International Journal of Machine Learning and Computing, Vol. 8, No. 3, June 2018

The Classification of the Documents Based on Word2Vec and 2-Layer Self Organizing Maps

Koki Yoshioka and Hiroshi Dozono

***Abstract*—Due to popularization of SNS and increase of use of WEB, people have to deal with large number of text data. However, it is difficult to process huge text data manually. For this problem, the classification methods based on machine learning is considered to be applicable. As a method of document classification, WEBSOM and its variations can visualize the relations among the documents as the similar documents are classified closely on the 2 dimension plane, and they will present good usability to the user because of their visualization ability. In this paper, the document classification method based on SOM and Word2Vec model, which can reduce the computational costs as to be executed on personal computers and can enhance the visualization ability. The performance of the proposed method is examined in the experiments using general collection of documents, and DNA sequences as the sample data.**

***Index Terms*—Self Organizing Map (SOM), Word2Vec, documents classification.**

I. Introduction

Recently, the cases of operating text data because of the increasing of SNS logs and WEB documents. It is difficult to operate the such increasing massive amount of text data by human. In the feature, the amount of data is expected to increase more and more, it is necessary to let the machine to classify the text data without human.

It is necessary to convert the text data to the numerical data which can be handled by machine easily to classify the text data using machine. As a typical method of text conversion, BoW(Bag of Word) model [1] is used to convert text data to feature vectors. However, the dimension of the vector becomes very large especially for large amounts of documents because one element of the vector is assigned to each word in the documents. Thus, it is necessary to compress the dimension of the vector. In this paper, the novel method for converting a document to a vector which resembles the text conversion of BoW model using Word2Vec [2] which is applied to natural language processing recently and Self Organzing Map (SOM) [3]. And the result of the conversion is examined in the clustering experiments using SOM.

As the representative example which applies SOM to document, classification, WEBSOM [4] which is proposed by T. Kohonen *et al.*, who is the founder of SOM can be given. WEBSOM uses the modification of BoW model for the representation of document. WEBSOM can classify the collection of the documents as a map which visualize the relation among the documents on 2-dimensional plane. And, some modified methods based on SOM is reported [5]-[7]. In these papers, SOM is considered to be suitable for document classification, because of the low computational complexity of SOM, and the ability of visualization of the relations among the documents in unsupervised learning manner. Furthermore, SOM is applied to the classification of DNA sequences based on BoW model in [8]. Recently, Deep Learning is applied to many areas, and it is also applied to document classification [9], [10]. Deep learning shows high performance in many application, however it needs huge computational cost, which requires special hardware, to establish high performance. In this paper, we propose the method which requires low computational costs, which can be processed in personal computers, with reducing the size of feature vector using Word2Vec and 1st layer SOM.

This paper is organized as follows. From Section II to Section IV, related techniques, BoW model, SOM, WEBSOM and Word2Vec are explained. In Section V, the method proposed in this paper is explained in detail. In Section VI and VII experimental results are presented. As the samples of general documents, livedoor news corpus [11] is used, and genome of 7 species are used in the experiment of the classification of DNA sequences.

II. Bag of Word(BoW) Model

For document classification, the vectorization of the documents to the vectors is necessary. For this purpose, the vector comprised of the frequencies of the words in a document can be used as the feature vector for classification. For each word included in the document, the frequency of word is counted as an element of the vector. Using this method, a word can be represented by a vector whose elements is 1 for the frequency of the word, and 0 for other words. As an example, Table I shows the example comprised of 3 Documents.

TABLE I: BoW Model of Documents

	word 1	word 2	word 3	word 4	word 5	···	word n
text 1	3	2	0	1	0		5
text 2	2	1	1	0	3		2
text 3	3	0	0	0	2		1

In this case, word 1 is used 3 times in the document 1, 2 times in document 2, and 3 times in document 3. The other words are converted in the same way. In this case, n words are counted in the vector. As the number of the documents increases, the number of the words which becomes the dimension of the vector increases.

For this problem, the compression method using Word2Vec and SOM is proposed in this paper.

Manuscript received March 7, 2018; revised May 11, 2018.

The authors are with the Graduate School of Science and Engineering Saga University, 1 Honjyo Saga, 840-8502, Japan (e-mail: 17578035@edu.cc.saga-u.ac.jp, hiro@dna.ec.saga-u.ac.jp).

doi: 10.18178/ijmlc.2018.8.3.695

III. Self Organizing Map(SOM)

A. Basic Self Organizing Map

Self-Organizing Map (SOM) is the model of the neurologic function of the cerebral cortex developed by T.Kohonen. As the class of neural networks, SOM is the feedforward type of two hierarchical without the interlayer using algorithms of unsupervised learning. SOM converts a nonlinear and statistics relations that exist between higher dimension data into the image with a simple geometrical relation. They can usually be used to make a higher dimension space visible because it is displayed as the lattice of the neuron of two dimensions.

Moreover, it becomes easy to be able to visualize higher dimension information because it is possible to cluster without the preliminary knowledge, and to understand the relations among these multi-dimensional data intuitively for human. SOM is basically a network in the double-layered structure in the input layer and the competitive layer as shown in Fig. 1.

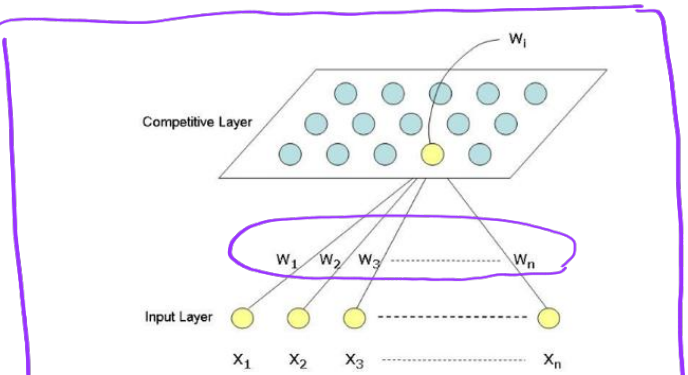

Fig. 1. Structure of SOM.

However, there are not connections between neurons in the same layer.

The first layer is input layer $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ of n dimension, and the second layer is called a competitive layer, and is generally two dimensions array for visualizing output data. It is assumed that it has the neuron that input data is given by n-dimensional real number vector $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ and two dimensions SOM were arranged on the grid point of m×m piece. Input data $\boldsymbol{x}$ is presented to all neurons. The neuron located in (j,i) in two dimension lattice array has $w_{ij} = (w_{ij}^1, w_{ij}^2 \ldots, w_{ij}^n)$ weight vectors which are tunable corresponding to the input vectors. This w_{ij} is called as reference vector.

The algorithm can be divided into two phases in a competitive phase and the cooperative phase. First of all, the neuron of the closest ,in a word, the winner neuron, is selected at a competitive phase. Next, the winner's weight are adjusted at the cooperation phase as well as the lattice near the winner.

B. WEB-SOM

WEB-SOM is a Self Organizing Map which classify the large amounts of documents like WEB data using BoW model of the documents as input vectors. To reduce the dimension of input vector, the words are once categorized by word category map, and the documents mapped to on the word category map are classified on documents map. WEB-SOM can be also used as search engine which can find similar WEB page.

In this paper, Word2Vec is introduced in the pre-processing of word category map to compress the dimension of word vector considering the meaning of the words more efficiently.

IV. Word2Vec

Word2Vec model is proposed by Mikolov et al, and is used to convert the words in the given set to the vectors based on the meaning of the words. It is applied to many researches of text processing recently. Word2vec can convert the similar words to similar vectors in Euclidean space, and linear operations between the vectors can represent the composition of the meaning of the words. Word2Vec used neural network as the internal model, and automatically organizes the words in the documents to vectors by machine learning.

V. Methods

A. Vectorization of the Word Based on Word2Vec Model

At first, all words in the set of documents are processed based on Word2Vec model, are given the vectors according to the meaning of the words. In this process, each word must be separated with space. In our research, the spaces should be inserted between words to process Japanese documents. For this purpose, morphological analysis system-MeCab is applied. As the dictionary of MeCab, ipadic-NEologd is used. After applying MeCab, the separated words are processed based on Word2Vec model.

B. Learning the Vectorized Words Using SOM

After applying Word2Vec model, the vectors are given to SOM as input vectors. To compress the number of words, the number of neurons are set as smaller than the number of words. SOM can map the similar words to same neurons using small number of neurons. The map organized by SOM using the vectors converted by SOM as input vectors is called "Word Map" in this paper.

C. Conversion of the Documents to Input Vectors

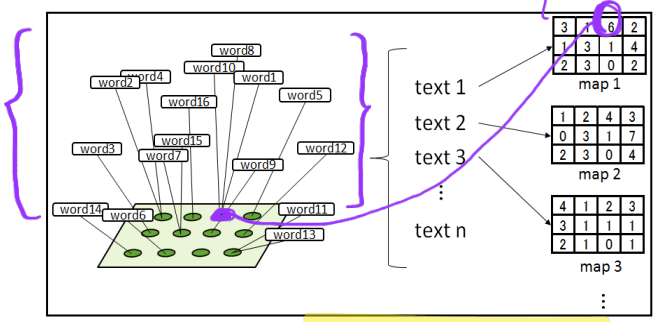

Fig. 2. Conversion of the document to word map vector.

After learning the words using SOM, each document is converted to the input vector. For each document, the words in the document are mapped on word map as to count the frequencies of becoming winner neuron. Fig. 2 shows an example. The vector 1 becomes winner for the vectors translated from the words in document 1 in 6 times, vector 8 and vector 9 become the winners in 2 times. Using the smaller number of neurons than that of words, dimension of the converted vectors of documents becomes smaller than

that of BoW model. The converted document is called as "word map vector". The word map vectors are given to 2nd-layer SOM which classifies the documents as shown in Fig. 3.

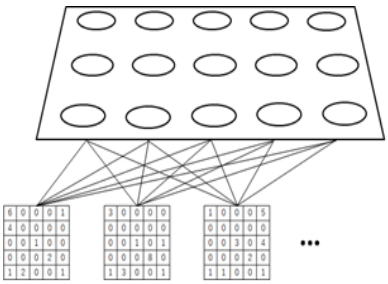

Fig. 3. Classification of the documets using word map vectors.

VI. Experimental Results Using News Corpus Data

In this section, the experimental results based on BoW model, and Doc2Vec model, and proposed method based on Word2Vec and SOM.

As mentioned before, livedoor news corpus data [11] is used in this experiment. Livedoor news corpus contains 7367 documents categorized in 9 categories, and total number of different words is 34768.

A. Experimental Result Using Input Vectors Based on BoW Model

At first, the experimental result learned by SOM using input vectors based on BoW model is shown. For each document, the input vector is composed of the frequencies of 34768 different words. Fig.4 shows the result. In this figure, each colored point represents a document categorized in the category shown in explanatory notes in right side. As the result of conventional SOM, the points in each color gather according to the similarities of input vectors. In Fig. 4, the points loosely gather on the map, and the documents in some categories are mixed on the map.

B. Experimental Result Using Input Vectors Based on Doc2Vec Model

Secondly, the experimental result learned by SOM using input vectors based on Doc2Vec model is shown. A document can be directly translated to a vector based on Doc2Vec model. Doc2Vec model can be also applied to document classification using machine learning method. In this subsection, the experimental result is shown for comparison. The dimension of the vector is set to 300. Fig. 5 shows the result. Compared with Fig. 4, the color points representing the documents gathers better. However some categories are scattered on the map similar to Fig. 4.

C. Experimental Result Using Input Vectors Based on Word2Vec and SOM

Next, the experimental result learned by 2nd-layer SOM using input vector based on Word2Vec model and 1st layer SOM is shown. All of words in the documents are processed based on Word2Vec model. The dimension of vector is set to 100. 34768 words are translated to 34768 vectors, and they are learned by 1st-layer SOM to organize word map. The size of word map is set to 10×10. After learning, each document is translated to word map vector with calculating the frequencies of becoming the winner during the mapping of the words in the document. The dimension of word map vector is 10×10=100. Compared with BoW model the dimension of input vector is compressed to 100 from 34768. All of the word map vectors are learned by 2nd layer SOM to classify the documents. Fig. 6 shows the result. Compared with Fig. 4 and Fig. 5, the documents in same categories represented with same color gathers better. Especially, the categories dokujo-tsushin and MOVIE ENTER gathers well. The categories of SMAX and IT life hack are mixed because they are considered to use same or similar words frequently.

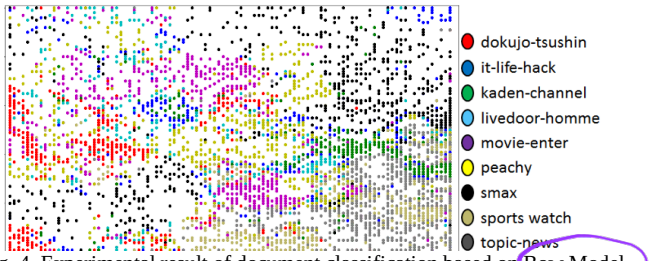

Fig. 4. Experimental result of document classification based on Bow Model.

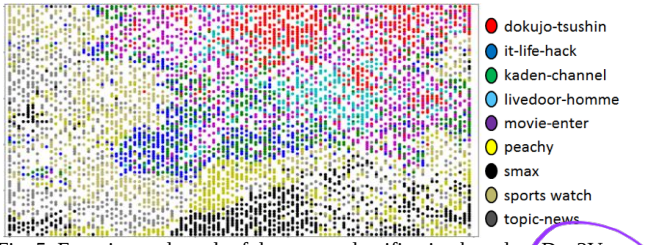

Fig. 5. Experimental result of document classification based on Doc2Vec.

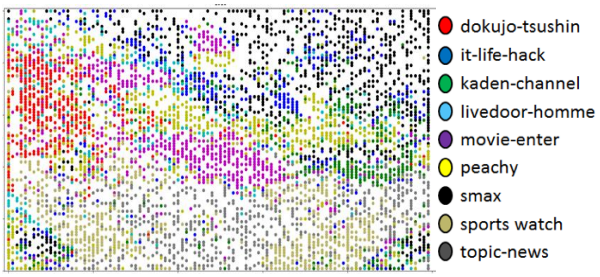

Fig. 6. Experimental result od document classification based on Word2Vec and SO.

VII. Experimental Results Using DNA Sequence

DNA sequence is the string of 4 character A, G, T, C, and it defines the whole of organism from development to death. The high detailed analysis of DNA sequence requires massive computational power. However, using some type of global feature of DNA sequences, they can be classified with far small computational power. As such global features, the vector of frequencies of N-mer(character) sequences is proposed in [8]This feature is based on BoW model of N-mer sequences. However, the dimension of vector becomes 4^N for N-mer sequences, thus the dimension increases exponentially by N. In this section, the result of classification of DNA sequences using the method proposed in the previous section is shown.

The DNA sequences are taken from KEGG online database. The sequences of the genes from amino acid metabolism for 7 species are used in the following experiments. As same as the procedure shown in 6.2, the sequences are processed based on Word2Vec model using

N-mer sequences as words. Fig. 7, Fig. 8 and Fig. 9 show the results with changing the N and 1st-later map size which becomes the dimension of input vector of 2nd layer. Fig.10 shows the result using the input vector based on Doc2Vec model for comparison.

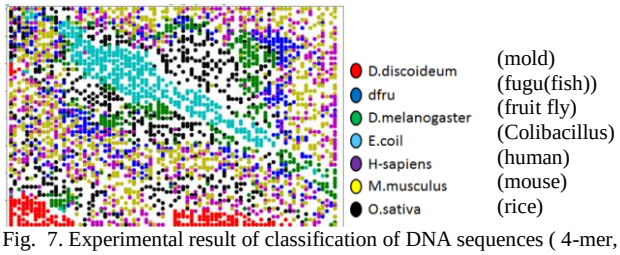

Fig. 7. Experimental result of classification of DNA sequences (4-mer, 1st layer SOM size 10 × 10).

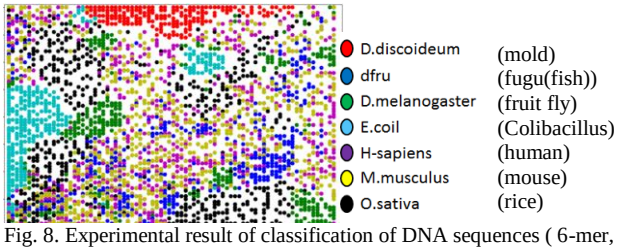

Fig. 8. Experimental result of classification of DNA sequences (6-mer, 1st layer SOM size 20 × 20).

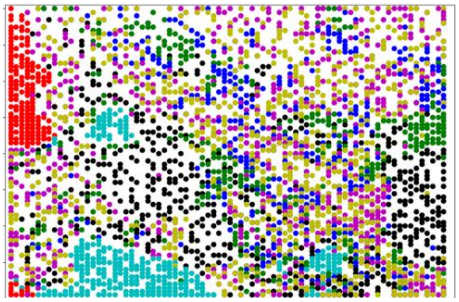

Fig. 9. Experimental result of classification of DNA sequences (6-mer, 1st layer SOM size 20 × 20).

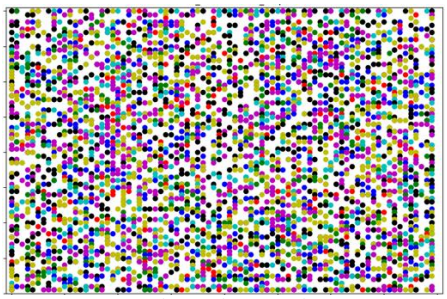

Fig. 10. Experimental result of classification of DNA sequence (Doc2Vec model.

In all Figs., the color points represent a gene of the specie shown in explanatory notes in right side. In Fig. 10, all of the species are randomly scattered on the map. Doc2Vec model is considered to be not able to extract the feature of DNA sequences. From Fig. 7 to Fig. 9, genes of mold, fruit fly, colibacillus and rice gather individually in some regions. Genes of human and mouse gather in mixed region because both of them are mammals. Genes of fugu gather closely to those of human and mouse because they are vertebrates. As shown in these figure, the organism are mapped according to the similarity among them. In Fig. 8 and Fig. 9, the input vectors are compressed greatly from 4^6=4096 and 4^8=65536 to 400 respectively.

VIII. Conclusion

In this paper, the classification method of documents based on Word2Vec and Self Organizing Maps in proposed. The dimension of input vector can be compressed with the conversion based on Word2Vec and 1st layer SOM compared with conventional BoW model, and 2nd layer SOM can classify the documents better than SOM using BoW model. The performance of this method is examined using the new corpus data. The documents in different categories gather individually on the map, and the documents using similar words gather together.

In this paper, the results of classification are visually evaluated based on the categories. The numerical evaluation of the classification results is considered to be required. Furthermore, the analysis of the classification in same categories, the analyses changing the size of vectors and processing of the word classes are required.

References

[1] Bag of Words (BoW)-Natural Language Processing. [Online]. Available: https://ongspxm.github.io/blog/2014/12/bag-of-words-natural-language-processing/
[2] M. Tomas, S. Ilya, C. Kai *et al.*, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, 2013.
[3] T. Kohonen, *Self Organizing Maps*, Springer, 2001
[4] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, "WEBSOM – Self-organizing maps of document collections," *Neurocomputing*, Vol. 21, issues 1–3, pp. 101-117, November 6, 1998.
[5] B. H. C. Shekar and G. Shobha, "Classification of documents using Kohonen's self-organizing Ma," *International Journal of Computer Theory and Engineering*, vol. 5, no. 1, pp. 610-613, 2009.
[6] N. Chowdhury and D. Saha, "Unsupervised text classification using kohonen's self organizing network," *LNCS*, vol. 3406, pp. 715–718, 2005.
[7] T. Chumwatana, K. Wong, and H. Xie, "A SOM-based document clustering using frequent max substring for non-segmented texts," *Journal of Intelligent Learning Systems & Applications*, vol. 2, pp. 117–125, 2010.
[8] T. Abe, T. Ikekura *et al.*, "Informatics for unreveiling hidden genome signatures," *Genome Res.*, vol. 13.
[9] X. Zhang and L. C. Yann, *Text Understanding from Scratch*, 2016.
[10] S. W. Lai, L. H. Xu *et al*., "Recurrent convolutional networks for text classification," in *Proc. the 29th AAAI Conference on Artificial Intelligence*, 2015, pp. 2267-2273.
[11] Livedoor new corpus. [Online]. Available: https://www.rondhuit.com/download.html

Koki Yoshioka was born on Oct. 2, 1993 in Nagasaki, Japan. He graduated Saga University located in Saga, Japan in March 2017, and continued to graduate course in the Department of Advanced Fusion. His research theme is machine learning, mainly on the application of self organizing map.

Hiroshi Dozono was born on Oct. 24, 1961 in Miyazaki, Japan. He graduated from Kyoto University in March 1984, and continued to graduate course and doctor cource in the Department of Applied Mathematics and Physics, and took PhD of engineering. He is now teaching in Saga University as an associate professor of the Department of Advanced Fusion. His research theme is soft computing including machine learning mainly on the application of self organizing map. He applied self organizing maps to the field of bioinformatics, controlling robot, biometric authentication and network security.