IEICE TRANS. INF. & SYST., VOL.E97–D, NO.8 AUGUST 2014

LETTER *Special Section on Cyberworlds*

Constructing Social Networks from Literary Fiction

Jong-kyu SEO[†a)], **Sung-hwan KIM**[†b)], ***Nonmembers***, ***and*** **Hwan-gue CHO**[†c)], ***Member***

SUMMARY A social network is a useful model for identifying hidden structures and meaningful knowledge among social atoms, which have complicated interactions. In recent years, most studies have focused on the real data of the social space such as emails, tweets, and human communities. In this paper, we construct a social network from literary fiction by mapping characters to vertices and their relationship strengths to edges. The main contribution of this paper is that our model can be exploited to reveal the deep structures of fiction novels by using graph theoretic concepts, without the involvement of any manual work. Experimental evaluation showed that our model successfully classified fictional characters in terms of their importance to the plot of a novel.

key words: *social network, text analysis*

1. Motivation

Linguistic analysis is the main method for identifying crucial features of literature [3], [4]. Besides linguistic analysis, automated text analysis such as co-occurrence analysis can also be applied to text categorization [2]. In this paper, we propose an algorithm to construct social networks from literary fiction by evaluating the strength of relationships between characters. We propose a graph model, called Social Network from Fiction (SNF), to reveal the structure of novels without any manual work.

2. Structure of Fictional Cyber Space

In our method, we locate the appearance of characters by scanning the entire text space. However, it is not possible to deduce semantically identical characters since the main characters have several aliases. Hence, we only consider the full names of characters as we have observed that full names are sufficient to obtain the correct social network model from the text.

Let $T = \langle S_1, S_2, \ldots, S_n \rangle$ be a document that consists of a sequence of statements $\langle S_i \rangle$, and let $C(T) = \{c_1, c_2, \ldots, c_k\}$ denote the set of all characters described in the text T.

We have to locate the appearance of characters in the text space before measuring the interactions among them. This can be achieved using known character indices or the Stanford Name Entity Recognizer. First, let $POS_T(c_i)$ be the set of indices for the statements that contain c_i.

$$POS_T(c_i) = \{j \mid \text{if } c_i \text{ is a word of } S_j\} \tag{1}$$

Manuscript received November 10, 2013.

[†]The authors are with Pusan National University, Korea.
a) E-mail: maniasjk@pusan.ac.kr
b) E-mail: sunghwan@pusan.ac.kr
c) E-mail: hgcho@pusan.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E97.D.2046

Next, we define the interaction of two characters(c_i, c_j) as follows:

$$Inter(c_i, c_j) = \sum_{l,m} \alpha^{|l-m|}, \; |l-m| \le \beta$$
$$\text{where } l \in POS_T(c_i), \; m \in POS_T(c_j) \tag{2}$$

According to the above equation, the interaction of two characters is measured on the basis of the statement distance between them. In other words, if two characters appear in the same statement, then we can say that they are highly related. α and β are control parameters that adjust the amount of relationships depending on the statement distance between the characters.

3. Experiments with Fictional Works

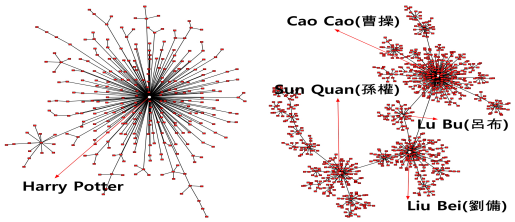
We conducted an experiment with more than 20 novels having different lengths, genres, and authors. Table 1 lists the basic statistics of some of these novels.

3.1 Spanning Tree of the Fictional Social Network

In the SNF graph $G(V, E)$, each character is mapped to a vertex. Each vertex is then connected to other vertices by weighted edges that correspond to the distances between character-pairs. Generally, the SNF graph appears complicated and hard to decipher. The set of sub-graphs (or trees) that connect all the graph vertices together are called the spanning trees of the graph. The minimum spanning tree (MST) of the graph is a spanning tree with a weight lower than every other tree in the graph. Spanning trees and MSTs can be used to significantly reduce the number of graph edges while maintaining the connectivity of the entire graph [1]. Figure 1 represents the MST graph for each novel. The names in each figure represent the protagonists in the novel. Our spanning tree could be used to show difference between heroic novels (Fig. 1 (a)) and epic novels (Fig. 1 (b, c, d)) in terms of the length of diametral path or the number of leaf nodes.

Table 1 Test novels for the experiment.

Title	Statements	Characters	Edges
War and Peace (W.P.)	30,912	234	4,303
Three Kingdoms (T.K.)	121,779	912	36,650
Harry Potter (H.P.)	85,006	287	8,526
The Earth (T.E.)	176,387	496	16,347



(a) Harry Potter (b) Three Kingdoms

(c) The Earth (d) War and Peace

Fig. 1 MST graph of each novel. From the MST graph, we can understand the characteristics of the distribution of the protagonists in each novel.

Table 2 The number of pronouns between characters.

	Harry Potter		War and Peace	
Pronouns	Counts	Ratio	Counts	Ratio
0	23,528	40%	3,826	29%
1	15,552	26%	2,694	20%
2	8,847	15%	1,843	14%
3	4,777	8%	1,238	9%
4	2,651	4%	917	7%
5	1,537	3%	626	5%

In this paper, we only considered the full name of all fiction characters without handling any aliases or personal pronouns such as {he, she, him, they, his et al...}. We counted the number of pronouns in between two adjacent character names we took in this paper. We found one interesting "power-law" features on the frequency of personal pronouns where about 1/3 of all intervals of adjacent character names has no personal pronouns. Our experiment (Table 2) showed that pronoun frequency decreases exponentially, which means most of character full names are adjacent enough to make the whole plot understandable by general readers. That implies the number of personal pronouns are strongly limited between two full named characters, so our model without considering personal pronouns could be one simplified model to approximate the whole plot.

3.2 Level of Importance of Fictional Characters

Let us denote the leaf vertices of the MST as $0-leafs$. When we remove all the $k-leaf$ vertices, we obtain a smaller tree in which the leaf vertices are called $(k+1)-leaf$ vertices. All $i-leaf$ vertices can be defined by using this inductive procedure. While conducting the experiment, we found that $j-leaf$ characters perform a more crucial role as compared to $k-leaf$ characters, where $j > k$. Figure 2 shows the $k-leaf$ algorithms in the tree.

Table 3 lists the counts that indicate the depths of characters in an MST graph. The results obtained from the experiment indicate that most novels are led by a few protagonists and many peripheral characters.

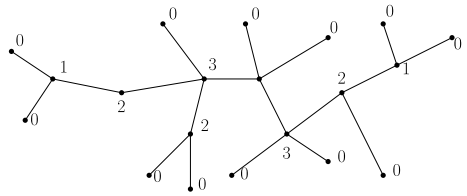


Fig. 2 k-leaf construction of a tree. Vertices are labeled with their corresponding k value.

Table 3 Characters at depth k of the MST. (a) {ZhuGe Liang(諸葛亮)}, (c) {Pierre}, (d) {Harry Potter}.

Depth	T.K.	T.E.	W.P.	H.P.
1	705	347	184	236
2	145	84	37	40
3	40	27	15	8
4	13	21	10	3
5	9	17	7	$^{(d)}1$
6	3	15	4	0
7	$^{(a)}1$	13	$^{(c)}1$	0
8	0	10	0	0
9	0	7	0	0
10	0	3	0	0
11	0	$^{(b)}1$	0	0

4. Conclusion and Future Work

In this paper, we proposed that the social network skeleton of fiction novels can be extracted simply by observing the syntactic structure of the text, such as the distance distribution of words (character names) in the case of lengthy novels. In future, we plan to investigate new features of SNFs that are based on power law distributions. These new features will help in developing a new inference algorithm for understanding the semantic structure of novels using SNFs.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2011-0015359).

References

[1] C. Chen and S. Morris, "Visualizing evolving networks: Minimum spanning trees versus pathfinder networks," 9th IEEE Symposium on Information Visualization, pp.67–74, 2003

[2] X. Luo and A. Zincir-Heywood, "Combining word based and word co-occurrence based sequence analysis for text categorization," Proc. Machine Learning and Cybernetics, vol.3, pp.1580–1585, Aug. 2004.

[3] J. Rydberg-Cox, "Social networks and the language of greek tragedy," J. Chicago Colloquium on Digital Humanities and Computer Science, vol.1, no.3, pp.1–11, 2011.

[4] D.K. Elson, D. Nicholas, and K.R. McKeown, "Extracting social networks from literary fiction," Proc. ACL, pp.141–147, 2012.