SeqTextVis: A Distributed Word Representation Learning for Time Sequential Text Data Visualization

Jheng-Long Wu[1,†] Mu-Hui Yu[2,*]

1) Department of Information Management, Chinese Culture University, Taipei, Taiwan
2) Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Abstract

Texts contain both implicit and explicit information. Explicit info can be easily observed and explored from co-occurrence matrix in text while finding and retrieving implicit info from raw text is very difficult, especially on time sequential text data. However, extracting both explicit and implicit info is important because it helps in explaining overall hidden relations among texts. In this paper, we propose SeqTextVis, a new method to learn from time sequential text data, and visualize explicit and implicit info based on distributed word representation learning methods: word2vec and node2vec. Word2vec, which is similar to co-occurrence approach, is a simple way to learn neighbor relations and extract explicit info; node2vec can extract implicit info by learning parsed structures. SeqTextVis combines with both advantages and time-decay feature to make dynamic graphs of both explicit and implicit relations chronologically. We used case study to explain correctness and effectiveness for SeqTextVis system evaluation.

Keywords: Distributed word representation, time sequential text data, explicit relation, implicit relation.

1 Introduction

Text data visualization always plays an important part in data science. Based on the authors' limited knowledge, there's still not any effective analysis and visualization method for information extraction on time sequential text data. Word embedding is a distributed representation method, which can map information, such as syntax and context, to vectors of real number [1,2,3]. Word2vec [1] is one popular distributed word embedding approach which can learn syntax of linguistics that we called explicit relations. Node2vec [4] is another advanced method to learn distributed representation by scalable feature learning for nodes. We called that implicit relations. Although text data visualization has been developed [5] for years, it's still hard to solve time sequence problem. Therefore, we propose a SeqTextVis system to capture words relations over time, and use a dynamic graph to visualize time sequential relations of words.

The purpose of this paper is to analyze and visualize raw text data with time sequence. There are two contributions: 1) Using word2vec and node2vec to learn relations with time sequence among words for extracting explicit and implicit info. 2) Visualizing these relations to a dynamic graph chronologically.

2 Approach

We use word2vec and node2vec with time decay feature to analyze time sequential text data, and a dynamic graph to visualize these trained word embedding. The overall concept and flow are shown in Figure 1. More detailed processes are follow as:

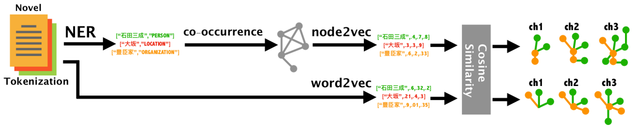


Figure 1: Training processes of our proposed SeqTextVis system.

2.1 Word2Vec to Learn Explicit Relations

We used two different word2vec model architectures to produce a distributed representation of words: continuous bag-of-words (CBOW) and continuous skip-gram. In the CBOW model, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction. In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words. We train word vectors on segmented dataset by two model of word2vec. In our experiment, skip-gram performs better than CBOW.

2.2 Node2Vec to Learn Implicit Relations

Node2vec learns low-dimensional representations for nodes in a parsing graph by optimizing a neighborhood preserving objective. The objective is flexible, and the algorithm accommodates for various definitions of network neighborhood by simulating biased random walks. Node representation uses breadth-first sampling and depth-first sampling to capture its neighborhoods information. Specifically, it provides a way of balancing the exploration-exploitation trade-off that in turn leads to representations obeying a spectrum of equivalences from homophily to structural equivalence. First, we use Name Entity Recognition (NER) model to tag words. The NER model is trained by "Sengoku period" (戦国時代) introduction from Wikipedia, recognizing 4 NE types including the time (T), location (L), person (P) and organization (O). Then we count the co-occurrence relations of tagged dataset to construct the parsing graph, which is syntax-level graph. After that, we use node2vec to train word vectors from parsing graph.

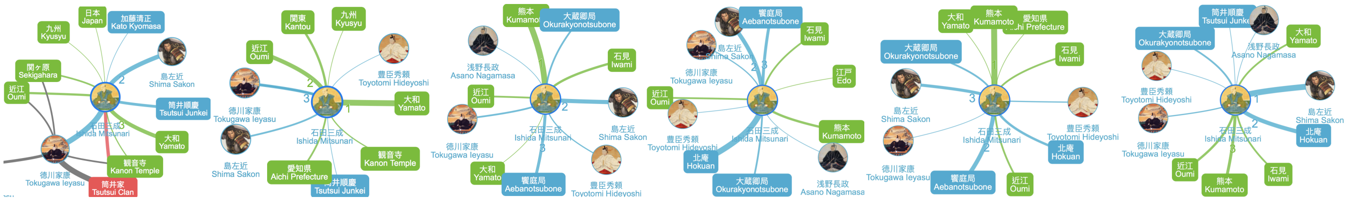
2.3 Time Sequential and Memory Decay Learning

For time sequential learning, we segment dataset into different chapters and train in accumulation. For example, we train the first graph by chapter 1, the second graph by chapter 1 and 2, and so on. For each new graph, it may increase or decrease some nodes and changes edges between nodes. After training, we compare those graph and find a reasonable explanation for those changes. Also, by setting memory decay rate and learning rate, it helps decay previous data and increase weights of incoming data.

2.4 Visualization on Time Sequential Relation

All word vectors generated by distributed word representation approaches are transformed into a dynamic graph over time. In the dynamic graph, nodes and edges are denoted as 4 NE types and cosine similarity value. We use these nodes and edges to construct the dynamic graph for time sequential relation on explicit relations and implicit relations.

*Corresponded author. E-mail: b03902023@csie.ntu.edu.tw
†E-mail: jlwu.studio@gmail.com



Chapter 1 Chapter 2 Chapter 3 Chapter 4 Chapter 5 Chapter 6

Figure 2: Implicit relation (information) over time from chapter 1 to chapter 6 by time decay node2vec.

3 Primary Experimental Findings

We used the dataset from Chinese version of novel "Sekigahara (関ヶ原)" written by Shiba Ryotaro (司馬遼太郎). There are 440,284 characters with 107 chapters, but we only used the first 6 chapters to learn relations. The role of "Isida Mitsunari (石田三成, R1)" is observed in result by SeqTextVis system.

3.1 Type of Relations Comparison

The co-occurrence analytics of R1 is shown in Table 1. For word2vec and node2vec, the top 5 related nodes are shown in Table 2. In Table 1 and 2, we found that most top 15 related nodes are "person" types, and so does word2vec because it is trained by context nodes. However, node2vec can find other NE types, which are not directly mentioned but also important such as Toyotomi reign (T1) because R1 lives under that reign. Moreover, the time decay node2vec can find 1) directly related words which also appear in co-occurrence results such as Shima Sakon (R2) and Oumi (L1, where is R1's hometown), 2) other location or time words as Kumamoto, and 3) some indirectly related words, like Hokuan (R3): although the relation of R1 and R3 isn't mentioned in novel, R1's most important general R2 is married with R3's daughter. R3 is also a famous doctor, treating many generals and leaders.

Name	Freq.	Name	Freq.	Name	Freq.	Name	Freq.
豐臣秀吉(P) Toyotomi Hideyoshi	184	島左近(P) Shima Sakon	172	徳川家康(P) Tokugawa Ieyasu	57	初芽(P) Hatsume	56
長政(P) Nagamasa	47	近江(L) Oumi	32	淀殿(P) Yododono	24	佐和山(L) Sawayama	21

Table 1: Co-occurrence analytics of R1.

Top	Word2vec (skip-gram)	Word2vec (CBOW)	Node2vec	Node2vec (time decay)
1	曲直瀬道三(P) Manase Dosan	曲直瀬道三(P) Manase Dosan	上野(L) Ueno	島左近(P) Shima Sakon
2	徳川家康(P) Tokugawa Ieyasu	井伊直政(P) Ii Naomasa	佐和山(L) Sawayama	北庵(P) Hokuan
3	井伊直政(P) Ii Naomasa	朝鮮(L) Korea	浅野家(O) Asano Clan	熊本(L) Kumamoto
4	京都(L) Kyoto	上野(P) Ueno	豊臣秀吉時代(T) Toyotomi Reign	近江(L) Oumi
5	伏見(L) Fushimi	徳川家康(P) Tokugawa Ieyasu	織田信長(P) Oda Nobunaga	石見(L) Iwami

Table 2: Top 5 related nodes of R1.

3.2 Relation Extraction on Explicit and Implicit

The top 5 related persons of R1 from co-occurrence measuring, word2vec and node2vec are shown in Table 3.

Top	Co-occurrence	Word2vec (skip-gram)	Word2vec (CBOW)	Node2vec	Node2vec (time decay)
1	**豊臣秀吉** **Toyotomi Hideyoshi**	曲直瀬道三 Manase Dosan	曲直瀬道三 Manase Dosan	織田信長 Oda Nobunaga	**島左近** **Shima Sakon**
2	**島左近** **Shima Sakon**	**徳川家康** **Tokugawa Ieyasu**	井伊直政 Ii Naomasa	**豊臣秀頼** **Toyotomi Hideyori**	北庵 Hokuan
3	**徳川家康** **Tokugawa Ieyasu**	井伊直政 Ii Naomasa	**徳川家康** **Tokugawa Ieyasu**	大野治長 Ono Harunaga	大蔵卿局 Okurakyonotsubone
4	初芽 Hatsume	**島左近** **Shima Sakon**	大野治長 Ono Harunaga	武田信玄 Takeda Shingen	**徳川家康** **Tokugawa Ieyasu**
5	長政 Nagamasa	池田輝政 Ikeda Terumasa	淺野幸長 Asano Yoshinaga	黒田長政 Kuroda Nagamasa	**豊臣秀頼** **Toyotomi Hideyori**

Table 3: The list for top 5 related persons of R1 by five methods.

Explicit Case: Toyotomi Hideyoshi (R4) is R1's lord, Shima Sakon (R2) is R1's most important general, and Tokugawa Ieyasu (R5) is R1's biggest enemy in novel. From Table 3, we found that R2, R4 and R5 appear in co-occurrence and word2vec top 5 related persons list but not in node2vec list.

Implicit Case: Toyotomi Hideyori (R6) is R4's son, and he is Toyotomi clan's heir. As Officer of Toyotomi clan, R1 wants to protect R6 because R5 is planning to destroy Toyotomi clan. As a matter, R1 has a close relation with R6, or Toyotomi clan. Nevertheless, R1 doesn't interact much with R6 in novel. That's why word2vec and co-occurrence can't detect this relation. But with neighborhood nodes, node2vec can find it.

With time decay: Although node2vec can't find explicit relation, time decay node2vec method can detect both two cases above. There's another implicit example: Okurakyonotsubone (R7) is close to R4's wife Yododono (R8). There's a fight in novel between R4's two wives, and R1 stands by R8. Also, R1 is close to R7's son. That's why R7 appear in time decay node2vec list.

Time sequential case: In Figure 2, we find other implicit relations in novel. First two Chapter introduces R1's background, so there are his hometown L1, his enemy R5, his general R2, and R2's ex-lord Tsutsui Junkei; in chapter 3, R1 is involved in the fight between R4's two wives, so many related women names appear. Chapter 4 to 6 talks about the story only in Toyotomi clan, so the main relations of R1 only change slightly.

The important thing is that the author of novel likes to comment and compare locations or people. So ancient and modern location names or different time period people may all appear in this novel.

4 Conclusion

We have presented the SeqTextVis system for time sequential text data to help understanding relations. The SeqTextVis is a very powerful system for analyze and visualize relations. Therefore, SeqTextVis is strong way to find insights. In the future, we propose that using an advanced approach to visualize the mix relations (both explicit and implicit) on a dynamic network graph.

References

[1] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean. Distributed representations of words and phrases and their compositionality. Proceedings of the NIPS 2013, page 3111-3119. December 2013.

[2] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. Proceedings of the EMNLLP 2014, pages 1523-1543, October 2014.

[3] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, and N.A. Smith. Retrofitting word vectors to semantic lexicons. Proceedings of the NAACL 2015. pages 1606–1615, May 2015

[4] A. Grover and J. Leskovec. Node2Vec: Scalable feature learning for networks. Proceedings of the SIGKDD 2016. pages 855–864. ACM, August 2016.

[5] L. Lin, S. Chen, F. Hong, C. Lai, S. Chen, and X. Yuan. GraphLDA: Latent dirichlet allocation-based visual exploration of dynamic graphs. Proceedings of the IEEE PacificVis 2017, April 2017.