Structural Analysis on Social Network Constructed from Characters in Literature Texts

Gyeong-Mi Park, Sung-Hwan Kim, and Hwan-Gue Cho, Pusan National University, South Korea

Abstract—Recently we witnessed that the social network analysis focusing on social entities is applied in the social science and web-science, behavioral sciences, as well as in economics, marketing. In this paper we present one method to construct and structure analysis the social network from literary fictions by a simple lexical analysis. And we will show that those literary social graph, shows the power law distribution of some features, which is the typical characteristics of complex systems. We newly proposed the concept of the kernel of literary social network by which we can classify the abstract level of protagonists appeared in fictions. And we also studied the connectivity of social network based on statement distance distribution of characters. Our study shows that the metric distance among characters written in linear text is very similar to the intrinsic and semantic relationship described by fiction writers, which implies the proposed social network from fictions could be another representation of literary fiction. So we can apply other scientific and quantitative approach by analyzing the concrete social graph model extracted from textual data.

Index Terms—Social Network, Complex System, Literary Fiction, Character Graph, Power law.

I. INTRODUCTION

Extracting useful information from a large textual repository is getting essential in data mining field. One difficulty in this work is how to deal with the various kinds of natural languages. Most work has based on English based texts, but recently some other languages have shown interesting result.

Some previous work proposed the open problems on the dynamic property of social network extracted from text[9][11]. In this view of new framework, we have tried to investigate the temporal structure of social network, since the literature fiction spans more than years or more than one human generation such as ”War and Peace” written by Tolstoy.

Our basic idea is that we can regard the complicated text(e.g. long literary fictions) as the typical complex system such as human society and the huge ecosystem in Earth. In novel lots of characters are interacting in the text space which consists of a sequence of words. So if we compute the distance of two characters over the linear text, then the new kinds of social relationship can be extracted. And we can construct the social network from the distance matrix of characters appeared in literatures. So we need to characterize the structure of literary fictions in terms of complex system.

By analyzing the social network extracted from text, we can determine the importance of fiction characters by a mechanical way of text analysis without any complicated semantic analysis. This will help us to understand the deep and common structure of literature regardless of written languages. According to the basic language theory of Noam Chomsky, there must be some common ground structure over almost natural languages and narrative stories made by writers. So mining the topological structure of social graph from text will help us to find other important features in a simple graph analysis and can summarize the whole story or can measure the complexity of literature fiction by measuring the degree of interactions among protagonists or between protagonists and other boundary persons. This implies the proposed approach will greatly help to reveal the semantic structure of fictions by constructing the concrete social network and topological analysis of the graph.

In next section, we will survey the related work on these topics. Section 3 will devoted to give some definitions and preliminary concepts. Section 4 will show the concrete algorithm for constructing the social network from text. Also we should open how we prepared the testing data(4 their-person novels) and how to preprocess to make them fit in experiment. Experiment results are shown in Section 6, where we can assure that these social network is quite similar to the general complex systems. And finally the summarized conclusion and future open problems will be exposed.

II. RELATED WORK

A. Computational Analysis on Literature

It is general that the computer-based literary textual analysis has typically performed at the word level[1]. This work has focused the discover authorial style and the lexical patterns of word use. Those computational linguistic work has contributed to validate or repute the basic literature theories that the novel is a literary form which tries to produce an accurate representation of the social world. According to that theory, it is believed that theories about the relation between novelistic form (the workings of plot, characters, and dialogue, to take the most basic categories) and changes to real-world social space.

Recently researchers started to study property of co-occurrence words in natural language space[7,8,10,12,13]. Since all words in a text are closed related though they are arranged in the linear sequence. The co-occurrence pattern of some pair of words may reveal the important features hidden in some texture[17]. One interesting application of these co-occurrence analysis gave the clue to identify spam or vicious notifying message[8,10]. One application of word appearance pattern is biomedical text mining whose goal is to find the biologically and medically meaningful knowledge only by analyzing the huge textual data(mainly academic papers)[2,16]. For example, hidden functions of genes can be