Journal of Next-generation Convergence Information Services Technology Vol.7, No.2, December (2018), pp. 177-186
차세대컨버전스정보서비스기술논문지 *http://dx.doi.org/10.29056/jncist.2018.12.05*

Analysis and Improvement Plans on Visualization Systems for Word Analysis: Focused on Word2Vec

Ho-Seong Kang[1], Jung-Yoon Kim[2]*

Abstract

As a number of studies on big data have been conducted, a great deal of information is extracted and shared by analyzing data. Studies on extracting useful information through artificial neural network based machine learning from accumulated voluminous document data have been made. Recently, Word2Vec that has overcome complexity of natural language processing machine learning has appeared as an efficient model to extract a link between words. A study on extracting similar words and visualizing information through Word2Vec model is being conducted. Visualization system helps people to easily perceive information by maximizing understanding of information extracted from big data. This study will analyze visualization system cases of Word2Vec model to find visualization system which is easy for people to understand.

Keyword : Visualization, Word2Vec, BigData, word analysis

1 Graduate School of Game, Gachon University, Seongnam-Si, Gyeonggi-Do, Korea [Graduate Student]
e-mail: sos6884@naver.com
2 Graduate School of Game, Gachon University, Seongnam-Si, Gyeonggi-Do, Korea [Professor]
e-mail: kjyoon79@gmail.com (Corresponding author)
* This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology(CT) Research & Development Program 2017 (R2017030062_00000002)
Received(September 29, 2018), Review Result(1st: October 14, 2018), Accepted(December 04, 2018), Published(December 31, 2018)

ISSN: 2384-101X JNCIST

1. Introduction

As information society has come, the amount of information people can get has increased infinitely. As big data technology has advanced, quality of information has improved. As the technology of storing and processing a great deal of data has advanced, studies on analyzing and processing information are being made. Advanced data processing module was developed by effectively studying a process of analyzing data through several proven artificial neural network. Based on the foregoing, various information can be extracted from a large volume of data.

Most researchers analyze data in various fields and especially informal data consisting of natural language has been analyzed[1].

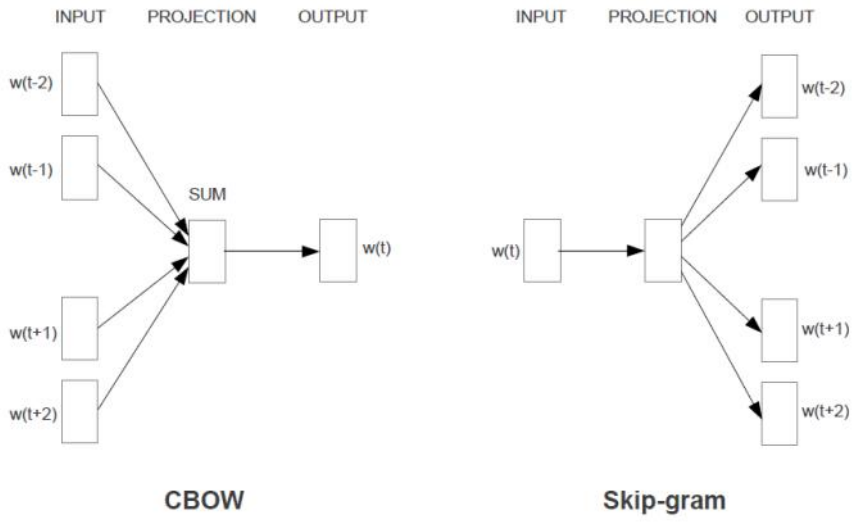
Word2Vec technology deserves to receive attention. Word2Vec model that was developed by improving complexity of NNLM(Feedforward Neural Net Language Model) and RNNLM(Recurrent Neural Net Language Model) in artificial neural network based machine learning draws attention. Word2Vec studying relations between words in a document and finding similar words forms foundation for big data study. Showing information analyzed through Word2Vec simply as text type result makes it difficult for people to understand. On the other hand, information reprocessed as diagram or image by converting information to figures makes it easy for people to understand. Once information is obtained through big data technology, is should be expressed through visualization for better understanding. Visualization helps people to understand information better[2].

This study analyzes visualization technology and framework of Word2Vec used for research considering the fact that there are lots of studies which make use of Word2Vec in machine learning and text analysis research.

2. Visualization framework analysis of Word2Vec

Word2Vec proposed a model which overcomes limits of machine learning models in natural language processing in a paper “Efficient Estimation of Word Representations in Vector Space” presented by Tomas Mikolov, at Google. Word2Vec converts meaning of words to vector form and calculates correlation between nearby words at vector level and analyzes its meaning through artificial neural network[3]. Word2Vec has two learning models. [Fig. 1] describes models of Skip-Gram and CBOW(Continuous Bag of Words). CBOW model can infer a word of “key” through words between blanks in a sentence “I entered opening closed door with a __.”. Unlike CBOW model, Skip-Gram

model infers nearby words from a word of “key”.



[Fig. 1] Learning architecture of the CBOW and Skip-gram models of Word2Vec[3].

Word2Vec can draw meaning of words in a document or a sentence in learning through two models. Information is visualized for better understanding. As several studies to express meaning of Word2Vec have been made, many visualization frames are used. This study selected three framework cases used for visualization of vector value or information to analyze visualization of Word2Vec.

First, TensorBoard learns data by using Word2Vec based on TensorFlow, neural network library in python environment and visualizes a result through its own framework.

Second, ScatterText implements scatter plot graph in python environment using vector data or text type data learned with Word2Vec to provide data visualization.

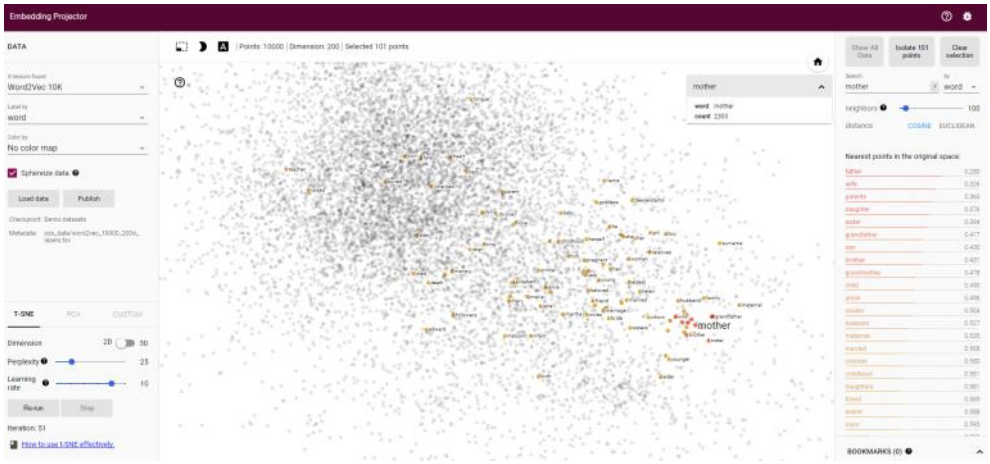
Lastly, D3.js expresses visualization data in web environment freely. There are several frameworks that visualized Word2Vec and framework which is popular with researchers or technicians was selected.

2.1.TensorFlow-TensorBoard

TensorFlow was developed by researchers and engineers on Google Brain team for research on machine learning and deep neural network. TensorFlow is open source software library using data flow graph. TensorFlow is released as apache 2.0 open source license on November 9, 2015[4][5]. TensorFlow is used by Google, Ebay and Intel with high level of awareness. Papers making use of

TensorFlow have been published. TensorFlow is mainly used in implementing machine learning and artificial neural network.

Visualization studies on informal data using visualization framework through TensorBoard provided by TensorFlow is being conducted.

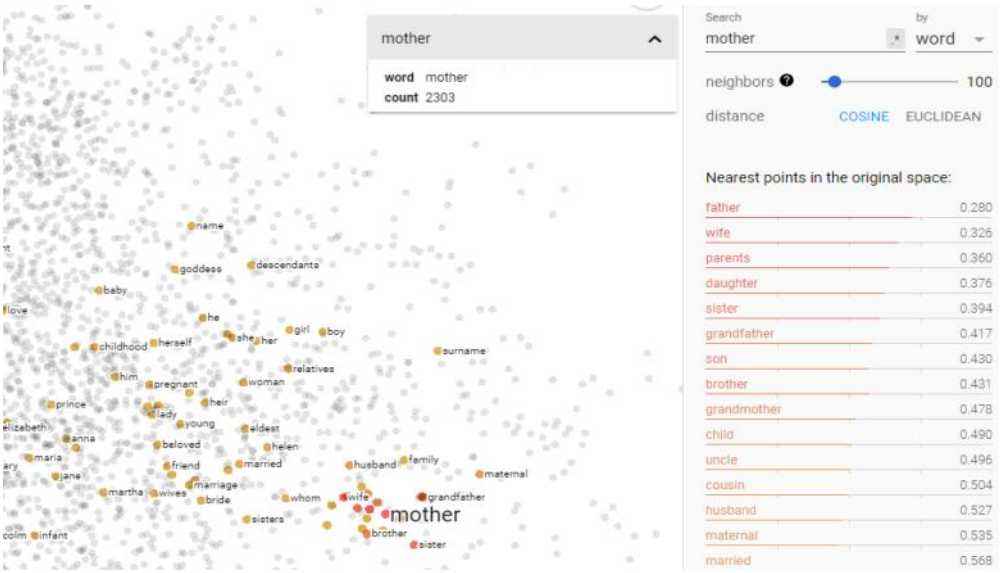


[Fig. 2] TensorBoard-Embedding Project, Google Tensorboard [6].

[Fig. 2] is visualization framework provided by TensorFlow. As shown in [Fig. 2], dataset in an example is learned with Word2Vec and drawn vector data is embedded to visualize each word in a scatter plot graph. There are T-SNE(T-distributed Stochastic Neighbor Embedding) method, PCA(Principal Component Analysis) method and user definition method as projection methods that can be selected from TensorBoard.

T-SNE method can check clustering in two dimensional environment and three dimensional environment by calculating correlations between words through nonlinear dimension reduction technique. PCA method can check distribution form of words by selecting ten component calculation using technique restoring high level data to low level data. User custom method is linear projection method for each axis using data label in each axis. Rendering a model with each projection method can display visualization result.

TensorBoard provides a search for lots of dots in scatter graph. [Fig. 3] is result value searched up to 100 words based on the level of similarity between words near ‘Mother’ in example of TensorBoard Embedding. Scatter plot graph shows a result with searched word and nearby words highlighted.



[Fig. 3] TensorBoard-Embedding Project Exploration, Google TensorBoard [6].

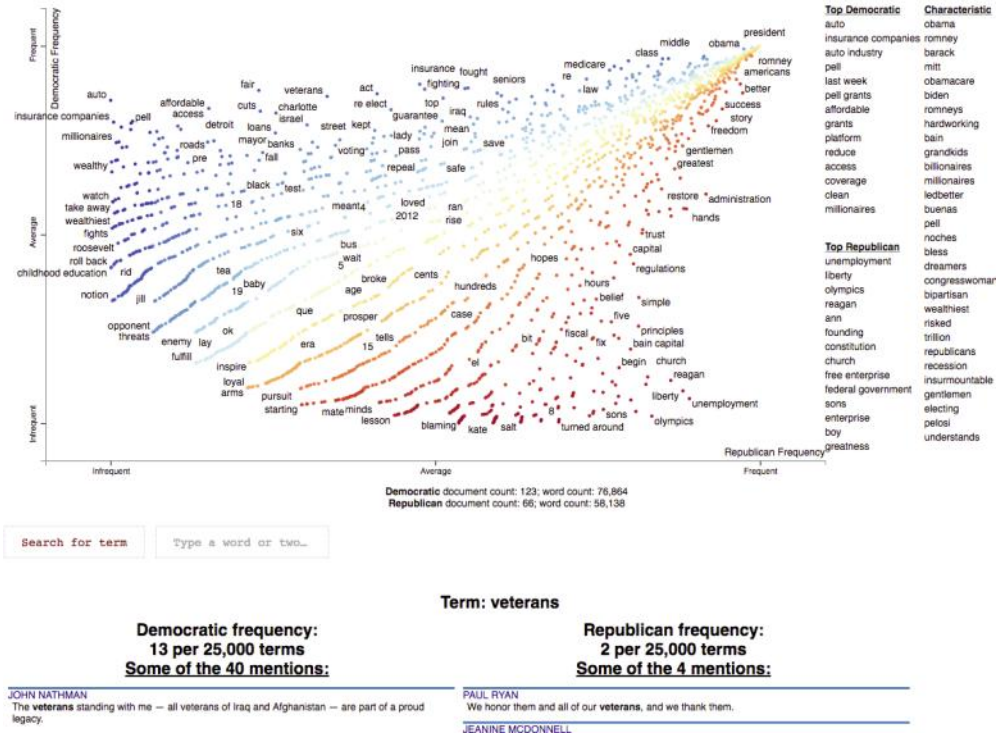
TensorBoard provides a result of Word2Vec through scatter plot form visualization. Various projection methods and search function make it possible to find correlations between words in a document.

Setting and visualizing data set require complex process. In python environment, data can be visualized easily but TensorFlow library is required and python environment which conform to library of TensorFlow must be built. Anaconda Cloud controlling python version and library can supplement weakness but it is difficult to use TensorBoard in an environment in which machine learning of Word2Vec has been built in C++or C# not python environment.

2.2 ScatterText

ScatterText is open source tool that visualizes data in a form of scatter plot with words in a document categorized in python environment developed by Jason S. Kessler[7]. ScatterText refines data in various methods and provides scatter plot graph framework with several directions.

ScatterText provides 0.0.2.28 version. ScatterText is optimized in Python 3.4 environment. ScatterText recommends using additional library such as spaCy, natural language module. ScatterText can be used efficiently through Anaconda Cloud to install relevant library and control version.



[Fig. 4] Example of data visualization in ScatterText [8].

[Fig. 4] is visualization example showing correlation between Democratic Party and Republican Party based on texts of a speech by Democratic Party and Republican Party using ScatterText library. X axis represents Republican Party and Y axis represents Democratic Party. It was found that words located near each axis are of high frequency.

Aside from above example, there is an example which visualizes in scatter plot graph form by classifying emoticon frequently used by men and women as negative/positive based on twitter data. Data visualization methods are provided based on various examples. ScatterText supports scatter plot graph visualization through T-SNE algorithm using Word2Vec. However, scatter plot graph support by ScatterText does not represent characteristics which only ScatterText has. ScatterText goes through natural language processing based on text form document data. ScatterText is framework optimized for visualizing information by analyzing words, correlation between words, influence of words in a document and tendency of words. The author of this study was faced with a problem of organizing dataset of documents written in Korean alphabet. ScatterText has been created based on spaCy, natural language processing library and does not support Korean alphabet processing. In order to solve above mentioned problem, Korean alphabet processing learning process in spaCy is required and another natural language

processing which compatibility is not clear should be combined.

2.3 D3.js

D3.js(Data-Driven Documents) is Javascript based library to implement interactive visualization in web environment developed by two persons including Mike Bostock[9]. D3.js provides community that can be used freely by developers opening various visualization examples and codes as shown in [Fig. 4]. D3.js expresses data visualization using HTML, SVG, CSS technologies on web. Several enterprises such as New York Times and Datameer use D3. D3.js is used as web live visualizing data in developing framework and conducting research. D3 is easy to implement visualization that web programming based knowledge should have when visualization library drawing basic graphs in python or R environment is used.

[Fig. 5] D3.js-based visualization examples and library list[9].

D3 does not provide framework of Word2Vec vector data visualization technology. Visualization framework can be implemented by researcher or developer selecting visualization method which he/she wants through D3's various visualization methods. Typical example is scatter plot type visualization method.

Advantage of D3 is that it can implement visualization in various forms. D3's visualization framework can be implemented through library Django in python environment. Using C# and ASP.NET can provide a user with visualization information based on data and web server. Unlike other cases, D3

does not have strength with regard to Word2Vec's visualization technology but can create new frameworks based on various visualization methods.

2.4 Comparative analysis of cases

[Table 1] below shows analysis of above mentioned three cases.

[Table 1] Visualization analysis results of Word2Vec

Type	TensorBoard	ScatterText	D3.js
Developer	Google Brain Team	Jason S. Kessler	Mike Bostock,Jeffrey Heer,Vadim Ogievetsky
Language	Python/C/Java	Python	Javascript
License	Open source	Open source	Open source
Dependency	Dependent on TensorFlow	Independent	Independent
Accessibility	Need to know how to use Python	Need to know how to use Python and Python Package	Need to know how to use web development language
Support Language	Most language supported	Supported language NLP Module spaCy (Unsupported Korean)	All languages supported
Visualization Environment	Web	Web/Html	Web
Visualization View	Scatterplot graphs (2D,3D)	Scatterplot graphs (2D)	Various visualization techniques
Visualization techniques	t-SNE,PCA,Custom	t-SNE,-axis for Semantic analysis	Variety
Visualization enabled data	Text/Image	Text/Image(emoticon)	Variety
Word2Vec Visualization Support	Supported	Supported	Unsupported
Pros	easy to visualize the associative information of words Providing a proven projection	Easy for visualizing propensity information 2D scatterplot optimized for graph representation	Various visualizations available Highly interactive using various visualization techniques High portability about the Other platform
Cons	Dependent on TensorFlow Time to load visualisations is long.	No support for NLP of Korean language data Need more libraries for visualization. Low interactive About the visualisation information	Need to develop a visualisation framework for word2Vec Need a high understanding of language in web development

3. Limits of Word2Vec's visualization framework cases

Word2Vec visualization method provides vector data based scatter plot graph. Scatter plot graph can check visualization information centering on similarity among words learned through Word2Vec but it is difficult to elicit words that are similar to word that a user or a researcher wants. Framework provided in cases focuses on visualization of information collected based on similarity between words which requires to see all words. Word2Vec visualization method searches a word to show similarity groups for relevant word but it is difficult to perceive similarity between words. Therefore, it is necessary to develop visualization method that focuses on words searched other than extraction of a word from all words. When searching a word of male, a word of female should be inferred and at the same time similar word list should be extracted and in vector operation of male+female-king, a list of result value for a word of queen should be inferred. Therefore, it is important to create visualization method centering on searched words in a new form when similar word list is extracted rather than to express in a form of grouping in scatter plot graph.

4. Conclusion and further studies

After having examined, tested and analyzed several cases of Word2Vec, limits of Word2Vec's visualization system have been found. Word2Vec visualization framework has complexity which is difficult for people to understand. Data visualization aims to provide visualization which is easy for people to understand. Ultimate purpose of this study is to improve so that people can recognize information of Word2Vec easily.

It is recommended that further studies should cover how to improve Word2Vec's visualization system based on analysis of above mentioned cases and design and produce visualization system for similarity between words. Implementing visualization system which is easy to understand based on future improvement will provide an opportunity for people to access various information more easily.

ISSN: 2384-101X JNCIST

Reference

[1] ZHAO ZILONG, “Comparing Big Data Research Trend in Korea and China by Text Mining”, Department of Management Information System Graduate School, Chungbuk National University, Master’s thesis, Republic of Korea, (2016): 3~24.

[2] H.S. Shin, J.M. Lim, J.S. Park, Information Visualization and Information Presentation for Visually Impaired People, 28.1, (2013): 81~91.

[3] Tomas Mikolov, Efficient Estimation of Word Representations in Vector Space, (2013).

[4] M. S Jo, Artificial Intelligence Open Source Library "Tensorflow" and Development of Artificial Intelligence Application Software, 34.10, (2017): 55~63.

[5] J.Y Kim, Introducing Google TensorFlow, 23.2, (2015): 9~15.

[6] http://projector.tensorflow.org, Retrieved: Agust 3, (2018).

[7] Jason S. Kessler, Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ, (2017).

[8] https://github.com/JasonKessler/scattertext, Retrieved: Agust 3, (2018).

[9] https://d3js.org/, Retrieved: Agust 3, (2018).