant Wikipedia concepts (referred to as WikiConcepts) to the documents, and uses Generalized Polya Urn (GPU) strategy to arrive at word distribution for each topic. The GPU strategy not only increases the probability corresponding to the observed word/WikiConcept but also the probability associated with its related words and WikiConcepts. In vanilla GPU-LDA, the text relatedness is a function of the co-document frequency of the two text entities. However, since the sprinkled WikiConcepts are very less likely to co-occur in the same document, WikiLDA defines relatedness between text entities as the cosine similarity between their ESA vectors.

The theme of a collection of documents in case WikiLDA can also be captured by the top 'k' most probable words in the corresponding topic. [1] concludes that WikiLDA performs significantly better than Standard Polya Urn LDA (SPU-LDA) and GPU-LDA with respect to the Topic Coherence score, due to which we take WikiLDA as our baseline.

3 Background

3.1 Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) is a non-introspective distributional semantic model. It represents the semantics of a word by transforming it to a high dimensional distributional vector in a predefined concept space dictated by some external knowledge source. As suggested in [5], Wikipedia is a good choice for external knowledge since it covers a wide range of topics. Additionally, since

Wikipedia is created by humans, it gives a natural measure of text relatedness. ESA with Wikipedia assumes the titles of the Wikipedia articles as orthogonal concepts which are considered to be human interpretable units of meaning.

In ESA, a word is expressed as a vector in the space spanned by the Wikipedia concepts, such that the i^{th} elements in the vector represent the strength of association of the word with the i^{th} concept. The measure of strength of association of a word in an article with a concept is given by the product of the 'Term frequency' of the word in the article and the 'Inverse document frequency' of the word in the Wikipedia corpus.

Using the vector representations of words in the wikipedia vocabulary, we can obtain the representation of a document by taking a weighted aggregation of the vectors corresponding to each word in the document.

This technique is termed 'explicit' since the dimensions of the vectors map directly to a human interpretable concept (wikipedia article title), unlike statistical techniques like latent semantic analysis whose dimensions are not human interpretable.

3.2 Cosine Similarity

In the context of text clustering, it is crucial to state when two documents are considered to be similar. We claim two documents to be similar if their relative association to the various concepts is similar. This requirement is effectively captured by the cosine-similarity measure. An advantage of using Cosine similarity is that it is not agnostic to the lengths of the two documents, unlike the estimation of similarity using euclidean distance. The cosine similarity measure for two vectors $\vec{u}$ and $\vec{v}$ is given by,

$$cos\_sim(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}||.||\vec{v}||}$$

where, $||\vec{u}||$ indicates the magnitude of the vector u, and $\cdot$ indicates the dot product operator.

3.3 K-Means Algorithm

The K-Means algorithm is an unsupervised learning algorithm used to cluster vectors based on some distance measure. It tries to minimize a criterion called the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified in advance. The algorithm follows the given three steps:-

1. The first step chooses the initial centroids. After initialization, K-means consists of looping between the following two steps.
2. The second step assigns each vector to its nearest centroid.
3. The third step creates new centroids by taking the mean value of all of the vectors assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these last

two steps until this value is less than a threshold. That is, it repeats until the centroids do not move significantly.

Given enough time, K-means will always converge, however, this may be to a local minimum. The time for convergence is highly dependent on the initialization of the centroids. As a result, the computation is often done several times, with different initializations of the centroids.

Since K-means typically minimizes euclidean distance and we wish to optimize for cosine similarity, we normalize the input vectors causing the overall result of minimizing the euclidean distance to be the same as that of maximizing the cosine similarity. The proof for this is as follows: Consider two vectors $\vec{u}$ and $\vec{v}$.
The normalized vectors will be given by,

$$\hat{u} = \frac{\vec{u}}{|u|} = [u_1, u_2 ... u_d]^T$$

and,

$$\hat{v} = \frac{\vec{v}}{|v|} = [v_1, v_2 ... v_d]^T$$

The Euclidean Distance between $\hat{u}$ and $\hat{v}$ is given by,

$$euclidean\_dist(\hat{u}, \hat{v}) = \sqrt{\sum_{i=1}^{d} (u_i - v_i)^2}$$

And, The cosine similarity between $\hat{u}$ and $\hat{v}$ is given by,

$$cosine\_similarity(\hat{u}, \hat{v}) = \sum_{i=1}^{d} (u_i \times v_i)$$

So the cosine distance between $\hat{u}$ and $\hat{v}$ will be given by,

$$cosine\_distance(\hat{u}, \hat{v}) = 1 - cosine\_similarity(\hat{u}, \hat{v})$$

Now, Consider the Euclidean distance,

$$\begin{aligned} euclidean\_dist(\hat{u},\hat{v}) &= \sqrt{\sum_{i=1}^{d}(u_i - v_i)^2} \\ &= \sqrt{\sum_{i=1}^{d}(u_i^2 + v_i^2 - 2u_i v_i)} \\ &= \sqrt{\sum_{i=1}^{d} u_i^2 + \sum_{i=1}^{d} v_i^2 - 2\sum_{i=1}^{d} u_i v_i} \\ &= \sqrt{1 + 1 - 2\sum_{i=1}^{d} u_i v_i} \\ &= \sqrt{2} \times \sqrt{1 - \sum_{i=1}^{d} u_i v_i} \\ &= \sqrt{2} \times \sqrt{cosine\_dist(\hat{u},\hat{v})} \end{aligned}$$

3.4 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised generative probabilistic model used to discover topics in a collection of documents. In LDA, each document is generated by choosing a distribution over topics and then choosing each word in the document from a topic selected according to the distribution. Generative process of LDA can be described as follows:

1. for t = 1...T
   (a) $\phi_t \sim Dirichlet(\beta)$
2. for each document d ϵ D
   (a) $\phi_d \sim Dirichlet(\alpha)$
   (b) for each word w at position n in d
      i. $z_d^n \sim$ Multinomial(θ_d)
      ii. $w_d^n \sim$ Multinomial(z_d^n)

3.5 Approaches for Text Clustering

This section briefly discusses the two approaches to text clustering and situate LDA, WikiLDA and our model in this context.

Generative Probabilistic Models These models assume a document to be generated from a single distribution of words(called topic in context of LDA), that is, a document is considered to be associated with only one topic. Text Clustering using Latent Dirichlet Analysis (LDA) is an example of such a model. WikiLDA also fall into the category of Generative Probabilistic models.

Similarity-based Approaches These models explicitly define a similarity function to measure similarity between two text objects. The goal is to find an optimal partitioning of data so as to maximize intra-group similarity and minimize inter-group similarity.

Clustering using the K-Means algorithm on ESA vectors is an example of similarity based approach in which the similarity is estimated using euclidean distance on normalized vectors, which is shown to produce results equivalent to cosine similarity.

4 Methodology

In our proposed technique to arrive at the theme of a collection of similar documents, we begin with representing every document as a normalized ESA vector. These vectors are assigned to clusters using the K-Means algorithm with the number of clusters as the hyperparameter. Once the clusters are made, one of the following heuristics can be used to arrive at the representative vector for each cluster.

Let the vectors belonging to a cluster be :

$$\begin{gathered} v_1 = [v_{1,1}, v_{1,2}, ..., v_{1,n}] \\ v_2 = [v_{2,1}, v_{2,2}, ..., v_{2,n}] \\ ... \\ v_m = [v_{m,1}, v_{m,2}, ..., v_{m,n}] \end{gathered}$$

1. *The mean vector:* The mean vector corresponding to a cluster is defined as the vector whose i^{th} element is obtained by taking the arithmetic mean of the i^{th} element of all the vectors belonging to the cluster. That is, the i^{th} element is given by,

$$v\_mean_i = mean(v_{1,i}, v_{2,i}...v_{m,i})$$

2. *The median vector:* The median vector corresponding to a cluster is defined as the vector whose i^{th} element is the median of the set formed by the i^{th} element of all the vectors belonging to the cluster. That is, the i^{th} element is given by,

$$v\_median_i = median(v_{1,i}, v_{2,i}...v_{m,i})$$

3. *The minimum vector:* The minimum vector corresponding to a cluster is defined as the vector whose i^{th} element is the minimum value in the set formed by the i^{th} element of all the vectors belonging to the cluster. The minimum vector captures the bare minimum strength of association that each of the documents has with the various Wikipedia concepts. That is, the i^{th} element is given by,

$$v\_min_i = min(v_{1,i}, v_{2,i}...v_{m,i})$$

The top 'k' concepts having the strongest association with the representative vector of a cluster are considered to capture the theme of the cluster. The results indicate that the minimum vector heuristic is agnostic to outliers and gives less coherent cluster labels.

We hypothesize that the effectiveness of this method is due to the following reasons:

1. Good quality external knowledge from Wikipedia.
2. ESA not being dependent on the word form.
3. Since documents are expressed in high dimensional space, documents dealing with different concepts will be placed far apart.

5 Dataset

5.1 NIPS

Neural Information Processing Systems (NIPS) is one of the top machine learning conferences in the world. It covers topics ranging from deep learning and computer vision to cognitive science and reinforcement learning. This dataset includes the extracted text for all NIPS papers ranging from the first 1987 conference to the 2016 conference.

Topics for ESA: The ESA model is based on concepts derived from Wikipedia articles on 26-05-2020 pertaining to topics obtained from PetScan with the queries being 'Machine Learning', 'Artificial Intelligence' and 'Deep Learning'. A total of 16088 Wikipedia Article Titles were extracted by setting the depth parameter to 3.

Input documents: Input dataset consists of texts extracted from 7241 NIPS conference papers.

5.2 20 Newsgroups

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

Topics for ESA: The ESA model is based on concepts derived from Wikipedia articles on 27-04-2020 pertaining to topics obtained from PetScan with the queries being 'Apple Inc', 'Christianity', 'Computer Hardware', 'Hockey', 'Medicine', 'Personal Computers', 'Politics', 'Science' and 'Weapons'.

Input documents: The input dataset consists of a subset of 20 Newsgroups data set containing documents related to 'Medicine', 'Christianity', and 'Hockey'.

6 Evaluation

It is desired that the words and/or phrases denoting the theme of the collection of documents are coherent. The notion of coherence between two pieces of text is approximated by the cosine similarity between their ESA vectors. That is, texts having higher cosine similarity are assumed to be more coherent.

Formally, the Coherence measure for the i^{th} collection of documents considering the top 'k' representative words/phrases is defined as the sum of cosine similarity between the $\binom{k}{2}$ distinct pairs of the words/phrases.

The Coherence measure can take a minimum value of 0 when all the words/phrases are orthogonal and a maximum value of $\binom{k}{2}$ when all the vectors are aligned in the same direction. Note that the ESA vector has elements which are non-negative, preventing the cosine similarity from becoming negative.

$$Coherence(T_i) = \sum_{i} \sum_{j>i} cosine\_similarity(ESA(w_i), ESA(w_j))$$

where, $T_i =< w_1, w_2, ..., w_k >$ such that w_j is less probable than w_i if $i < j$.

7 Results

7.1 NIPS Dataset

The performance of the model is evaluated based on the Coherence measure with k = 20 averaged over the number of clusters (in case of clustering using ESA) or topics (in case of LDA and its variants). The hyperparameter for K-Means algorithm is set to 50. It is important to note that the Coherence values reported for Clustering using ESA is obtained by using a subset of Wikipedia articles relevant to NIPS, whereas a snapshot of complete Wikipedia was used for ClassifyLDA, TS-LDA and WikiLDA as reported in wikilda paper.

Table 1. Average Coherence scores for SPU-LDA, GPU-LDA, WikiLDA and Clustering using ESA on NIPS dataset

Model	Result
SPU-LDA	1.852
GPU-LDA	2.624
WikiLDA	9.182
ESA clustering using Mean Vector	46.36
ESA clustering using Median Vector	46.74
ESA clustering using Minimum Vector	38.97

The following tables include words/WikiConcepts indicative of the theme of some collections of similar documents. It can be observed that Clustering using

ESA captures the theme in terms of WikiConcepts, whereas WikiLDA expresses it in terms of words and WikiConcepts. LDA is found to perform poorly since it expresses the theme only in terms of words.

Table 2. Top 5 strongest concepts for some collection of similar documents obtained using ESA on NIPS Dataset using minimum vector

No.	Result
1	'Natural-language understanding', 'Artificial intelligence', 'Word-sense disambiguation', 'N-gram', 'Language model'
2	'Deep learning', 'Types of artificial neural networks', 'Artificial neural network', 'Convolutional neural network', 'Recurrent neural network'
3	'Principal component analysis', 'Robust principal component analysis', 'Non-negative matrix factorization','Nonlinear dimensionality reduction', 'Latent semantic analysis'

Table 3. Top 5 strongest concepts for some collection of similar documents obtained using ESA on NIPS Dataset using mean vectors

No.	Result
1	'Statistical machine translation', 'Machine translation', 'Example-based machine translation', 'Paraphrasing (computational linguistics)', 'Language model'
2	'Deep learning', 'Convolutional neural network', 'Types of artificial neural networks', 'Deep belief network', 'Artificial neural network'
3	'Principal component analysis', 'Sparse PCA', 'Low-rank approximation', 'Robust principal component analysis', 'Matrix regularization'

Table 4. Top 5 strongest concepts for some collection of similar documents obtained using ESA on NIPS Dataset using median vectors

No.	Result
1	'Statistical machine translation', 'Language model', 'N-gram', 'Example-based machine translation', 'Pattern theory'
2	'Deep learning', 'Types of artificial neural networks', 'Convolutional neural network', 'Artificial neural network', 'Convolutional deep belief network'
3	'Principal component analysis', 'Low-rank approximation', 'Matrix regularization', 'Robust principal component analysis', 'K-SVD'

Table 5. Top 9 most probable words for some topics obtained using WikiLDA on NIPS dataset

No.	Result
1	document, topic, dirichlet, entity, Latent Dirichlet allocation, lda, Dirichlet-multinomial distribution, Dirichlet process
2	speech, hmm, speak, frame, utterance, band, Viterbi algorithm,Hierarchical hidden Markov model, Markov model
3	policy, reward, agent, Markov decision process, mdp, Reinforcement learning,reinforcement, Q-learning, plan

Table 2-4 lists some of the results obtained by our model using different heuristic, namely minimum, mean, and median vectors as representative vectors, whereas Table 5 shows the results obtained using WikiLDA on NIPS dataset. As it can be observed in Table 2 that the use of concepts makes the underlying theme of documents more coherent. For instance in WikiLDA, in the first topic, word 'document' on its own is not very descriptive of the underlying theme whereas in our model, in the first result, every term directly or indirectly adds to the general theme of 'Natural Language Understanding'.

7.2 Newsgroups Dataset

Table 6-8 shows the result obtained by our method using different heuristics on the Newsgroups dataset whereas Table 9 shows the result obtained by applying LDA on the same dataset. Here the first result of each table, represents the theme of documents pertaining to 'Christianity', however it can be seen that the results obtained by using concepts is much more representative compared to when using purely words. For instance the word 'does', in first result in Table 9, does not add anything meaningful to the theme of 'Christianity', whereas in our model each term is directly or indirectly representative of the underlying theme.

Table 6. Top 5 strongest concepts corresponding to true cluster labels - Christianity, Hockey and Medicine in The 20 Newsgroups Dataset using median vector

No.	Result
1	'Christianity', 'Apostasy in Christianity', 'Salvation in Christianity', 'Nontrinitarianism', 'Christianity and other religions'
2	'History of the National Hockey League', 'Ice hockey', 'Ice hockey in the United States', 'Original Six', 'Ice hockey in popular culture'
3	'Logology (science)', 'History of medicine', 'Science', 'History of science', 'Self-experimentation in medicine'

Table 7. Top 5 strongest concepts corresponding to true cluster labels - Christianity, Hockey and Medicine in the 20 Newsgroups Dataset using minimum vectors

No.	Result
1	'Christianity', 'Born again', 'Biblical inerrancy', 'Relationship between religion and science', 'Salvation in Christianity'
2	'Overtime (ice hockey)', 'History of the National Hockey League', 'Towel Power', 'Ice hockey', 'Goal (ice hockey)'
3	'Apostasy in Christianity', 'Logology (science)', 'Amazon (company)', 'Misinformation related to the 2019–20 coronavirus pandemic', 'Misinformation related to the 2019–20 coronavirus pandemic'

Table 8. Top 5 strongest concepts corresponding to true cluster labels - Christianity, Hockey and Medicine in The 20 Newsgroups Dataset using mean vector

No.	Result
1	'Christianity', 'Apostasy in Christianity', 'Salvation in Christianity', 'Accommodation (religion)', 'Nontrinitarianism'
2	'Ice hockey', 'History of the National Hockey League', 'Ice hockey in the United States', 'Original Six', 'Ice hockey in popular culture'
3	'Logology (science)', 'History of medicine', 'Medicine', 'Science', 'History of science'

Table 9. Top 10 most probable words corresponding to true cluster labels - Christianity, Hockey and Medicine in the 20 Newsgroups Dataset

No.	Result
1	'god', 'people', 'jesus', 'believe', 'does', 'bible', 'say', 'does', 'life', 'church'
2	'year', 'game', 'team', 'games', 'season', 'play', 'hockey', 'players', 'league', 'player'
3	'mr', 'line', 'rules', 'science', 'stephanopoulos', 'title', 'current', 'define', 'int', 'yes'

8 Conclusion

We proposed a new method for document clustering using ESA which takes into consideration the rich semantic knowledge from Wikipedia. This method is extended to arrive at sets of concepts that capture the theme of the clusters. The theme obtained from Clustering using ESA was compared with that

obtained from LDA and WikiLDA and empirical study indicated that the proposed method performed better. The quantitative measure for this, called the Coherence, also suggests the same.

9 Future Work

We aim to empirically analyze the quality of clusters formed using the K-Means algorithm on the ESA representation of the documents and the theme concepts obtained using the various heuristics. An in-depth analysis becomes crucial to check if one representative vector for a collection of documents is sufficient to capture the important themes covered. Further, attempts need to be made to handle the the sparse nature of the ESA vectors without compromising on the interpretability.

References

1. Hingmire, Swapnil, et al. “WikiLDA: Towards more effective knowledge acquisition in topic models using Wikipedia.” Proceedings of the knowledge capture conference. 2017.
2. Hingmire, Swapnil, et al. “Document classification by topic labeling.” Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013.
3. Hingmire, Swapnil, and Sutanu Chakraborti. “Sprinkling topics for weakly supervised text classification.” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014.
4. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation.” Journal of machine Learning research 3.Jan (2003): 993-1022.
5. Gabrilovich, Evgeniy, and Shaul Markovitch. “Computing semantic relatedness using wikipedia-based explicit semantic analysis.” IJcAI. Vol. 7. 2007.