Graph based Version for Clustering Texts in Current Affair Domain

Taeho Jo
School of Game
Hongik University
Sejong, South Korea
tjo018@hongik.ac.kr

***Abstract*—This research proposes the graph based AHC algorithm where graphs are clustered depending on their similarities. The motivations for this research are the two facts: one is that the graphs are the visualized text representations as well as other types of data and the other is that various similarity metric between graphs were defined previously. In this research, we encode texts into graphs where vertices are given as words and edges are given as their semantic similarities, and modify the AHC algorithm into the graph based version. We adopt the clustering index which is based on both the intra-cluster similarity and the inter-cluster one as the evaluation metric, and evaluate the traditional version and the proposed version in clustering news articles. In the next research, we consider other kinds of graphs which represent texts and data structures for representing graphs in the implementation level.**

I. Introduction

The text clustering is referred to the process of segmenting a group of texts into subgroups of content based similar ones. Even if a word belongs to a text in the broad view, a text is an article which consists of more than one paragraph, so both of them should be distinguished from each other. In this research, we encode texts into graphs, define the similarity metric between them, and apply the AHC algorithm. The results from clustering texts enable the browsing for accessing to texts by adding the cluster naming. In this section, we describe briefly the motivation, the idea, and the validation of this research.

Let us consider the motivations for doing this research. The text clustering is necessary for organizing unlabeled texts into clusters and for automating preliminary tasks such as category predefinitions and sample labeled text allocations for doing the text categorization. The AHC algorithm is a simple approach to the data clustering for starting to modify machine learning algorithms. A graph is more graphical representation than a numerical vector for visualizing a word. This research is intended to avoid the poor discrimination from the sparse distribution in each numerical vector, by encoding them into alternative ones.

In this research, we propose the modified AHC version as the approach to the text clustering. The texts are encoded into graphs where nodes are words and edges are similarities between texts, and the similarity metric between them is defined. By adopting the proposed similarity metric, for computing the similarity between the test example and each training example, the AHC algorithm is modified. The modified version is applied to the text clustering which is covered in this research. The discriminations among numerical vectors are improved by encoding texts into alternative representations to numerical vectors.

In this research, we will validate empirically the proposed approach to the text clustering as the better version than the traditional AHC version. We use the collections of news articles: NewsPage.com and 20NewsGroups. The traditional AHC version and the proposed version are compared with each other. We observe the better results of the proposed AHC version in clustering texts. In addition, it is possible to represent words or texts more graphically from encoding them into graphs.

Let us mention the organization of this research. In Section II, we explore the previous works which are relevant to this research. In Section III, we describe in detail what we propose in this research. In Section IV, we validate empirically the proposed approach by comparing it with the traditional one. In Section V, we mention the significances of this research and the remaining tasks as the conclusion.

II. Previous Works

Let us survey the previous cases of encoding texts into structured forms for using the machine learning algorithms to text mining tasks. The three main problems, huge dimensionality, sparse distribution, and poor transparency, have existed inherently in encoding them into numerical vectors. In previous works, various schemes of preprocessing texts have been proposed, in order to solve the problems. In this survey, we focus on the process of encoding texts into alternative structured forms to numerical vectors. In other words, this section is intended to explore previous works on solutions to the problems.

Let us mention the popularity of encoding texts into numerical vectors, and the proposal and the application of string kernels as the solution to the above problems. In 2002, Sebastiani presented the numerical vectors are the standard representations of texts in applying the machine learning algorithms to the text classifications [4]. In 2002, Lodhi et al. proposed the string kernel as a kernel function of raw texts in using the SVM (Support Vector Machine) to the

text classification [5]. In 2004, Lesile et al. used the version of SVM which proposed by Lodhi et al. to the protein classification [6]. In 2004, Kate and Mooney used also the SVM version for classifying sentences by their meanings [7].

It was proposed that texts are encoded into tables instead of numerical vectors, as the solutions to the above problems. In 2008, Jo and Cho proposed the table matching algorithm as the approach to text classification [8]. In 2008, Jo applied also his proposed approach to the text clustering, as well as the text categorization [12]. In 2011, Jo described as the technique of automatic text classification in his patent document [10]. In 2015, Jo improved the table matching algorithm into its more stable version [11].

Previously, it was proposed that texts should be encoded into string vectors as other structured forms. In 2008, Jo modified the k means algorithm into the version which processes string vectors as the approach to the text clustering[12]. In 2010, Jo modified the two supervised learning algorithms, the KNN and the SVM, into the version as the improved approaches to the text classification [13]. In 2010, Jo proposed the unsupervised neural networks, called Neural Text Self Organizer, which receives the string vector as its input data [14]. In 2010, Jo applied the supervised neural networks, called Neural Text Categorizer, which gets a string vector as its input, as the approach to the text classification [15].

The above previous works proposed the string kernel as the kernel function of raw texts in the SVM, and tables and string vectors as representations of texts, in order to solve the problems. Because the string kernel takes very much computation time for computing their values, it was used for processing short strings or sentences rather than texts. In the previous works on encoding texts into tables, only table matching algorithm was proposed; there is no attempt to modify the machine algorithms into their table based version. In the previous works on encoding texts into string vectors, only frequency was considered for defining features of string vectors. In this research, we propose that texts should be encoded into graphs, and modify the AHC algorithm into the version which processes graphs instead of numerical vectors, as the approach to the text clustering.

III. Proposed Approach

This section is concerned with what we propose in this research. Words are encoded into graphs and the graph similarity is defined. The AHC algorithm is modified into the version which receives a group of graphs as clustering targets and computes the proposed similarity metrics among data items. The modified version is applied to the topic based text clustering. In this section, we describe what is proposed in this research.

Let us explain the process of encoding a text into a graph. The text is given as the encoding target and it is indexed into a list of words as its graph vertices. The similarities among the words are computed as edge candidates, and some among them are selected by their similarities as edges. In this process, a weighted undirected graph which consists of weighted edges is generated as the text representation. In the graph, the words are given as vertices and the semantic similarities among words are given as edges.

The two graphs are expressed as two edge sets, $G_1 = \{e_{11}, e_{12}, \cdots, e_{1n}\}$ and $G_2 = \{e_{21}, e_{22}, \cdots, e_{2n}\}$, assuming the identical number of edges in both of them, and an individual edge is expressed into an entry: $e_{ki} = (v_{ki1}, v_{ki2}, w_{ki})$. In computing the similarity between two edges, we consider the three cases: $sim(e_{1i}, e_{2j}) = \frac{1}{2}(w_{1i} + w_{2j})$ in case of both identical vertices, $sim(e_{1i}, e_{2j}) = w_{1i}w_{2j}$ in case of either of identical vertices, and $sim(e_{1i}, e_{2j}) = 0$, in case of no identical vertex. The similarity between an individual edge and a graph is computed by Equation (1),

$$sim(e_{1j}, G_2) = \max_{j=1} nsim(e_{1i}, e_{2j}) \quad (1)$$

The similarity between two graphs is computed by Equation (2),

$$sim(G_1, G_2) = \frac{1}{n} \sum_{i=1} nsim(e_{1j}, G_2) \quad (2)$$

The weight which is associated with each edge is always given as a normalized value between zero and one, so the product of two weights results in a reduced value, when either of two vertices in an edge is same to those of another edge.

The proposed AHC algorithm is presented in Figure 1. The proposed system encodes the words which are given as the clustering targets into graphs and starts with singletons as many as items. It computes the similarities of all possible cluster pairs by Equation(2) and merges the most similar clusters into a single cluster. It iterates the above steps until the number of clusters is reduced to the desired one. By discriminating the similarities and the attributes, we derive AHC variants from this version.

Let us make some remarks on what is proposed in this research. Even if the AHC algorithm is a very simple machine learning algorithm, it is useful for implementing a light version of classification system. Even if it takes much tome for computing the proposed similarity metric, it tackled against the poor discriminations from the sparse distribution of numerical vectors. We may use texts which are called associated ones as vertices as well as words for representing texts into graphs. The proposed AHC algorithm is described in more detail in [16].

IV. Experiments

This section is concerned with one more set of experiments where the better performance of the proposed version is validated on another version of 20NewsGroups.

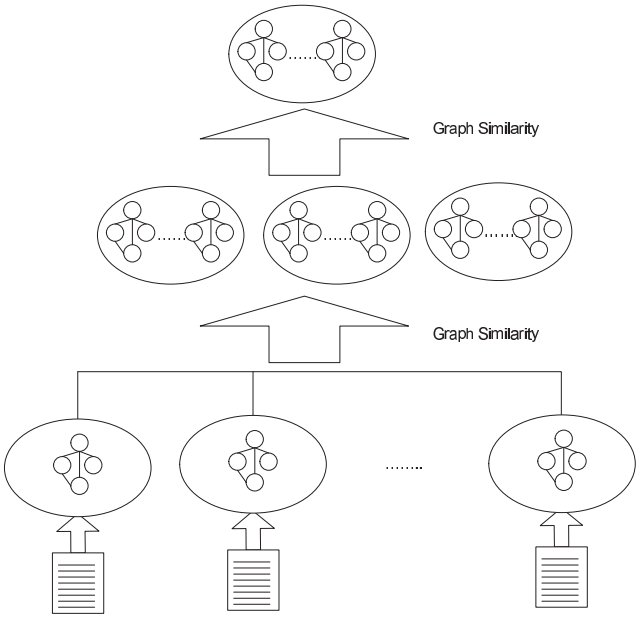

Figure 1. The Proposed Version of AHC Algorithm

In this set of experiments, the four specific categories are predefined in this collection. Texts are exclusively clustered into the four subgroups like the previous sets of experiments. We use the clustering index as the metric for evaluating clustering results. Therefore, in this section, we observe the performances of the both versions of AHC algorithm with the different input sizes.

In Table I, we specify the specific version of 20NewsGroups which is used as the test collection, in this set of experiments. Within the general category, sci, we predefine the four categories: 'electro', 'medicine', 'script', and 'space'. In each category, we select 300 texts among approximately 1000 texts, at random. We evaluate the results from clustering texts by the clustering index which is used as the evaluation metric, in the previous sets of experiments. We use the classified texts for evaluating the results, hiding their labels, while clustering texts.

Table I
THE NUMBER OF TEXTS IN 20NEWSGROUPS II

Category	#Texts	#Used Texts
Electro	1000	300
Medicine	1000	300
Script	1000	300
Space	1000	300
Total	4000	1200

The process of doing this set of experiments is same to that in the previous sets of experiments. We select the balanced number of texts from the collection over categories, and encode them into the representations with the input sizes which are identical to those in the previous set of experiments. Using the two versions of AHC algorithm, we cluster the 300 examples into the four clusters, identically to the previous set of experiments. We use the clustering index whose bases are the intra-cluster similarity and the inverse inter-cluster similarity, for evaluating the both versions of AHC algorithm. We evaluate the results from clustering items, using the labeled examples, following the external validity.

We present the experimental results from clustering the texts using the both versions of KNN algorithm on the specific version of 20NewsGroups. The frame of illustrating the classification results is identical to the previous ones. In each group, the gray bar and the black bar stand for the achievements of the traditional version and the proposed version, respectively. The y-axis in Figure 2, indicates the clustering index which is used as the performance metric. In clustering texts, each of them is allowed to belong to only one cluster like the cases in the previous sets of experiments.

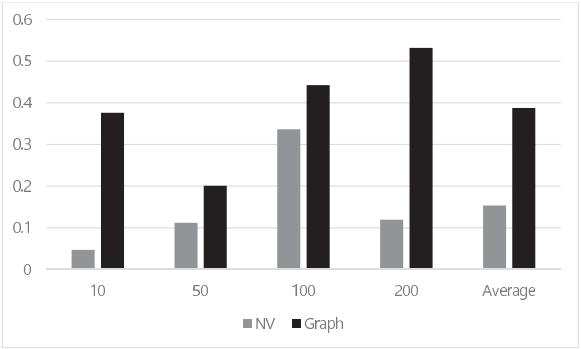

Figure 2. Results from Clustering Texts in Text Collection: 20NewsGroups II

Let us discuss on the results from clustering the texts on the specific version of 20NewsGroups, as shown in Figure 2. The accuracies of the both versions range between 0.05 and 0.52. The proposed version shows its outstandingly better performance in all of the four input sizes. It shows its strong difference in the input size, 200. From this set of experiments, it is concluded that the proposed version shows its better performance by averaging over the accuracies of the four cases.

V. CONCLUSION

Let us mention the remaining tasks for doing the further research. We apply and validate the proposed research in clustering technical documents in specific domains such as medicine or engineering rather than news articles in various domains. We define and characterize more advanced operations mathematically on graphs which represent texts. We modify more advanced machine learning algorithms into their graph based version, using the more sophisticated operations. We implement the text clustering system as a system module or an independent program by adopting the proposed approach.

VI. Acknowledgement

This work was supported by 2018 Hongik University Research Fund.

References

[1] T. Jo, *The Implementation of Dynamic Document Organization using Text Categorization and Text Clustering*, PhD Dissertation of University of Ottawa, 2006.

[2] N.F. Noy and C. D. Hafner, "State of the Art in Ontology Design", AI Magazine, Vol 18, No 3, 1997.

[3] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologies*, Mrgan Kaufmann, 2011.

[4] F. Sebastiani, "Machine Learning in Automated Text Categorization", pp1-47, ACM Computing Survey, Vol 34, No 1, 2002.

[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification with String Kernels", pp419-444, Journal of Machine Learning Research, Vol 2, No 2, 2002.

[6] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch String Kernels for Discriminative Protein Classification", pp467-476, Bioinformatics, Vol 20, No 4, 2004.

[7] R. J. Kate and R. J. Mooney, "Using String Kernels for Learning Semantic Parsers", pp913-920, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006.

[8] T. Jo and D. Cho, "Index based Approach for Text Categorization",International Journal of Mathematics and Computers in Simulation, Vol 2, No 1, 2008.

[9] T. Jo, "Single Pass Algorithm for Text Clustering by Encoding Documents into Tables", pp1749-1757, Journal of Korea Multimedia Society, Vol 11, No 12, 2008.

[10] T. Jo, "Device and Method for Categorizing Electronic Document Automatically", Patent Document, 10-2009-0041272, 10-1071495, 2011.

[11] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", pp839-849, Soft Computing, Vol 19, No 4, 2015.

[12] T. Jo, "Inverted Index based Modified Version of K-Means Algorithm for Text Clustering", pp67-76, Journal of Information Processing Systems, Vol 4, No 2, 2008.

[13] T. Jo, "Representationof Texts into String Vectors for Text Categorization", pp110-127, Journal of Computing Science and Engineering, Vol 4, No 2, 2010.

[14] T. Jo, "NTSO (Neural Text Self Organizer): A New Neural Network for Text Clustering", pp31-43, Journal of Network Technology, Vol 1, No 1, 2010.

[15] T. Jo, "NTC (Neural Text Categorizer): Neural Network for Text Categorization", pp83-96, International Journal of Information Studies, Vol 2, No 2, 2010.

[16] T. Jo, "Graph based AHC Algorithm for Text Clustering", accepted, The Proceedings of Computer Science and Computational Intelligence, 2018.