Clustering Words Semantically by Graph based Version of AHC Algorithm

Taeho Jo
President
Alpha AI Publication
Cheongju, South Korea
tjo018@naver.com

***Abstract*—This article proposes the modified AHC (Agglomerative Hierarchical Clustering) algorithm which clusters graphs, instead of numerical vectors, as the approach to the word clustering. The graph is more graphical for representing a word and the synergy effect between the text clustering and the word clustering is expected by combining them with each other. In this research, we propose the similarity metric between two graphs representing words, and modify the AHC algorithm by adopting the proposed similarity metric as the approach to the word clustering. The proposed AHC algorithm is empirically validated as the better approach in clustering words in news articles and opinions. In this article, a word is encoded into a weighted and undirected graph and it is represented into a list of edges.**

I. Introduction

Word clustering refers to the process of segmenting a group of words into subgroups of content based similar words. An arbitrary group of words is given as the input and they are encoded into their structured forms. We define the similarity measure between the structured forms of words and compute their similarities among them. The words are arranged into their own subgroups based on their similarities. In this research, we assume that the unsupervised learning algorithms are used as the approach to the word clustering.

We mention the facts which provide the motivations for doing this research. It caused much computation time to encode words into numerical vectors, because too many features are required for the robustness [2]. The sparse distribution in each numerical vector which represents a text or word as results from using too many features degrades the discriminations among string vectors [2]. Recently, it became the popular trend to encode the knowledge into graph called ontology [1][27]. Hence, in his research, motivated by the facts, we attempt to encode words into graphs for doing the word clustering task.

Let us mention what is proposed in this research as some ideas. In this research, each word is encoded into a graph which its vertices indicate text identifiers and its edges indicate the their semantic relations. We define the similarity measure between two graphs which is given as a normalized value between zero and one for clustering words by the AHC (Agglomerative Hierarchical Clustering) algorithm. We modify the AHC algorithm into the graph based version where a graph is directly given as input data, and apply it as the approach to the word clustering. This research provides the graphical representations of words as well as the solution to the above problems in encoding words into numerical vectors.

Let us mention some benefits which are expected from this research. We may expect the more semantic and graphical representations as shown by graphs from this research. We may also expect the strong discrimination among the proposed representations of words by removing completely the sparse distribution. We may expect the better word clustering performance by solving the problems from encoding words into numerical vectors. However, we need to define more operations on graphs, in order to modify more advanced machine learning algorithms.

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 significance of this research and the remaining tasks as the conclusion.

II. Previous Works

This section is concerned with the previous works which are relevant to this research. In Section II-A, we explore the previous cases of applying the AHC algorithm to text mining tasks. In Section II-B, we survey the schemes of encoding texts or words into structured data. In Section II-C, we describe the previous machine learning algorithms which receive alternative structured data such as tables and string vectors to numerical vectors. Therefore, in this section, we provide the history about this research, by surveying the relevant previous works.

A. Application to Clustering Tasks

This section is concerned with the previous works on applying the modernized AHC version to the word clustering and the text clustering. Even if the word and the article belong to a text in the wide view, they should be distinguished from each other in the specific view. The version of AHC algorithm was modernized by modifying it into the version which solves the problems in encoding words into numerical vectors. The word clustering and the text

clustering are typical tasks to which the modernized version of AHC algorithm is applied. This section is intended to explore the previous works as the cases of applying the type of AHC algorithm to both tasks.

Let us mention the previous works on applying the modernized AHC algorithm to the word clustering. The similarity metric between two vectors was modified into one which considers similarities among features, in applying the AHC algorithm to the word clustering [5]. The AHC algorithm which clusters tables directly, as the modernized version, was applied to the word clustering [7]. The AHC algorithm as the approach to the word clustering was modified into the version which clusters string vectors [8]. The literatures which were mentioned above deal with the modernized version of the AHC algorithm for improving the clustering performance.

The word clustering may be expanded into the text clustering as a task of text mining. The modernized AHC algorithm which uses the similarity metric considers the feature similarities was applied to the text clustering, as well as the word clustering [17]. Another type of modernized AHC algorithm which clusters table directly was adopted for implementing a text clustering system [18]. One more type of modernized AHC algorithm which clusters string vectors was considered as the approach to the text clustering [19]. The AHC algorithms which are modernized with their different directions were applied to both the word clustering and the text clustering.

The clustering index was used as the metric for evaluating clustering results in this study. It was initially mentioned for evaluating the dynamic document organization system in 2006 [2]. It was described in detail as the metric for evaluating clustering results, in 2007 [24]. In 2019, it was proposed for tuning parameters of clustering algorithms [20]. It is the metric into which the intra-cluster similarity and the inter-cluster discrimination are integrated.

Let us mention some points which make this research distinguishable from the above literatures. We explored the previous cases of applying the modernized versions of AHC algorithm with their different directions to the word clustering and the text clustering. We mentioned the historical notes about the clustering index which is used for evaluating the clustering results through some literatures. The proposed mention of AHC algorithm which is the approach to the word clustering, is one which clusters graphs representing words, directly, based on their similarities. In this study, we will apply the proposed version to the word clustering, in order to validate its performance.

B. Word and Text Encoding

This section is concerned with the previous works on encoding texts or words into other types of structured data. In previous works, some issues in encoding words and texts were discovered. The works challenged against the issues by encoding them into alternative structured forms to numerical vectors. In this section, we mentioned the tables, the string vectors, and the graphs, as structured data which are alternative to numerical vectors. This section is intended to explore the previous cases of encoding texts or words into one of the three types of structured forms.

Let us mention the previous cases of encoding texts or words into table for modernizing other machine learning algorithms. Words were encoded into tables in applying the KNN algorithm to the text categorization [11]. Words were encoded so, in doing it to the keyword extraction [12]. Texts were encoded into tables in doing it to the text categorization [13]. In the above literatures, texts and words were encoded into tables in using the KNN algorithm.

Let us consider the previous cases of encoding texts or words into string vectors. In modifying the KNN algorithm as the approach to the word categorization, words were encoded into string vectors [14]. In doing it as the approach to the keyword extraction, words were also encoded into string vectors [15]. Texts were encoded into string vectors for modifying the KNN algorithm as the approach to the text categorization [16]. The above literatures presented the previous cases of encoding raw data into string vectors.

Let us consider the previous works on encoding words or texts into graphs. It was proposed that words should be encoded into graphs, in using the KNN algorithm for classifying them [9]. Words were proposed to be encoded so in using it for extracting keywords [10]. It was proposed that texts were encoded into graphs in using it for classifying texts [21]. From the representative literatures, we present the previous cases of encoding raw data into graphs.

We mentioned the previous works about the three schemes of encoding words or texts for performing the text mining tasks. We adopt the third scheme where words are encoded into graphs in this study. We define the similarity matric between graphs and modify the AHC algorithm into the version which processes graphs directly. We apply the modified version of AHC algorithm for implementing the semantic word clustering system. We evaluate the modified version using the clustering index, comparing with the traditional version.

C. Non-Numerical Vector based Clustering Algorithms

This section is concerned with the previous works on the non-numerical vector based clustering algorithms. In the previous section, we presented the cases of encoding words or texts into non-numerical vectors. In this section, we mention the three clustering algorithms, the string kernel based clustering algorithm, the table matching algorithm, and the Neural Text Self Organizer, as the typical non-numerical vector based clustering algorithms. Because the word clustering which is covered in this research is relevant to the text clustering, in this section, we focus on the text clustering in surveying the previous works. This section is

intended to survey the previous works which propose one among the three algorithms as the approach to the text clustering.

Let us survey the previous works on proposing and using the string kernel. It was initially proposed for improving the SVM (Support Vector Machine) performance as the approach to the text categorization by Lodhi et al. in 2002 [26]. The k means algorithm was modified using the string kernel as the approach to the text clustering and implemented in R by Karatzoglou and Feinerer in 2006 [25]. The spectral algorithm was modified using the string kernel as the approach to the text clustering and validated empirically as the better approach than the traditional k means algorithm by Shi et al. in 2010 [28]. In the above literatures, the string kernel was utilized for improving the clustering algorithm performances, as well as the SVM one.

Let us explore the previous works on another kind of approach to the text categorization which is called table based matching algorithm. It was initially proposed by Jo and Cho in 2008 [22]. It was applied to the soft categorization of texts as the extended text categorization by Jo in 2008 [3]. It was improved into the more robust and stable approach by Jo in 2015 [6]. In using the table based matching algorithm which is mentioned in the above literatures, texts should be encoded into tables.

Let us mention the Neural Text Self Organizer as the neural network model in the style of the Kohonen Networks which was specialized for the text clustering. It was initially proposed as the approach to the text clustering by Jo and Japkowicz, in 2005 [23]. It was mentioned in surveying text clustering methods by Zheng et al. in 2006 [29]. Its better performance than the k means algorithm and the Kohonen Networks was confirmed in clustering texts in various domains in 2010 [4]. In using it, texts should be encoded into string vectors as non-numerical vectors.

We mentioned the three non-numerical vector based clustering and classification algorithms. In the string kernel based clustering algorithm, raw texts are used directly, in the table based matching algorithm, they are encoded into tables, and in the Neural Text Self Organizer, and they are encoded into string vectors. In this research, words are encoded into graphs; it is different from the cases in the above approaches. The AHC algorithm is modified into the version which clusters graphs directly. Its performance will be validated in the clustering, compared with the traditional one.

III. Proposed Approach

This section is concerned with encoding words into graphs, modifying the AHC (Agglomerative Hierarchical Clustering) algorithm into the graph based version and applying it to the word clustering, and consists of the three sections. In Section III-A, we deal with the process of encoding words into graphs. In Section III-B, we describe formally the process of computing the similarity between to graphs. In Section III-C, we do the graph vector based AHC version as the approach to the word clustering, and in Section III-D, present the architecture of the system which we try to implement by adopting the proposed AHC algorithm. Therefore, this section is intended to describe the proposed AHC version as the word clustering tool.

A. Word Encoding

This section is concerned with the process of mapping words into graphs. We presented the previous cases of encoding words or texts into graphs in Section II-B. The three steps which are presented in Figure 1-3, are involved in encoding words into graphs. In each graph, the vertex set indicate text identifiers which are related with the word, and the edge set indicates similarities among texts. This section is intended to describe the three steps which are involved in the word encoding.

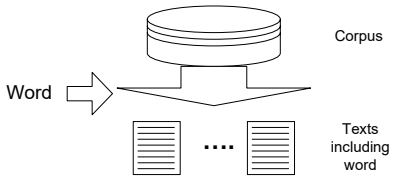

Figure 1. Word Indexing

The process of defining the vertex set in encoding a word into a graph is illustrated in Figure 1. The corpus as the source and a word as an input are initially given. Texts which include the word are extracted from the corpus as

a list of vertices. Only some among extracted texts with higher weights of the word are selected, if too many texts are extracted. The criteria for selecting texts as vertices becomes the issue in this step.

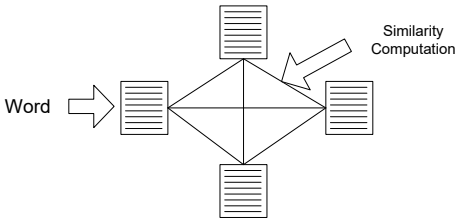

Figure 2. Word Representation: Graph

Edges in representing a word into a graph are illustrated in Figure 2. In the corpus, texts which are related with the word are given as vertices of the graph which represents it. All possible edges are generated as the complete links and some with their weaker similarities are eliminated. Each edge is labeled with its own weight which indicates the similarity between two texts which are vertices. The number of edges is controlled, depending on the number of vertices.

The process of computing a weight for each edge in the complete links among vertices is illustrated in Figure 3. The vertices are assumed to be N texts which are relevant to the word, and $N \times N$ similarity matrix whose columns and rows correspond to the texts. The diagonal elements are filled with 1.0 and the off-diagonal elements are filled with the similarities between two texts which are given as normalized values between 0 and 1.0 in the similarity metric. Each text is viewed as a set of words as results from indexing it and the similarity between texts is computed based on their

$$\begin{array}{c} \\ \text{Text 1} \\ \text{Text 2} \\ \text{....} \\ \text{Text N} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Text 1} & \text{Text 2} & \text{....} & \text{Text N} \end{array} \\ \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{bmatrix} \end{array} \qquad s_{ij} = \frac{2\left|Text_i \cap Text_j\right|}{\left|Text_i\right| + \left|Text_j\right|}$$

Figure 3. Similarity Matrix

intersection. The values except diagonal elements are given as weights of the edge candidates, some with higher weights are selected as real edges.

In this section, we described the three steps which are involved in encoding a word into a graph, as presented in Figure 1-3. In the graph which represents a word, the vertex set is given as a set of text identifiers which is represent to it, and the edge set is given as a set of similarities among text identifiers. Each graph is interpreted into a set of edges each of which consists of the two text identifiers and the weight. The weight which is assigned to each edge as the similarity between two texts is always given as a normalized value between zero and one. We need to define the operations on graphs for modifying the machine learning algorithms into versions which process them directly.

B. Graphs Similarity

This section is concerned with the similarity metric between two graphs. In the previous section, we studied the process of converting words into graphs. We need to define the similarity metric between two graphs for modifying the AHC algorithm into the version which clusters graphs

directly. We view a graph as an edge set and start with defining the similarity between two edges. This section is intended to describe the similarity metric between two graphs.

Let us mention the computation of similarity between two edges as the basis for computing one between two graphs. Each edge is expressed as an entry of three values as shown in equation (1),

$$e \equiv (node_1, node_2, weight) \tag{1}$$

the two edges, e_1 and e_2 are expressed as equation (2) and (3),

$$e_1 = (node_{11}, node_{12}, weight_1) \tag{2}$$

$$e_2 = (node_{21}, node_{22}, weight_2) \tag{3}$$

and $weight_1$ and $weight_2$ are given as normalized values between zero and one.

We consider the three possible cases between two edges as both nodes are same to each other, either of them is same, and neither of them is so. The similarity between two edges is defined on the three conditions:

- In the two edges, if both nodes are same to each other, the similarity between them is defined by equation(4),

$$sim(e_1, e_2) = \frac{1}{2}(weight_1 + weight_2) \tag{4}$$

- In the two edges, if only either of two nodes are same to each other, the similarity between them is defined by equation(5),

$$sim(e_1, e_2) = (weight_1 \times weight_2) \tag{5}$$

- In the two edges, if no node are same to each other, the similarity between them is zero.

The edge similarity will be used for computing the similarity between an edge and a graph, next.

The similarity between two edges is expanded into one between an edge and a graph. The graph, G is expressed as a set of edges, $G = \{e_1, e_2, \ldots, e_{|G|}\}$. The similarity, $sim(e_i, G)$, is computed by equation (6),

$$sim(e_i, G) = \max_{k=1}^{G} sim(e_i, e_k) \tag{6}$$

The similarity between an edge and a graph is the maximum among its similarities with ones in the graph, as shown in equation (6). When only edge e_r in the graph, G, have both identical, assuming that all weights are constant between zero and one, the similarity between an edge and a graph is expressed by equation (7),

$$sim(e_i, G) = sim(e_i, e_r) \tag{7}$$

The similarity between an edge and a graph is expanded into one between two graphs,further. The two graphs are notated by G_1 and G_2, and they are viewed as edge sets, as shown in equation (8) and (9),

$$G_1 = \{e_{11}, e_{12}, \ldots, e_{1|G_1|}\} \tag{8}$$

$$G_2 = \{e_{21}, e_{22}, \ldots, e_{2|G|_2}\} \tag{9}$$

For each edge in the graph, G_1, its similarity with the graph, G_2 is computed by equation (6). The similarity between the two graphs, G_1 and G_2 is computed by equation (10),

$$sim(G_1, G_2) = \frac{1}{|G_1|} \sum_{i=1}^{|G_1|} sim(e_{1i}, G_2) \tag{10}$$

The similarity between the two graphs, G_1 and G_2, is always given as a normalized value between zero and one.

We mentioned the similarity between two graphs as a normalized value between zero and one, and let us assume that all edges are weighted as 1.0 in both graphs. If $G_1 = G_2$, the similarity between two graphs is given as 1.0 by equation (11),

$$\begin{aligned} sim(e_{1i}, G_2) &= 1.0 \\ sim(G_1, G_2) = \frac{1}{|G_1|} \sum_{i=1}^{G_1} sim(e_{1i}, G_2) &= \frac{|G_1|}{|G_1|} = 1.0 \end{aligned} \tag{11}$$

If no vertex shared by two graphs, the similarity between two graphs given as zero by equation (12),

$$\begin{aligned} sim(e_{1i}, G_2) &= 0.0 \\ sim(G_1, G_2) = \frac{1}{|G_1|} \sum_{i=1}^{G_1} sim(e_{1i}, G_2) &= \frac{0}{|G_1|} = 0.0 \end{aligned} \tag{12}$$

The similarity between two graphs is always given as a normalized value between zero and one by equation (13),

$$\begin{aligned} G_1 \cap G_2 \subseteq G_1, G_1 \cap G_2 \subseteq G_2 \\ 0 \leq sim(e_{1i}, G_2) \leq 1.0 \\ 0 \leq \frac{1}{|G_1|} \sum_{i=1}^{G_1} sim(e_{1i}, G_2) \leq 1.0 \\ 0 \leq sim(G_1, G_2) \leq 1.0 \end{aligned} \tag{13}$$

Each edge is usually weighted between zero and one, so the similarity between two graphs is clearly given as a normalized value.

C. Proposed Version of AHC Algorithm

This section is concerned with the proposed version of the AHC algorithm which is shown in Figure 4, as the approach to the semantic word clustering. We described the process of encoding words into graphs in Section III-A, and assume that items in the group are given as graphs. The similarity metric which is described in Section III-B is used for computing the similarity between clusters in executing the AHC algorithm.

Variants may be derived by considering various schemes of merging clusters and computing the cluster similarities, in addition. This section is intended to describe the AHC algorithm which clusters graphs directly, and its variants.

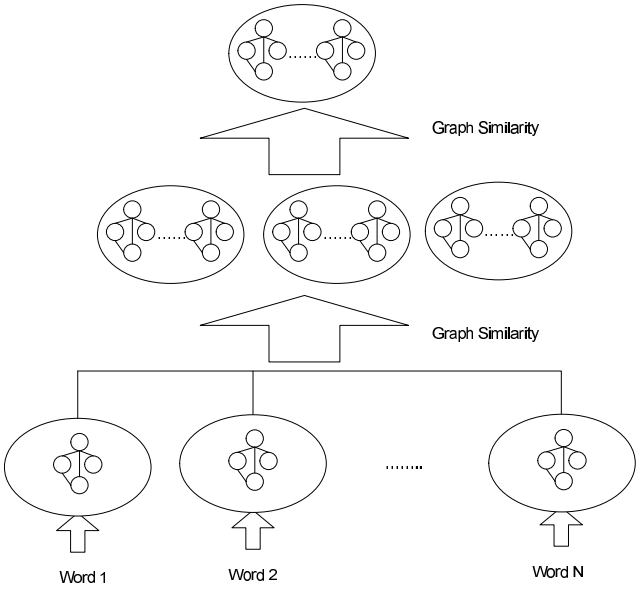

Figure 4. Proposed Version of AHC Algorithm

Let us mention the computation of the similarity between two clusters. The two clusters are notated by sets of graphs:$C_1 = \{G_{11}, G_{12}, \ldots, G_{1|C_1|}\}$ and $C_2 = \{G_{21}, G_{22}, \ldots, G_{2|C_2|}\}$. All possible pairs of graphs are generated from the two clusters, and for each pair, its similarity is computed by the equation which was defined in Section III-B. The similarity between the two clusters is computed by equation (14),

$$sim(C_1, C2) = \frac{1}{|C_1||C_2|} \sum_{i=1}^{|C_1|} \sum_{j=1}^{|C_2|} sim(G_{1i}, G_{2j}) \quad (14)$$

The similarity between two graphs is always given as a normalized value between zero and one, so the similarity between two clusters which is computed by equation (14) is also given as a normalized value.

Let us mention the process of clustering data items by the AHC algorithm. The tables which are mapped from words in the group are notated by the set, $\{G_1, G_2, \ldots, G_N\}$, and the set of initial clusters is expressed as $\{C_1^1, C_2^1, \ldots, C_{N_1}^1\}$, where $C_i = \{G_i\}$, the super script 1 means the initial iteration, and $N_1 = N$ which is the number of clusters in the first iteration. All possible pairs of clusters, $Pair(C_i^k, C_j^k), i < j$, are generated, and the similarity between two clusters $sim(C_i^k, C_j^k)$ is computed for each pair by equation (14). Clusters in the pair with the maximal similarity are merged into a cluster as shown in equation (15),

$$\begin{aligned} Pair_{\max}(C_i^k, C_j^k) &= \underset{i<j}{\overset{N_k}{\operatorname{argmax}}}\, sim(C_i^k, C_j^k) \\ C_{merge}^{k+1} &= merge(Pair_{\max}(C_i^k, C_j^k)) \end{aligned} \quad (15)$$

and the number of clusters in the $k+1$ th iteration is $N_{k+1} = N_k - 1$ by decrementing the number of clusters by merging it. The AHC algorithm proceeds clustering by iterating the computation of similarities between clusters in all possible pairs and merge of pair with the maximal similarity into one cluster.

Let us mention the clustering index which is used for evaluating the traditional version and the proposed one of the AHC algorithm. The intra-cluster similarity of the cluster, C_i, and the inter-cluster similarity of the two clusters, C_i and C_j are notated respectively by $intra\_sim(C_i)$ and $inter\_sim(C_i, C_j)$ and the clustering results are expressed as a set of clusters, $C = \{C_1, C_2, \ldots, C_{|C|}\}$. The intra-cluster similarity over the clustering results,C, is computed by equation (16),

$$intra\_sim(C) = \frac{1}{|C|} \sum_{i=1}^{|C|} intra\_sim(C_i) \quad (16)$$

and the inter-cluster similarity over entire cluster, C is computed by equation (17),

$$inter\_sim(C) = \frac{2}{|C|(|C|-1)} \sum_{i<j}^{|C|} inter\_sim(C_i, C_j) \quad (17)$$

The clustering index is computed by equation (18),

$$CI(C) = \frac{2 \cdot intra\_sim(C) \cdot (1 - inter\_sim(C))}{intra\_sim(C) + (1 - inter\_sim(C))} \quad (18)$$

The desired goal of clustering data items is to maximize the intra cluster similarity and minimize the inter cluster similarity.

We described the proposed version of the AHC algorithm as the approach to the data clustering. Raw data is encoded into graphs for using the proposed version for clustering data items. We use the similarity metric between graphs for computing similarities among items. The similarity between clusters is the average over all possible similarities of data items. The desired number of clusters is set as the termination condition in proceeding clustering by the AHC algorithm.

D. Word Clustering System

This section is concerned with the semantic word clustering system which adopts the graph based AHC algorithm. In Section III-C, we described the proposed version of AHC algorithm which clusters graphs directly. Words are encoded into graphs and clustered into subgroups as the main functions. Clustering data items is executed by iterating

computing the similarities among clusters and merging clusters. This section is intended to describe the word clustering system with respect to its functions and architecture.

The words are gathered as clustering targets. Because unsupervised learning algorithms are used for clustering data, the words are assumed to be unlabeled. The words are encoded into tables by the process which was mentioned in Section III-A. The similarity metric which is described in Section III-B is defined and the AHC algorithm which is described in Section III-C is adopted as the clustering method. The number of clusters should be set as the termination condition in the system.

The entire architecture of the proposed word clustering system is illustrated in Figure 5. All words which are given as the input are encoded into graphs. They are clustered by the AHC algorithm which was described in Section III-C in the similarity computation module and the clustering module. The graph clusters are restored into the word clusters by the decoder. There are the four modules in the system: the encoding module, the similarity computation module, the clustering module, and the decoding module.

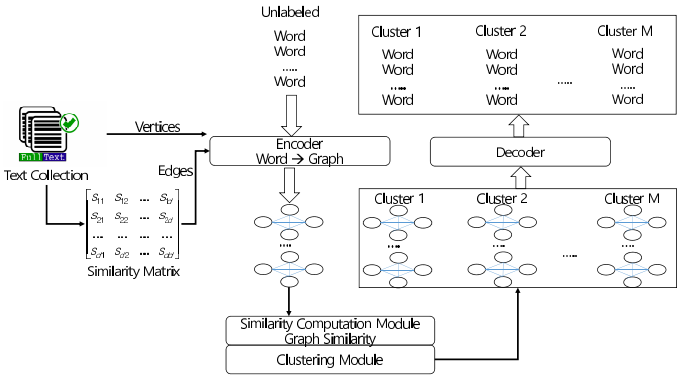

Figure 5. Proposed System Architecture

The execution process of the proposed system is illustrated as a block diagram in Figure 6. The words which are clustered are encoded into graphs by the encoding module. The graphs are clustered by the AHC algorithm by iterating computing the similarity among clusters and merging clusters. Clusters each of which contain semantic similar words are given as the final output in the system. In advance we need to decide the number of clusters as an external parameter.

Let us make some remarks on the proposed system which is illustrated in Figure 5 as the architecture. Words are encoded into graphs, instead of numerical vectors. Graphs which represent words are clustered by the proposed AHC algorithm, directly. The clustering performance is improve by what is proposed in this research as shown in Section IV. In the next research, we present the graphical user interface and the source codes which are necessary for implementing the system as a complete one.

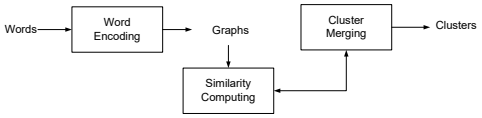

Figure 6. Execution Process of Proposed System

IV. Experiments

This section is concerned with the empirical experiments for validating the proposed version of AHC algorithm, and consists of the five sections. In Section IV-A, we present the results from applying the proposed version of AHC to the word clustering on the collection, NewsPage.com. In Section IV-B, we show the results from applying it for clustering words from the collection, Opinosis. In Section IV-C and IV-D, we mention the results from comparing the two versions of AHC algorithm with each other in clustering words from 20NewsGroups.

A. *NewsPage.com*

This section is concerned with the experiments for validating the better performance of the proposed version on the collection: NewsPage.com. We set the number of clusters as four, following the number of categories for evaluating the performance, and gather words from the collection, category by category, as the labeled ones. In the clustering process, each word is arranged into one of the four clusters, exclusively, in this set of experiments. We use the clustering index which was proposed in [2] for evaluating the clustering performances. Therefore, this section is intended to observe the performance of the traditional and proposed versions of AHC algorithm with different input sizes.

In Table I, we specify NewsPage.com as the text collection which is used as the source for extracting classified words, in this set of experiments. The text collection, NewsPage.com, was also used for evaluating approaches to text

categorization, in previous works [5]. We extract the 300 important words from each topic for building the collection of classified words for evaluating the approaches to word clustering. We segment the entire collection which consists totally of 1200 words into the four subgroups, depending on their semantic similarities. In each category, words are selected by their frequencies concentrated on the given topic combined with subjectivity, from the text collection.

Table I
THE NUMBER OF TEXTS AND WORDS IN NEWSPAGE.COM

Category	#Texts	#Words
Business	500	300
Health	500	300
Internet	500	300
Sports	500	300
Total	2000	1200

Let us mention the experimental process for validating empirically the proposed approach to the task of word clustering. We extract the important words from each category in the above text collection, and encode them into numerical vectors and graphs. The 1200 examples are clustered into the four clusters by the both versions of AHC algorithm. We use the clustering index which combines the two measures, the intra-cluster similarity and the inter-cluster similarity, for evaluating the both versions. The clustering index is described in detail in [24], and used previously for evaluating the clustering algorithms [2].

In Figure 7, we illustrate experimental results from clustering words using the both versions of AHC algorithm. The y-axis indicate the clustering index and is the measure for evaluating the clustering results. In the x-axis, each group indicates the input size as the dimension of numerical vectors which represent words. In each group, the gray bar and the black bar indicate the results of the traditional version and the proposed version of AHC algorithm, respectively. The most right group in Figure 7 indicates the average aver the results of the left four groups.

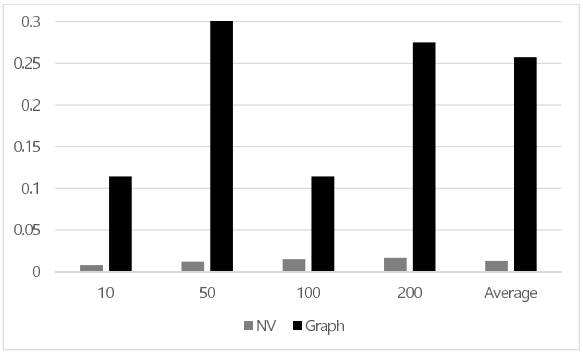

Figure 7. Results from Clustering Words in Text Collection: NewsPage.com

Let us make the discussions on the results from doing the word clustering, using the both versions of AHC algorithm, as shown in Figure 7. In the proposed version of AHC algorithm, the clustering index which is the performance measure of these clustering tasks is in the range between 0.1 and 0.3. The proposed version of the AHC Algorithm works much better in the all input sizes, as shown in Figure 7. The reason of the better performance is the improved discriminations among representations of words, by encoding words into graphs as alternative structured forms to numerical vectors. From this set of experiments, we conclude that the proposed version works much better than the traditional one, in averaging over the four cases.

B. Opinopsis

This section is concerned with the set of experiments for validating the better performance of the proposed version: Opniopsis. In this set of experiments, the three categories are predefined in the collection, and we collect words category by category as the classified ones. A group of words is exclusively segmented into the three clusters. In this set of experiments, we also use the clustering index. Therefore, in this section, we observe the performances of the both versions of AHC algorithm with the different input sizes on another collection.

In Table II, we illustrate the text collection, Opinosis, which is used as the source for extracting the classified words, in this set of experiments. The collection, Opinosis, was used in previous works for evaluating approaches to text categorization. We extract the 300 important words from each topic as the collection of classified words, for evaluating the approaches to word clustering. The group of totally 900 words is segmented into the three subgroups by the clustering algorithms, according to the number of the predefined categories. The words are extracted by both their frequencies which are concentrated in their own categories, in this set of experiments.

Table II
THE NUMBER OF TEXTS AND WORDS IN OPINIOPSIS

Category	#Texts	#Words
Car	23	300
Electronic	16	300
Hotel	12	300
Total	51	900

We perform this set of experiments by the process which is described in section IV-A. We extract the 300 important words by scanning individual texts in each category, and encode them into numerical vectors and graphs, with the input sizes: 10, 50, 100, and 200. The group of total 900 examples is clustered by the both versions of AHC algorithm into the three clusters, using the cosine similarity and the proposed one. In this set of experiments, we use also the clustering index which combines the intra-cluster similarity and the inverse inter-cluster similarity with each

other, for evaluating the both versions. We adopted the external evaluation where the labeled examples are used for evaluating clustering algorithms which is mentioned in [2].

In Figure 8, we illustrate the experimental results from clustering words using the both versions of AHC algorithm. Like Figure 7, the y-axis indicates the value of clustering index, and x-axis indicates the group of the two versions of AHC algorithm by an input size. In each group, the grey bar and the black bar indicate the achievements of the traditional version and the proposed on of AHC algorithm. In Figure 8, the most right group indicates the averages over the achievements of both versions of the left four groups. Therefore, Figure 8 shows the results from clustering words into the three subgroups by both versions, on the collection: Opinosis.

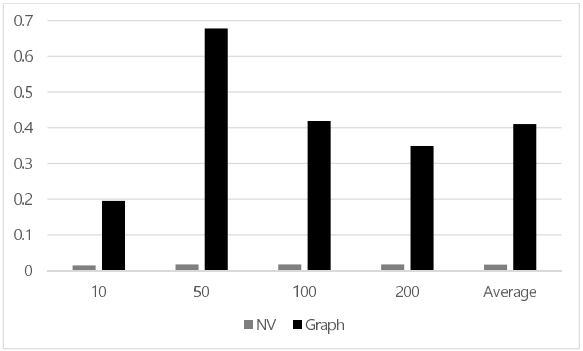

Figure 8. Results from Clustering Words in Text Collection: Opiniopsis

We discuss the results from doing the word clustering, using the both versions of AHC algorithm, on Opinosis, shown in Figure 8. The values of clustering index of both versions range between less than 0.1 and 0.7. The proposed version of AHC algorithm works better than the traditional ones in all input sizes. The reason of its better performance is the improved discriminations among graphs as alternative representations of words to numerical vectors. From this set of experiments, we conclude that the proposed one works outstandingly better in averaging over the four cases.

C. 20NewsGroups I: General Version

This section is concerned with one more set of experiments for validating empirically the better performance of the proposed version on the text collection: 20NewsGroups I. In this set of experiments, we predefine the four general categories and gather words from the collection category by category as the classified ones. The task of in this set of experiments is to cluster the gathered words into the four clusters based on their semantic similarities, exclusively. The both versions of AHC algorithm are evaluated by the clustering index, like the previous set of experiments. Therefore, in this section, we observe the performances of the both versions with the different input sizes.

In Table III, we specify the general version of 20NewsGroups which is used for evaluating the two versions of AHC algorithm. In 20NewsGroup, the hierarchical classification system is defined with the two levels; in the first level, the six categories, alt, comp, rec, sci, talk, misc, and soc, are defined, and among them, the four categories are selected, as shown in Table III. In each category, we select 1000 texts at random, and extract 300 important words from them as the labeled words. In the process of gathering the classified words, they are selected by their frequencies which are concentrated in their corresponding categories. Therefore, following the external evaluation, we use the classified words for evaluating clustering results.

Table III
THE NUMBER OF TEXTS AND WORDS IN 20NEWSGROUPS I

Category	#Texts	#Words
Comp	1000	300
Rec	1000	300
Sci	1000	300
Talk	1000	300
Total	4000	1200

The experimental process is identical is that in the previous sets of experiments. In each category, we extract the 300 important words and encode them into numerical vectors and graphs with the input sizes, 10, 50, 100, and 200. The totally 1200 words are clustered by the two versions of AHC algorithm, based on their similarities. We use the clustering index which combines the intra-cluster similarity and the inverse inter-cluster similarity with each other, for evaluating the both versions, identically to the previous sets of experiments. We use the labeled words and their target labels are hidden during clustering process.

In Figure 9, we illustrate the experimental results from clustering the words using the both versions of AHC algorithm on the broad version of 20NewsGroups. Figure 9 has the identical frame of presenting the results to those of Figure 7 and 8. In each group, the gray bar and the black bar indicates the achievements of the traditional version and the proposed version of AHC algorithm, respectively. This figure presents the results from clustering words into the four clusters by changing their input sizes. We adopt the external evaluation as the paradigm of evaluating the clustering results, in this set of experiments.

Let us discuss the results from doing the word clustering using the both versions of AHC algorithm on the broad version of 20NewsGroups, as shown in Figure 9. The clustering indices of the both versions range between less than 0.1 and 0.3. The proposed version shows the much better results in all of the input sizes. The reason of the better results is the improved discrimination among word representations. From this set of experiments, we conclude the proposed version win completely over the traditional one, in averaging their four achievements.

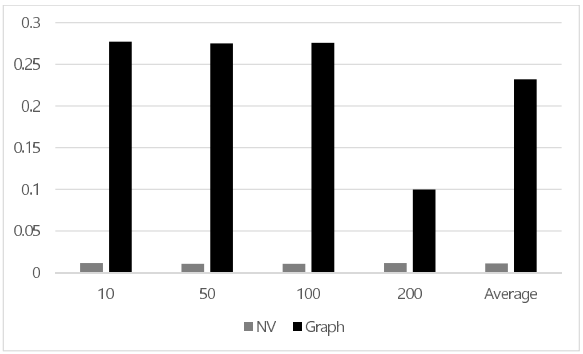

Figure 9. Results from Clustering Words in Text Collection: 20NewsGroup I

D. 20NewsGroups II: Specific Version

This section is concerned with one more set of experiments where the better performance of the proposed version is validated on another different version of 20NewsGroups. In this set of experiments, the four specific categories are predefined and words are gathered from each topic as the classified ones. The task of this set of experiments is to cluster exclusively words into four clusters. We use the clustering index like the previous sets of experiments as the evaluation metric. Therefore, in this section, we observe the performances of the both versions of AHC algorithm, with the different input sizes.

In Table 4, we specify the second version of 20NewsGroups which is used in this set of experiments. Within the general category, sci, the four categories, electro, medicine, script, and space, are predefined. We build the collection of labeled words by extracting the 300 important words from approximately 1000 texts in each specific category. In this set of experiments, the group of 1,200 words is clustered into the four groups. We use the classified words for evaluating the results from clustering them, like the case in the previous set of experiments.

Table IV
THE NUMBER OF TEXTS AND WORDS IN 20NEWSGROUPS II

Category	#Texts	#Words
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 extract the identical number of words from all texts in each category, and encode them into numerical vectors. We cluster 1200 words by the two versions of AHC algorithm into the four clusters. We use the clustering index based on the intra-cluster similarity and inverse inter-cluster similarity, for evaluating the both versions. We evaluate the results from clustering items, using the labeled examples, following the external validity.

We present the experimental results from clustering the words using the both versions of AHC 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 10, indicates the clustering index which is used as the performance metric. In clustering words, each of them is allowed to belong to only one cluster like the cases in the previous sets of experiments.

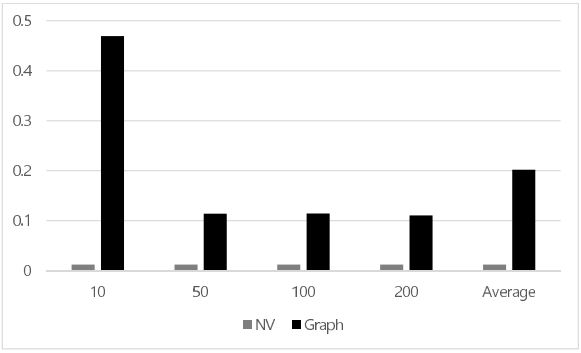

Figure 10. Results from Clustering Words in Text Collection: 20NewsGroup II

Let us discuss the results from clustering the words using the both versions of AHC algorithm on the specific version of 20NewsGroups, as shown in Figure 10. The clustering indices of both versions range between less than 0.1 and 0.47. The proposed version shows its strongly better performances in the all input sized, as shown in Figure 10. The reason of the better performances is the discriminations among feature vectors which is improved by encoding words into graphs, instead of numerical vectors. From this set of experiments, it is concluded that the proposed version of AHC algorithm is much feasible to the task of word clustering.

V. CONCLUSION

Let us discuss the entire results from clustering word using the two versions of AHC algorithm. In these sets of experiments, the traditional and proposed version are compared with each other in the tasks of word clustering. The proposed version shows the better results in all of the four collections. The clustering indices of the traditional version is always less than 0.1, while those of the proposed version range between 0.1 and 0.68. Through the four sets of experiments, we conclude that the proposed version improve the word clustering performance very strongly as the contribution of this research.

Let us mention some remaining tasks for doing the further research. We need to validate more the proposed approach in

clustering words in specific domains such as medicine, engineering, and economics, and customize it correspondingly. We need to consider other schemes of encoding words into graphs and other similarity measures between graphs. We modify other machine learning algorithms into their graph based versions where a graph is given by itself as the input data. We implement a word clustering system by adopting the proposed approach.

REFERENCES

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

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

[3] T. Jo, "Table based Matching Algorithm for Soft Categorization of News Articles in Reuter 21578", 875-882, Journal of Korea Multimedia Society, Vol 11, No 6, 2008.

[4] T. Jo, "NTSO (Neural Text Self Organizer): A New Neural Network for Text Clustering", 31-43, Journal of Network Technology, Vol 1, No 1, 2010.

[5] T. Jo, "AHC based Clustering considering Feature Similarities", 67-70, The Proceedings of 11th International Conference on Data Mining, 2015.

[6] T. Jo, "Normalized Table Matching Algorithm as Approach to Text Categorization", 839-849, Soft Computing, Vol 19, No 4, 2015.

[7] T. Jo, "Table based AHC Algorithm for Clustering Words", 574-579, The Proceedings of 18th International Conference on Advanced Communication Technology, 2016.

[8] T. Jo, "String Vector based AHC as Approach to Word Clustering", 133-138, The Proceedings of 12th International Conference on Data Mining, 2016.

[9] T. Jo, "Graph based KNN for Content based Word Classification", 24-29, The Proceedings of 12th International Conference on Multimedia Information Technology and Applications, 2016.

[10] T. Jo, "Extracting Keywords by Graph based KNN", 96-101, The Proceedings of 12th International Conference on Multimedia Information Technology and Applications, 2016.

[11] T. Jo, "Table based K Nearest Neighbor for Word Categorization in News Articles", 1214-1217, The Proceedings of 25th International Conference on Computational Science & Computational Intelligence, 2018.

[12] T. Jo, "Keyword Extraction in News Articles using Table based K Nearest Neighbors", 1230-1233, The Proceedings of 25th International Conference on Computational Science & Computational Intelligence, 2018.

[13] T. Jo, "Modification into Table based K Nearest Neighbor for News Article Classification", 49-50, The Proceedings of 1st International Conference on Advanced Engineering and ICT-Convergence, 2018.

[14] T. Jo, "Modification of K Nearest Neighbor into String Vector based Version for Classifying Words in Current Affairs", 72-75, The Proceedings of International Conference on Information and Knowledge Engineering, 2018.

[15] T. Jo, "Modifying K Nearest Neighbor into String Vector based Version for Extracting Keywords from News Articles", 43-46, The Proceedings of International Conference on Applied Cognitive Computing, 2018.

[16] Taeho Jo, "Improving K Nearest Neighbor into String Vector Version for Text Categorization", pp 1091-1097, ICACT Transaction on Communication Technology, Vol 7, No 1, 2018.

[17] T. Jo, "Feature Similarity AHC Algorithm for Clustering News Articles", 49-54, The Proceedings of 2nd International Conference on Advanced Engineering and ICT-Convergence, 2019.

[18] T. Jo, "Applying Table based AHC Algorithm to News Article Clustering", 8-11, The Proceedings of International Conference on Green and Human Information Technology, Part I, 2019.

[19] T. Jo, "Introduction of String Vectors to AHC Algorithm for Clustering News Articles", 150-153, The Proceedings of 21st International Conference on Artificial Intelligence, 2019.

[20] T. Jo, "Text Mining: Concepts and Big Data Challenge", Springer, 2019.

[21] T. Jo, "Graph based Version of K Nearest Neighbor for classifying News Articles", 4-7, The Proceedings of International Conference on Green and Human Information Technology Part I, 2019.

[22] T. Jo and D. Cho, "Index Based Approach for Text Categorization", 127-132, International Journal of Mathematics and Computers in Simulation, Vol 2, No 1, 2007.

[23] T. Jo and N. Japkowicz, "Text Clustering using NTSO", 558-563, The Proceedings of Internaitonal Joint Conference on Neural Networks, 2005.

[24] T. Jo and M. Lee, "The Evaluation Measure of Text Clustering for the Variable Number of Clusters", 871-879, Lecture Notes in Computer Science, Vol 4492, 2007.

[25] A. Karatzoglou and I. Feinerer, "Text Clustering with String Kernels in R", 91-98, Advances in Data Analysis, 2006.

[26] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification with String Kernels", 419-444, Journal of Machine Learning Research, Vol 2, No 2, 2002.

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

[28] Q. Shi, X. Qiao, and X. Guangquan, "Using String Kernel for Document Clustering", pp40-46, I.J. Information Technology and Computer Science, Vol 2, 2010.

[29] Y. Zheng, X. Cheng, R. Huang, and Y. Man, "A Comparative Study on Text Clustering Methods", 644-651, Advanced Data Mining and Applications, 2006.