Simultaneous Visualization of Documents, Words and Topics by Tensor Self-Organizing Map and Non-negative Matrix Factorization

doi: https://doi.org/10.1109/SCISISIS50064.2020.9322683

Simultaneous Visualization of Documents, Words and Topics by Tensor Self-Organizing Map and Non-negative Matrix Factorization

Kazuki Noguchi
Dept. of Human Intelligence Systems
Kyushu Institute of Technology
Kitakyushu, Japan
noguchi.kazuki309@mail.kyutech.jp

Takuro Ishida
Dept. of Human Intelligence Systems
Kyushu Institute of Technology
Kitakyushu, Japan
ishida.takuro249@mail.kyutech.jp

Tetsuo Furukawa
Dept. of Human Intelligence Systems
Kyushu Institute of Technology
Kitakyushu, Japan
https://orcid.org/0000-0002-4469-7749

***Abstract*—The purpose of this work is to develop a simultaneous visualization method of documents, words, and topics. The task of the proposed method is to map a set of documents to a pair of low-dimensional latent spaces corresponding to documents and words, by which the relations between them are visualized. In addition, the method also decomposes the mapping as the sum of topics, so that the topic distributions are visualized on the latent spaces. To achieve the task, we combined the tensor self-organizing map and the non-negative matrix factorization. We applied the method to NeurIPS data set, and the result shows that the method enables us to understand the tripartite relation between document, words and topics easily.**

***Index Terms*—tensor self-organizing map, document analysis, topic, non-negative matrix factorization**

I. Introduction

Document visualization methods aim to map a set of documents to a low-dimensional, usually 2-dimensional space, in which similar documents are arranged to be nearby [1]. Similarly, word visualization methods aim to map a set of words to a low-dimensional space so that the similarities of words are indicated [2].

While the present methods mainly aim to visualize either documents or words, the purpose of this work is to develop a simultaneous visualization method of both documents and words, which enables us to make the cross-domain analysis (namely, document and word domains) as well as the intra-domain analysis. Thus, the proposed method does not only aim to generate a pair of low dimensional representations of documents and words, but also visualizes the relation between them, for example, visualizing the words contained in a document, and vice versa. Therefore such method is expected to enhance the information retrieval ability a lot, allowing us bidirectional data exploration.

Such cross-domain analysis can be further enhanced by introducing the concept of topic, which categorizes both documents and words. Thus, instead of examining any combinations of document and word pair exhaustively (which usually becomes too much to examine), we can see the rough picture of their relations from the viewpoint of topics. Therefore, it is expected that the visualization method using topics extends the data exploration ability from bi-directional to tri-directional.

In summary, the tasks of the proposed method are as follows. (i) From a bag-of-word document data, generating two low-dimensional visualizations corresponding to documents and words, which are referred to as 'maps' in this paper. Thus, the first task of the method is to generate a document map and a word map simultaneously, by which similarities of documents/words are indicated. By using these maps, intra-analysis can be executed. (ii) The second task of the method is to visualize the mutual relation between documents and words. Thus visualizing the regions of the words contained in a particular documents, and vice versa. More simply, by specifying a point in the document map, the corresponding regions are indicated in the word map. Similarly, by specifying a point in the word map, the corresponding regions are indicated in the document map. (iii) The third task of the method is to visualize the regions in the document/word maps corresponding to each topic.

In order to achieve the above tasks, the proposed method combines the tensor self-organization map (TSOM) [3] and the non-negative matrix factorization (NMF) [4]. The TSOM visualizes documents and words simultaneously, whereas the NMF makes the topic decomposition of the maps obtained by TSOM.

II. Problem Formulation

Let $\mathcal{D}$ and $\mathcal{W}$ be the set of documents and words respectively. For $d \in \mathcal{D}$ and $w \in \mathcal{W}$, let $u_d \in \mathcal{U}$, $v_w \in \mathcal{V}$ be the corresponding latent variables, representing the positional vectors in the low dimensional spaces for visualization, namely, the latent spaces $\mathcal{U}$, $\mathcal{V}$. Typically $\mathcal{U}$ and $\mathcal{V}$ are two dimensional unit square spaces $[0,1]^2$, which are also used in this paper.

The input is the bag-of-words (BoW) data $\mathbf{N} = (n_{dw})$, namely, the frequency of word $w \in \mathcal{W}$ in document $d \in \mathcal{D}$. In this work, we preprocessed BoW data by the tf-idf [5], which is a non-negative index characterizing the importance

This work is supported by KAKENHI (18K11472). This work is also supported by ZOZO research.

of the words in each document. Thus, $x_{dw} = \text{tf-idf}(n_{dw})$ is the actual input.

When $\mathbf{X} = (x_{dw})$ is given, our aim is to estimate the latent variables $\mathbf{U} = (u_d)$ and $\mathbf{V} = (v_w)$, so that $\mathbf{X}$ is decomposed as

$$x_{dw} \simeq \sum_{t=1}^{T} \varphi(u_d|t)\, \psi(v_w|t), \tag{1}$$

where t represents the topic, and $\varphi(u_d|t)$, $\psi(v_w|t)$ are the smooth functions of u_d and v_w with respect to topic t.

Though the traditional tf-idf is employed in this paper, the probabilistic approach is also possible. Let $P(w|d)$ is the occurrence probability of word w in document d. In this case (1) becomes

$$P(w|d) \propto P(w)\, p(v_w|u_d) = P(w) \sum_t p(v_w|t)\, p(t|u_d), \tag{2}$$

which is regarded as a topic model with the latent variables.

III. Proposed Method

To perform the above task, we employed the following objective functions.

$$F_{\mathrm{T}} = \sum_d \sum_w \iint h(u|u_d)\, h(v|v_w)\ (x_{dw} - f(u,v))^2\ \mathrm{d}u\, \mathrm{d}v \tag{3}$$

$$F_{\mathrm{N}} = \iint \left(f(u,v) - \sum_t \varphi(u|t)\, \psi(v|t) \right)^2 \mathrm{d}u\, \mathrm{d}v, \tag{4}$$

such that $\varphi(u|t), \psi(v|t) \geq 0$, where $h(\cdot|\cdot)$ is the kernel function. In this work, Gaussian function is used as the kernel. Note that (3) (4) are the objective functions of the TSOM and the NMF respectively [3] [4]. Therefore, the objective function F_{T} is optimized by TSOM, the F_{N} is optimized by NMF. If the task is defined as (2), then the Euclidean distance is replaced by KL-divergence, and the corresponding methods become the TSOM for probability set [6] and the latent Dirichlet allocation (LDA) [7].

The details of the proposed method are described as follows.

Step 1: Latent variable estimation by TSOM

After the latent variables are initialized randomly, the map $f(u,v)$ and the latent variables $\{u_d\}, \{v_w\}$ are estimated alternately until they converge. $f(u,v)$ is estimated as

$$f(u,v) = \frac{1}{H_1(u) H_2(v)} \sum_d \sum_w h(u|u_d)\, h(v|v_w)\, x_{dw}, \tag{5}$$

where $H_1(u) = \sum_d h(u|u_d)$ and $H_2(v) = \sum_w h(v|v_w)$. Then the latent variables are estimated as

$$u_d = \arg\min_u \int \left(g_1(v|d) - f(u,v)\right)^2 \mathrm{d}v, \tag{6}$$

$$v_w = \arg\min_v \int \left(g_2(u|w) - f(u,v)\right)^2 \mathrm{d}u, \tag{7}$$

where

$$g_1(v|d) = \frac{1}{H_2(v)} \sum_w h(v|w)\, x_{dw} \tag{8}$$

$$g_2(u|w) = \frac{1}{H_1(u)} \sum_d h(u|d)\, x_{dw}. \tag{9}$$

(5)–(7) are executed iteratively until F_{T} converges.

Step 2: Topic decomposition by NMF

After estimating the simultaneous mapping $f(u,v)$ in Step 1, it is decomposed to topics by NMF. To apply NMF, the latent space is discretized to K nodes, and the continuous functions are transformed to matrices. Thus, (1) becomes $\mathbf{X} \simeq \Phi\Psi^\top$. In this paper, we used the algorithm proposed by Lee and Seung [4], as follows.

$$\Phi^{\text{new}} := \Phi \odot (\mathbf{X}\Psi) \oslash (\Phi\Psi^\top\Psi) \tag{10}$$

$$\Psi^{\text{new}} := \Psi \odot (\mathbf{X}^\top\Phi) \oslash (\Psi\Phi^\top\Phi) \tag{11}$$

Here $\odot$ and $\oslash$ are the elementwise product/division of matrices respectively.

IV. Visualization

The proposed method provides several exploring methods of the obtained model. (i) The method provides two low dimensional representations corresponding documents and words, namely, the document map and the word map. These maps can be used as the ordinary visualization using dimensionality reduction. Thus, the documents which consist of similar words are located nearer in the document map. Similarly, the words map represents the similarities between words. (ii) For cross-domain analysis, the conditional component plane (CCP) is useful [3]. By selecting a point in the document map, say u_p, CCP displays $f(v|u_p) \equiv f(u_p, v)$ on the word map as the gray scale. Thus CCP visualizes the region of words, which are contained in the documents at u_p in the document map. Similarly, by selecting a point in the word space v_p, CCP displays $f(u|v_p) \equiv f(u, v_p)$ on the document map as the gray scale indicating the documents, containing the words at v_p. Therefore CCP is a powerful method of bi-directional search. (iii) The last visualization method is the topic component plane (TCP), which visualizes both $\varphi(u|t)$ and $\psi(v|t)$ in the document and word maps. CCP is useful to see the property of a specific point in the maps, whereas TCP is convenient to overview the two maps at the same time. Since the number of topics is usually more or less a dozen or so, TCP summarizes the entire maps concisely.

V. Results

The proposed method is applied to visualize the NeurIPS dataset. In this work, we used 402 papers accepted in 2015, and 250 representative keywords are used. The number of topics was 10, determined by the reconstruction error of NMF.

The obtained word and document maps are shown in Fig. 1 with CCP gray scales. For example, the words related to the bandit problem are located at point B in the word map, and

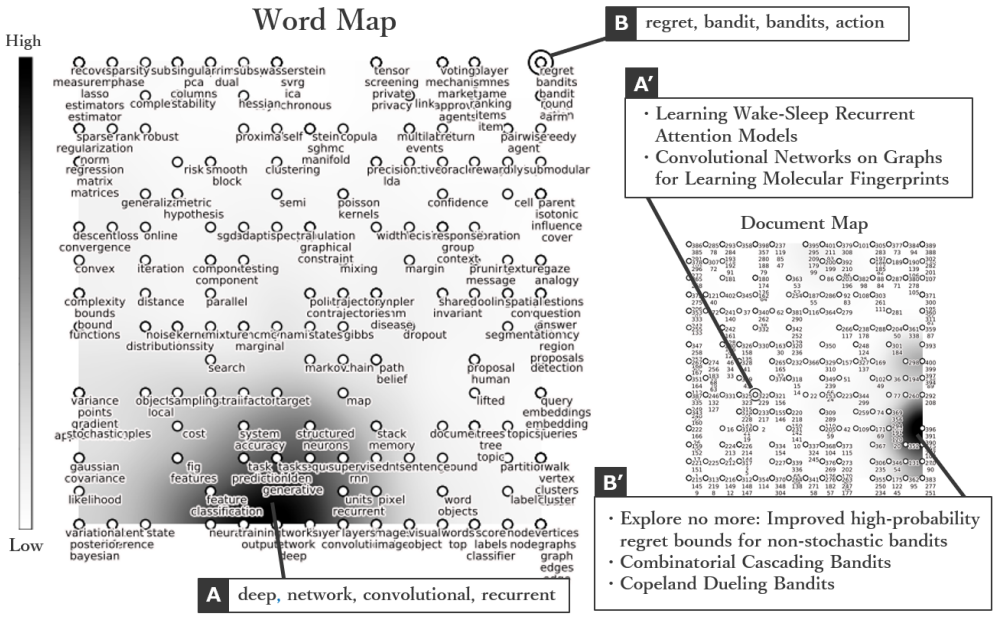

Fig. 1. The word map and the document map generated from NeruIPS dataset. The gray scale represents the conditional component plane (CCP). At the conditioning point B in the word map, words related to bandit problem are located. The corresponding the region of the documents containing these words (B') is indicated by the gray scale in the document map. Similarly, the conditioning point A' in the document map corresponds to the region around point A in the word map.

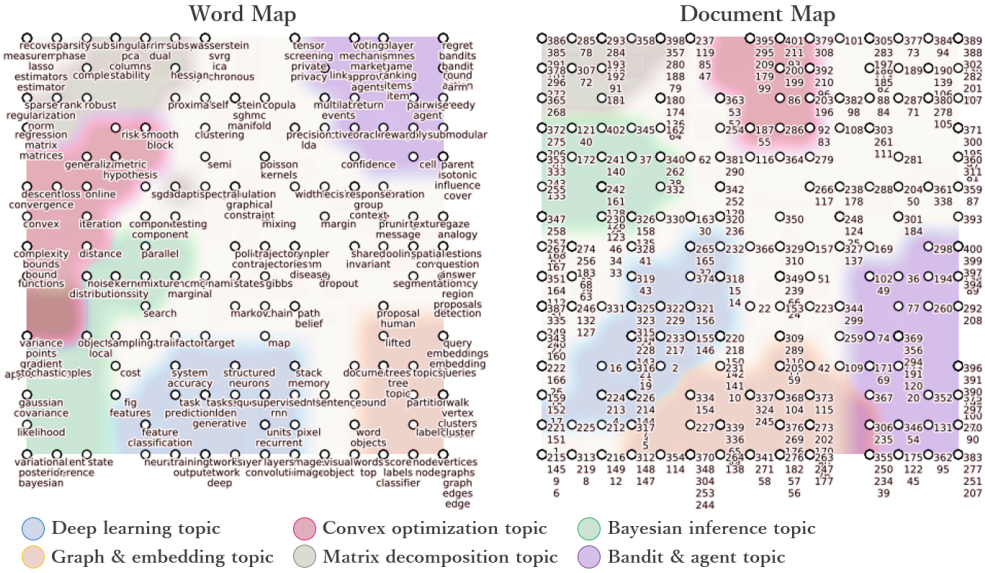

Fig. 2. The regions topics in the document and the word maps. 6 out of 10 topics are indicated.

the corresponding region of the documents containing these words is shown in the document map as grayscale.

Fig. 2 shows the topic distributions in both word and document maps. Though topics are overlapped to each other, they softly divide the latent spaces. Mediated by topics, it is easy to overview the relations between documents and words.

Fig. 3 and Fig. 4 are examples of TCP. The gray areas in these figures indicate the regions of words and documents related to the deep learning topic and the Bayesian inference topic respectively. By combining CCP and TCP, users can explore the documents and the words maps easily. This is the advantage of the proposed method.

REFERENCES

[1] S. Huang, M. O. Ward, and E.A. Rundensteiner, "Exploration of dimensionality reduction for text visualization," Tech. Rep. TR - 03–14, Worcester Polytechnic Institute, Computer Science Department, 2003.

[2] H. Heuer, "Text comparison using word vector representations and dimensionality reduction," ArXiv 2016.

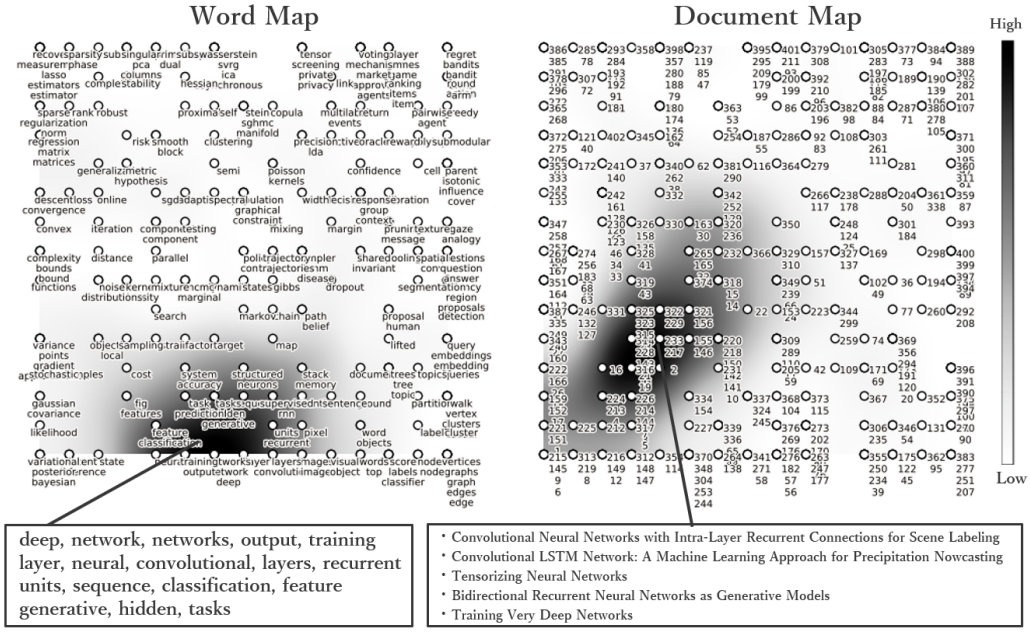

Fig. 3. The words and the documents of the deep learning topic.

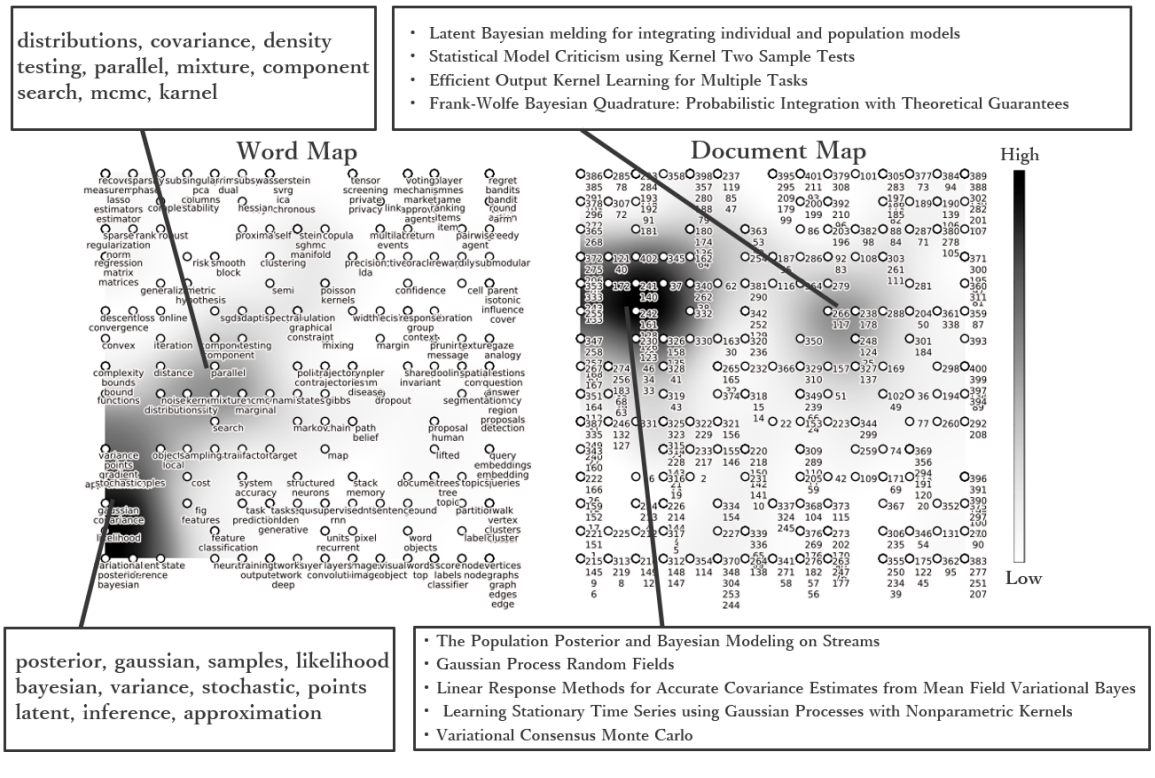

Fig. 4. The words and the documents of the Bayesian inference topic.

[3] T. Iwasaki and T. Furukawa, "Tensor SOM and tensor GTM: Nonlinear tensor analysis by topographic mappings," Neural Networks, vol.77, pp.107-125,2016.

[4] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in Adv. NIPS, 556/562 (2000)

[5] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TFIDF, LSI and multi-words for text classification," Expert Systems with Applications, Volume 38, Issue 3, 2011.

[6] T. Ishida, H. Hatano, and T. Furukawa, "Simultaneous visualization of documents and words by using tensor self-organizing map," SCIS and ISIS 2018 in conjunction with ISWS 2018.

[7] D .M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet allocation," Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.