Topic Grouping by Spectral Clustering

Young-Seob JEONG, Won-Jo LEE, Ho-Jin CHOI

Department of Computer Science,
KAIST(Korea Advanced Institute of Science and Technology),
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea

{pinode, mochagold, hojinc}@kaist.ac.kr

Abstract— **With the growing number of web documents, it becomes difficult to analyze and obtain information from such an array of documents. Furthermore, unsupervised methods are preferable, as most web documents are unlabeled. Probabilistic topic modeling is one such method. It discovers latent structures among unstructured documents. While many traditional topic models usually assume that the topics are independent of each other, some models have been proposed to obtain correlations between the topics or a hierarchy of the topics. These models are designed to obtain both the topics and the correlations without using any other method. Therefore, very few studies apply other methods to determine a correlation between topics. In this paper, we apply spectral clustering to group the topics obtained from a traditional topic model, in this case the Latent Dirichlet Allocation model. To the best of our knowledge, this is the first approach that uses spectral clustering for the grouping of topics. We demonstrate the experimental results with various settings.**

***Keywords*— Topic model, Spectral clustering, Topic grouping**

I. INTRODUCTION

The enormous number of unstructured web documents is becoming an increasingly important source of knowledge and wisdom, making unsupervised topic-modeling methods preferable, as most of these documents are unlabeled. Probabilistic topic modeling is one such unsupervised method, and many studies have been conducted in relation to it. Here, a topic is defined as a latent structure beyond the observed documents and is represented as a distribution of the vocabulary of the documents.

Recent studies analyze the correlations or the hierarchies among topics [3-5]. These approaches are commonly based on the hypothesis that topic correlation or topic hierarchy can be obtained from the co-occurrence of topics, and these extended topic models are designed for discovering both the topic and the correlation concurrently. However, too few studies have utilized other methods for discovering correlations, e.g., topic groups, from among the topics obtained from topic models. Therefore, in this paper, we employ a spectral clustering technique [8] to group topics obtained from a typical topic model, in this case the Latent Dirichlet Allocation (LDA) model [2]. The spectral clustering technique groups the nodes of a given graph, leading to the construction of a graph consisting of topic nodes, after which it computes the edge weights of all possible pairs of topic nodes.

The contribution of this work is twofold. First, this is the first approach that uses a spectral clustering technique to group the topics obtained from a topic model. Second, four different distance functions are tested and compared for computing the similarities or weights between the topic nodes.

The rest of the paper is organized as follows. Section 2 provides background knowledge and related work. Section 3 presents the proposed approach, and section 4 shows the results of the experiments. Finally, section 5 concludes the paper.

II. RELATED WORK

A. *Topic Models*

Since probabilistic latent semantic indexing (PLSI) [1] and the LDA model [2] were proposed, many probabilistic topic models have been proposed. Each topic model has its own unique structure, as it has a unique hypothesis regarding the data. With the same data, therefore, different topic models produce different results. For example, for documents with sequential information, e.g., time tags, many topic models, including the Dynamic Topic (DT) model [3], the Sequential LDA (S-LDA) model [4], and the Sequential Entity Group Topic (S-EGT) model [5] aim to discover temporal patterns of topics. Although they have similar purposes, the three models give different results, as they have different hypotheses even with the same data.

There are some topic models that discover correlations or hierarchies of topics. For example, the Correlated Topic (CT) model [6] employs a Gaussian kernel to model correlations between topics. The Hierarchical LDA (H-LDA) model [7] gives a hierarchy of topics, where only the leaf nodes of the hierarchy are word distributions. Each document has a distribution of topics along with a path from the root node to the leaf node. All of these models are designed to obtain correlations or hierarchies of topics. Therefore, the topics are obtained concurrently with the correlation or hierarchy. In other words, these models accomplish the two tasks of obtaining topics and obtaining correlations or hierarchies concurrently. The existing models are also commonly based on the hypothesis that the topic correlation or topic hierarchy can be obtained from the co-occurrences of the topics, as the topics are obtained from the co-occurrences of the words. However, there are too few studies concerning how to obtain correlations between topics, i.e., topic groups, using other methods. Accordingly, in this paper, we apply a graph clustering algorithm, e.g., spectral clustering, for such a task, based on the new hypothesis that the topic correlation can be obtained from the topics themselves without using the co-occurrences of the topics. To this end, we use the traditional LDA model in this paper, as the objective of the paper is to investigate how well the topics can be grouped using the graph clustering algorithm, especially using spectral clustering.

B. Graph Clustering

Clustering algorithms have been researched steadily as the Internet has come into wider use. As the size of the data on the Internet is continually increasing, graphical representations of the data become complicated and very large. It is necessary, therefore, to develop clustering techniques for graphs, and a number of algorithms have been introduced, along with several concepts. These include a spectral clustering algorithm [8], METIS [9], modularity maximization [10], and a co-clustering [11] algorithm. METIS employs multilevel k-way partitioning and uses three phases, termed coarsen, initial-partitioning, and refinement. Modularity maximization gives high modularity to a network which has strong intra-connections and weak interconnections to other networks, resulting in clustered networks having high modularities. The co-clustering method, given a similarity matrix, simultaneously groups the row items and the column items. Identical to other clustering algorithms, spectral clustering is based on the finding 'good cuts' in the graph. It seeks sparse external edges from among the clusters and dense internal edges. One issue which arises when finding clusters is the issue of imbalanced small clusters. To deal with these, several metrics have been proposed, such as the ratio cut [12] and the normalized cut metrics [8]. As the optimization of the metrics is NP-hard, several approximation algorithms and heuristic methods have been proposed [9, 12, 13]. In this paper, we focus on the normalized Laplacian to find approximate solutions for spectral clustering [8].

To measure clustered graphs, conductance [14] is a widely used metric. Essentially, conductance seeks to verify how fast a random walker moves from a particular cluster to another cluster. The conductance value is small when the internal connections within a cluster are strong and the external connections are weak. Formally, given the graph G = (V, E) where V is the set of vertices and E is the set of edges, the conductance C(A) for A $\subseteq$ V is defined as follows:

$$C(A) = \frac{cut(A,\overline{A})}{min\{\, eV(A), eV(\overline{A})\}} \tag{1}$$

where $cut(A,\overline{A}) = \left|\{u,v\} : u \in A, v \in \overline{A}\right|$ and $eV(A) = \sum_u \sum_v \{u,v\}$.

III. Topic Grouping

Topic grouping clusters the topics obtained from a particular topic model. Although there are several approaches that can be used to discover correlations among topics, this is the first approach to use a graph clustering technique for such a task. While existing approaches commonly hypothesize that topic correlations can be obtained from the co-occurrences of the topics, this approach is based on the new hypothesis that the topic correlations can be obtained from the topics themselves. In other words, the similarities between the topics may convey enough information for grouping them.

The proposed approach consists of four steps. The first step is the topic modeling step that we use the LDA model. The LDA model generates the latent compact representation (i.e., topics) from the observed word sequences, where each topic is a word distribution. After we obtain *T* topics from the LDA model, we get a matrix of *T*×*T* weights between the topics in the second step named as the weight computation step. The weight is typically the similarity between each pair of topics; therefore, it is necessary to convert the distances between the topics into the similarities when the distances are given. For example, we can use the Euclidean distances between all possible pairs of topics to create the *T*×*T* symmetric distance matrix, which will be converted into the similarity matrix. We apply four distance functions for this step in our experiments: the Euclidean distance, Hellinger distance, Jensen-Shannon (JS) divergence, and Kullback-Leibler (KL) divergence functions.

The third step is the topic grouping step that the spectral clustering is applied to the obtained similarity matrix; thus, *T* topics (i.e., nodes) will be grouped into *G* clusters. Given the number of clusters *G*, this step seeks graph cuts that reduce the weight sum of the cut edges. The topics of each cluster have heavy weights within the cluster but low weights between different clusters. In other words, the topics of each cluster are similar to each other, while the topics of different clusters are dissimilar to each other. Note that this step is conducted independently of the topic modeling step. While the topics are the latent semantic representation of the observed documents, the topic grouping step simply involves clustering the topics using spectral clustering, which is only concerned with the weights between the nodes (i.e., topics) and now with the latent semantics. This independent grouping of topics is based on our hypothesis that the topic correlation can be obtained from the topics themselves. For the last step of the evaluation, the conductance of the clustered nodes is computed, where lower is better.

IV. Experiments

To demonstrate the usefulness of the proposed approach, we conduct experiments on NIPS papers from a five-year period from 1987 to 1991. In total, there are 1,740 documents, 397,229 sentences, and 2,896,233 words. We removed stop-words and performed stemming using the Porter stemmer. The size of the vocabulary is 172,981. The sentences were recognized by '.', '?', '!', and 'newline'. For the parameters of the LDA model, we set α=0.1 and β=0.1 symmetrically, as the initial values do not have much of an effect on the result for a large dataset. The total number of iterations for the LDA model is 1,000. We set the sigma parameter of spectral clustering to 0 such that the weight computation step would be done by the self-tuning of the spectral clustering. Note that we do not investigate the best parameter setting for spectral clustering in this paper, whereas we do investigate how useful spectral clustering is for grouping topics with diverse distance functions. Thus, self-tuning is reasonable for the overall process.

To measure the grouped topics obtained from the proposed approach, we performed two experiments: (1) topic comprehension of each cluster, and (2) a conductance comparison step. The first experiment sought to assess how much the grouped topics are comprehensible by humans. The second experiment simply computes the conductance of the obtained topic clusters, where lower conductance is better, as noted earlier.

A. Comprehension of Topic Groups

To measure the comprehension of the grouped topics, it is initially necessary to determine the appropriate setting of the number of topics *T*. We varied this from 10 to 50, finding that there is less redundancy of the topics when *T* = 40. Here, redundancy means that multiple topics exist which are nearly identical to each other. Some of the discovered topics when *T* = 40 are listed in Table 1, where the top 10 words for each topic are shown. According to the top 10 words, we manually determined the labels of the topics. As represented in the table, the top 10 words are representative of every topic. Thus, each topic is comprehensible by humans. For example, for the top 10 words of the 20th topic, humans would most likely produce the label ‘speech recognition’.

TABLE 1. SOME OF THE FOUND TOPICS WHEN T=40

Label (Topic index)	**Top 10 words**
Component identification (n1)	source, component, ica, blind, separate, signal, independent, eeg, pca, matrix
Cancer diagnosis (n2)	diagnosis, patient, cancer, clinic, medic, hpnn, blood, util, context, idnn
Protein sequence (n3)	protein, chain, mouse, region, sequence, precursor, acid, human, amino, receptor
Learning algorithm (n4)	function, algorithm, learn, set, result, vector, case, estimate, error, optimize
Kernel methods (n7)	kernel, svm, support, regression, vapnik, margin, sv, machine, pca, ridge
Neural network theory (n8)	bound, network, theorem, function, threshold, compute, proof, polynomial, neural, class
Autonomic vehicle drive (n9)	hint, road, monotonic, vehicle, steer, drive, autonomic, lane, umli, cdm
Memory management (n10)	memory, code, capacity, decode, associate, store, recall, retrieve, bit, item
Visual model for cell (n11)	visual, model, cell, field, motion, spatial, orient, direct, response, recept
Neural network (n12)	network, state, dynamic, system, oscillate, attractor, neuron, neural, time, fix
Image object recognition (n13)	image, object, feature, recognition, face, pixel, view, vision, segment, visual
Reinforcement learning (n14)	state, learn, action, reinforce, policy, control, time, optimize, algorithm, reward
Neural network training (n18)	network, learn, input, train, neural, unit, output, perform, set, weight
Language grammar structure (n19)	symbol, rule, language, string, grammar, structure, state, connection, tree, represent
Speech recognition (n20)	speech, word, recognition, hmm, speaker, character, train, phoneme, segment, system
Tree classification (n23)	classification, train, set, tree, decision, algorithm, data, test, error, label
Sound filter (n24)	signal, auditory, sound, frequency, filter, channel, spectral, cochlear, spectrum, noisy
Neuron (n25)	neuron, cell, model, spike, fire, time, active, synaptic, input, response
Perceptron (n28)	student, teacher, general, physics, perceptron, replica, overlap, decay, average, train
Control model for robot (n29)	control, model, motor, robot, trajectory, arm, movement, forward, inverse, dynamic
Circuit implementation (n33)	circuit, chip, analog, implement, current, voltage, vlsi, output, neuron, input
Mixture model (n34)	model, data, mixture, gaussian, probabilistic, bayesian, distribution, likelihood, prior, posterior

Although each topic is comprehensible, we need to investigate how comprehensible the topic groups are by humans. We varied the number of groups *G* from 2 to 10, finding that it is most reasonably clustered when *G* = 4. The settings of the parameters *T* and *G* should be determined manually, as it will be useless if the topics of each cluster are not reasonable to humans.

The result of the topic grouping with *T* = 40 and *G* = 4 is depicted in Fig. 1, where the similarities between the topics are obtained using the Hellinger distance function. Although we used four distance functions, we chose the Hellinger distance for this experiment because its conductance generally was best among the four. We will discuss conductance in the following subsection. The bottom cluster, shown in purple, contains many topics, whereas the other clusters contain only 4 ~ 7 topics. Some of the topics are listed in Table 1. The cluster shown in red has four topics: learning algorithm, reinforcement learning, neural network training, and control model for robot. This implies that the red cluster can be regarded as learning or training, because the corresponding topics are commonly related to the learning of something. This also means that one of the main research areas of NIPS is to investigate learning or training algorithms. The cyan-colored cluster has the topics component identification, autonomic vehicle drive, image object recognition, language grammar structure, tree classification, perceptron, and a mixture model. Although the topics of this cluster are related to methods and tasks, they belong to the same cluster. Note that the similarity between each pair of topics is obtained using a particular distance function. The distance between the pair of topics will be smaller if the representative words of the topics do not co-occur frequently. Here, the top 10 words of the methods and the tasks co-occurred frequently in same context, which implies that the methods were frequently used for solving the tasks. To be specific, the four methods of component identification, tree classification, perceptron, and the mixture model are frequently used to solve the three tasks of autonomic vehicle drive, image object recognition, and language grammar structure. The topics of the cluster shown in green commonly discuss neural networks, indicating that neural networks are one of the most popular topics of NIPS. As described above, the four clusters are easily comprehensible by humans; hence, the proposed approach is useful for grouping topics obtained from a topic model, e.g., the LDA model in this case.

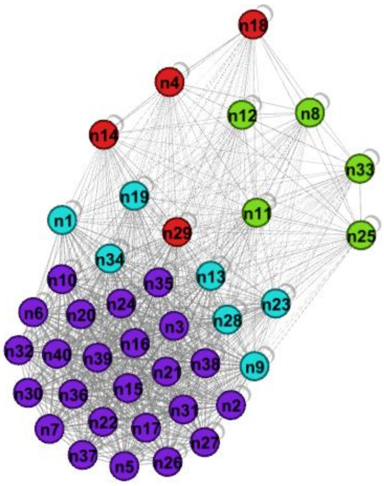


Figure 1. Result of topic grouping through spectral clustering, where G = 4 and T = 40. Node nx denotes the x-th topic, and each color is associated with a topic cluster. An undirected line between two nodes represents the similarity between them

B. Conductance of Topic Clusters

To measure how well the graph nodes are clustered, the measure of conductance is traditionally used. When *T* is fixed to 40, we varied *G* from 2 to 10. The resulting conductances are listed in Table 2. The conductance values are obtained in two steps. First, we compute the sum of the conductances of all possible cluster pairs, as conductance is essentially computed only for two clusters. Second, we repeat the first step five times, resulting in five sums of conductances for each possible pair. We then compute the averaged values of the sums. This step is required because the proposed approach relies on a stochastic method, i.e., the LDA model, and the k-means algorithm. The LDA model gives slightly different results even with the same data, while the k-means algorithm may give different results with different initial settings. Thus, we performed the overall process five times independently, and computed the average values.

As shown in Table 2, the conductance is generally small with the Hellinger distance. Therefore, we can conclude that the Hellinger distance is the best distance function for the topic grouping step. While we applied the four distance functions for the spectral clustering topic grouping step, there are other clustering algorithms, e.g., the METIS algorithm [9]. There are also many other distance functions or similarity functions, such as cosine similarity. We plan to conduct more experiments using these other distance functions and clustering algorithms in a future work.

TABLE 2. THE CONDUCTANCE VALUES WHEN T=40

G	Euclidean distance	Hellinger distance	JS divergence	KL divergence
2	0.8772	**0.8469**	0.89962	0.9195
4	5.4343	**5.1647**	5.54488	5.6331
6	14.1162	**13.6364**	14.2842	14.1549
8	26.841	**26.107**	26.6421	26.6987
10	43.2762	**42.2536**	43.5689	43.0337

V. CONCLUSIONS

Correlations between topics have been successfully discovered by existing topic models, but there are too few studies involving the application of other techniques for such a task. While existing approaches are commonly based on the hypothesis that correlations or hierarchies of topics can be obtained using the co-occurrences of the topics, we propose a four-step process of topic grouping based on the new hypothesis that the topic correlations can be obtained from the topics themselves. We employ spectral clustering for grouping the topics generated from LDA model. We demonstrated that the proposed approach is useful for grouping the topics by empirical results using four distance functions. In a future work, we will apply other clustering algorithms, for example, the METIS algorithm, to the proposed approach.

ACKNOWLEDGMENT

This work was supported by the Korea Ministry of Sciece, ICT and Future Planning, under the title “Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services” (No. 10044577)

REFERENCES

[1] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 50-57, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” in *Proceedings of Neural Information Processing Systems (NIPS)*, pp. 601-608, 2001.

[3] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 113-120, 2006.

[4] L. Du, W. L. Buntine, and H. Jin, “Sequential latent Dirichlet allocation: discover underlying topic structures within a document,” in *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pp. 148-157, 2010.

[5] Y. S. Jeong and H. J. Choi, “Sequential entity group topic model for getting topic flows of entity groups within one document,” in *Proceedings of the Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pp. 366-378, 2012.

[6] D. M. Blei and J. D. Lafferty, “Correlated topic models,” in *Proceedings of Neural Information Processing Systems (NIPS)*, 2005.

[7] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, “Hierarchical topic models and the nested Chinese restaurant process,” in *Proceedings of Neural Information Processing Systems (NIPS)*, 2003.

[8] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.

[9] G. Karypis and V. Kumar, “Multilevel k-way hypergraph partitioning,” in *Proceedings of the Design and Automation Conference (DAC)*, pp. 343-348, 1998.

[10] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical Review E*, vol. 70, no. 6, pp. 1-6, 2004.

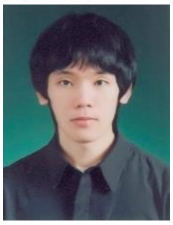
[11] I. S. Dhillon, S. Mallela, and D. Modha, “Information-theoretic co-clustering,” in *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 89-98, 2003.

[12] S. Wang and J. M. Siskind, “Image segmentation with ratio cut,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 675-690, 2003.

[13] K. Lang and S. Rao, “A flow-based method for improving the expansion or conductance of graph cuts,” *Integer Programming and Combinatorial Optimization*, pp. 383-400, 2004.

[14] R. Kannan, S. Vempala, and A. Vetta, “On clusterings: good, bad and spectral,” *Journal of the ACM*, vol. 51, no. 3, pp. 497-515, 2004.

Young-Seob Jeong is currently a PhD student in the Dept. of Computer Science at KAIST. His current research interests include topic modeling, deep learning, and action prediction based on various sensor data.



Won-Ji Lee received a bachelor's degree of computer science in 2012 from Hanyang University. He is currently a MS candidate student in the department of computer science at Korea Advanced Institute of Science and Technology.

Ho-Jin Choi is currently an associate professor in the Dept. of Computer Science at KAIST. In 1982, he received a BS in Computer Engineering from Seoul National University, Korea, in 1985, an MSc in Computing Software and Systems Design from Newcastle University, UK, and in 1995, a PhD in Artificial Intelligence from Imperial College, London, UK. From 1982 to 1989, he worked for DACOM, Korea, and between 1995 and 1996, worked as a post-doctoral researcher at Imperial College. From 1997 to 2002, he served as a faculty member at Korea Aerospace University, Korea, then from 2002 to 2009 at Information and Communications University (ICU), Korea, and since 2009 he has been with the Dept. of Computer Science at KAIST. Between 2002 and 2003, he visited Carnegie Mellon University (CMU), Pittsburgh, USA, and has been serving as an adjunct professor of CMU for the program of Master of Software Engineering (MSE). Between 2006 and 2008, he served as the Director of Institute for IT Gifted Youth at ICU. Since 2010, he has been participating in the Systems Biomedical Informatics National Core Research Center at the Medical School of Seoul National University. Currently, he serves as a member of the boards of directors for the Software Engineering Society of Korea, for the Computational Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics .