# Complex System Analysis of Social Networks Extracted from Literary Fictions

Gyeong-Mi Park, Sung-Hwan Kim, Hye-Ryeon Hwang, and Hwan-Gue Cho

*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 the social network from literary fictions by a simple lexical analysis, not using the complex natural language processing tools. And we will show that those social graphs, saying literary social graph, shows the power law distribution of some features, which is the typical characteristics of complex systems. We showed that the social network extracted from literary data reflects the similar network structure which was semantically designed by authors of fictions. And 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. 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 including East-Asian languages have shown interesting result. Due to the recent prevailing SNS (Social Network Service), people try to extract the sentiment from text data shared in on-line users. After obtaining attitude from one individual, we are able to identify the connection structure of elements in a community. The main issues of these sentiment analysis is how to identify the polarity of adjectives based on conjunctions linking them in a large corpus. And another hot research area is mining information over online discussion by observing discussion threads. By mining used words or related replying patterns among discussion thread, we can identify the friendly group or conflicting group.

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 third-person novels) and how to preprocess to make them fit in experiment. Experiment results are shown in Section 6, where we can assure that this 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 to discover authorial style and the lexical patterns of word use. This 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

Manuscript received October 15, 2012; revised December 29, 2012.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.2012-0005518). Corresponding should be addressed to Hwan-Gue Cho (hgcho@pusan.ac.kr).

Gyeong-Mi Park, Sung-Hwan Kim and Hwan-Gue Cho are with Dept. of Computer Engineering, Pusan National University, Busan, South Korea (e-mail: miya11@pusan.ac.kr; sunghwan@pusan.ac.kr; hgcho@pusan.ac.kr).

DOI: 10.7763/IJMLC.2013.V3.282

co-occurrence words in natural language space [2]-[4]. Since all words in a text are closed related to how 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. One interesting application of this co-occurrence analysis gave the clue to identify spam or vicious replying message [3]. 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) [5], [6]. For example, hidden functions of genes can be estimated or predicted by biomedical text mining tools, which helps biologists to make a new drug fast and efficiently.

### *B. Extracting Social Networks from Text*

Typical social network studies based on textual data focus on the structured data including email and Tweet messages and phone SMS messaging texts for social network construction [7], [8]. The main goal of this work is to find the most influential person. And the propagating patterns of texts (messages) are the key subjects of this work. And pure combinatorial studies has been started to discover the frequency of some designated words [1], [9]. One of typical work on this line is the Zipf's law. In any natural language, the generic Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. That implies the most frequent word appears approximately twice as often as the second most frequent word, three times as often as the third most frequent word.

Till now, little work has been done on extracting social relations between individuals from text. One notable work has published on constructing conversational networks from literature [10]-[13]. Elsan et al proposed a method for extracting social networks from nineteenth-century British novels [14]. Link two characters are given if they are in conversation or not in the text. So they successfully made the conversational network to imply the closeness of protagonists. They tried to validate some literature-related hypotheses. It means they replaced the social network by conversational network to evaluate literary theories. In their graph model, the named characters are represented vertices represent characters and edges are added to indicate dialogue interaction among characters. They set the edge weight proportional to the amount of interaction. The two main constraints of conversational network are followings. 1. The characters should be in the same place at the same time; 2. The characters should speak turn by turn.

Their conclusion is twofold. First there is an inverse correlation between the amount of dialogues in a novel and the number of characters in that novel. Second a significant difference of social interaction in the 19 century English novel is basically is geographical. One difference of Elsan's approach and ours is that we do not have interests in validating any literature theory. Another interesting work should be noted. Hassan et al have studied the positive and negative influential social network by analyzing the sentiment words in literature [1]. They have used a larger amount of online discussion posts. Their experiments showed that their method easily constructed networks from text with high accuracy, which connects out analysis to social psychology theories of signed network, namely the structural balance theory.

Our purpose is to give any engineering or computational method to reveal the hidden structure of literary fictions. And also we will show the social network extracted from fictions is also one of complex system where rich get richer and poor get poorer by showing various statistics gathered in word analysis.

## III. Preliminary

### *A. Distance metric for Characters on Literary*

Since every social network is characterized by entities (nodes) and the distance function, it is important to define the valid metric function. In our social network model from literary text, the node (entity) is textual words denoting the persons described in the corresponding text. Then we need how to make them connected by a proper metric (distance function). We will explain this in the following.

Let T denote one literary fiction text. Each *T* consists of a set of consecutive statements$< s_i >$. And each statement $s_i$can be decomposed of a sequence of words $< \mathrm{w_{i,j}} >$, where $\mathrm{w_{i,1}}$ is the first word of statement $\mathrm{s_i}$. And finally all words $\mathrm{w_{i,j}}$ consist of a sequence of letters$< l_{\mathrm{i,j,k}} >$. $|\mathrm{S_i}|$ denotes the number of words( is the statement $\mathrm{s_i}$. In a similar way let $|\mathrm{T}|$ denote the number of statements in that text *T*. And we define statement rank of word $w^* = w_{i,j}$, *staterank*$(w^*) = i$ and the word rank, *wordrank*$(w_{i,j})$ is defined as the followings.

$$\mathrm{wordrank}(w_{i,j}) = \sum_{k=1}^{i-1} |sk| + j - 1, \qquad (1)$$

By the preprocessing step, we already prepared the name lists of all characters appeared in the fiction. In the following we denote $\{C_i\}$ the set of all characters in T.Since $C_i$alsoappeared in text with different form, we need to locate the *wordrank*().

Now we have to define the distance metric of statements and words. $dist_X(p,q)$ denote the distance between two entities p and q over $X$ metric space, where $X$ could be thestatement space or word space. Let me show a few examples for this distance function.

In the following let $dist_{state}(w_{i,p}, w_{j.q})$ denote the numberof statement which lies in between two different statements $s_i$ and $s_j$. That is defined $j - i + 1$ if $i < j$. That means if two words locate in a same statement, then the statement distance between two words is zero. And we define$dist_{word}(X,Y) = |wordrank(Y) - wordrank(X)|$

As you expect, if two characters X and Y are closed relatedin story, then $dist_{state}(X,Y)$would be short compared to the pair of loosely related characters. So the relatedness of two characters are expected to depend on the average distance of $dist_{state}(X,Y) < C_0$ , also the frequency of $dist_{state}(X,Y) < C_0$ where $C_o$ is one threshold constant to determine if the person is in the working narrative space.

Let us give one example to help the definitions above. Here we have a tiny text with 5 simple statements.

TABLE I: STATEMENT DISTANCE SAMPLE

| # | word sequence of each statement |
|---|---|
| 1 | Harry walked across Hedwig's cage. |
| 2 | Hedwig and Harry had been absent for two nights. |
| 3 | Dursley worried about Hedwig and Lily. |
| 4 | Dursley had pretended for ten years. |
| 5 | Because Lily and James Potter had not died. |

Harry appears two times at statement 1, 2 which are denoted $H_1, H_2$. And we denote $D_1, D_2, D_3$ for three Hedwig from the first statement. Then we see that$dist_{state}(H_1, H_2) = 0, dist_{state}(H_1, D_3) = 3$ and $dist_{state}(H_{1,}D_3) = 16$ since there are 16 words between $H_1$ and $D_3$ over the whole textspace. And it is easy to know that workdrad$(D_1 = 4$,wordrank$(H_2) = 8$.

## IV. CONSTRUCTING SOCIAL NETWORK FROM TEXT

Now we will explain how to extract the social network from literary fiction. Let $T$ be a text to be analyzed and $G_T(V, E)$ be the corresponding graph constructed from T. All words representing characters are mapped to vertices of $G, V$. Next we decide how to make each characters (vertices) connected by assigning edge relation. Let $v_i, v_j$be two different vertices representing two characters. We give the edge connection as $(v_i, v_j)$ if the weight between $v_i$ and $v_j$ ,weight$(v_i, v_j)$, is less than a threshold constant c*, which isgiven by user to meet the own purpose.Since there are twodistance measures based or statement or word entity, we applied $\mathrm{dist}_{\mathrm{state}}()$ measure for weight function in this paper, which was decided by comparative work against $\mathrm{dist}_{\mathrm{word}}()$. More formal computation is given. Since two characters X and Y appears more than once adjacently, we find all adjacent pair of X and Y in text. The edge weight of for two characters *X* and *Y* is given followings.

$$\mathrm{weight(X, Y)} = \sum_{w_x \equiv X, w_{y=Y}} \alpha^k, \qquad (2)$$

where $k = dist_{state}(w_x, w_y)$, and $0 < \alpha < 1$ , a control constant. In this experiment we set $\alpha = 0.7$. So if two characters are in a statement, then statement distance should be zero(0), so the weight is 0.7. If the statement distance between two words (characters) is 5, then the weight should $0.7^5 = 0.16807$. If we set the control variable $\alpha = 0.9$ then, $\alpha^5 = 0.59049$. We can get the extremal case if $\alpha = 1$, which disregards the effect of distant word in weight calculation. Next if weight$(X, Y) > c_0$ then we give the connecting edgebetween $Y$ and $Y$. In implementation, we do not considerany pair of characters which are separated with more than 10 statements.

To summarize, we can construct the social network graph by adjusting two control variables, $c_0, \alpha$. If we increase $c_0$or α then we get the denser social network structure and we will get the sparse network if we lower $c_0$. So generally let us $G_T(V, E, \alpha, c_0)$represent the social network extractedfrom the literature T with two control parameters $\alpha, c_0$.

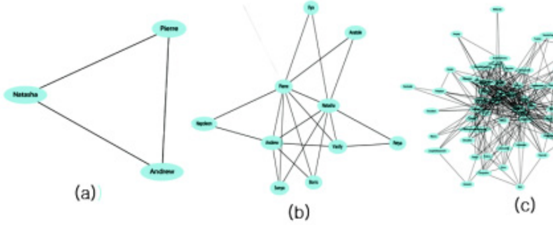

Fig. 1. Three social networks extracted from fictions (a) social network extracted from crime and punishment with the threshold constant ca =23.56. (b) Another social network from crime and punishment with more a loose threshold cb =23.56.constant compared to case (a). (c) We can get the nearly complete social network of the fiction with a threshold cb = 23.5

### A. Kernel Construction for Literary Social Network

Since all fictions consist of a few protagonists and other boundary persons, it is interesting to observe the topological property of each character according to the role importance in the story. So we propose the concept of network kernel which reflects the role importance of literary text space.

In the following we show one example social network $G_T(V, E, \alpha, c_0)$ .We call the $1 - degree\ preson(node)$ whose graph degree is 1. So there are $1 - deree\ person = \{u, t, h, a, p, q, s, e, r\}$ who can be considered some boundary persons in fiction since their interaction is quite low.

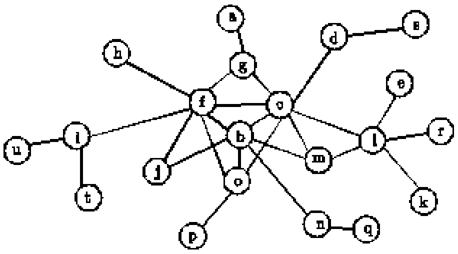

Fig. 2. $\mathrm{G_T(V, E, \alpha, c_0)}$, the original social networkgraph with 21 characters, that is $\mathrm{Kernel_0(G, t)}$.

Now we will explain how to construct the $Kernel_k(G, t)$of social networks $G$, constructively and successively. We set $Kernel_0 = G_T(V, E, \alpha, c_0)$for base condition. And wedenote *(k+1)-th* Kernel as $Kernel_{k+1}($ which was derived from $Kernel_{k+1}(G, t)$. The procedure is simple. We delete all vertices of $Kernel_k(G, t)$ whose graph degree is less than or equal to t. So $Kernel_{k+1}(G, t)$ should be smaller than $Kernel_{k+1}(G, t)$ . When we get $Kernel_{k+1}(G, t) = Kernel_k(G, t)$, we say the kernel is stabilized. In Fig. 3, we show $Kernel_1(G, 1)$.

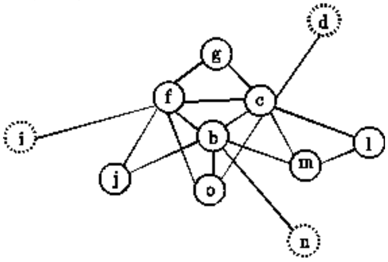

Fig. 3. $\mathrm{Kernel_1(G, 1)}$showing all degree1 nodes(boundary nodes) are removed from$\mathrm{Kernel_0(G, t)}$. The shaded nodes{i, n, d}will beremoved if we construct $\mathrm{Kernel_2(G, 1)}$.

Next we remove all nodes whose degree is less than or equal to 2, which results in the following graph $Kernel_2(G, 2)$ . Note that $Kernel_1(G, 1) = Kernel_2(G, 2)$
$Kernel_3(G, 1)$. If we apply this Kernel extraction process,

the core characters can be isolated from the lots of fictional person without analyzing the complicated semantic analysis of lexical model. The structure of social network Kernel is expected to reveal the hidden structure of fiction story. Some novels are saturated fast if lots of characters are interacted evenly such as the Romance of Three Kingdoms, the Chinese history fiction.

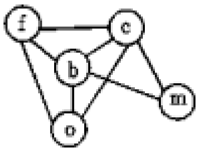

Fig. 4. $\mathrm{Kernel}_2(G,2)$. which was obtained by removing all nodes of $\mathrm{Kernel}_1(G,1)$. with degree 1 and2. And this structure will be keltin constructing$\mathrm{Kernel}_3(G,1)$.

## V. Experiments

### A. Data Preparation

Now we will explain how to locate the word position of fiction characters. First of all we have obtained the list of characters appeared in example testing fictions from Wikipedia and other internet sources. And it should be noted that there is an auxiliary booklet explaining all characters appeared in Korean novel "Land", where more than 500 persons are listed with brief explanation on the role and relationship to other characters and protagonists.

And 26(28) person-pronouns are selected for English-(Korean) literature. The interesting characteristics of person pronouns appeared in between Level-0 name is described in experiment section.

TABLE II: Test Data 1 Novel Text

| Title | Language | Words | Statements | Characters |
|---|---|---|---|---|
| Three Kingdoms | Korean | 642,222 | 121,779 | 1, 285 |
| The Earth | Korean | 1,446,879 | 176,387 | 515 |
| War and Peace | English | 584,618 | 30,912 | 139 |
| Harry Potter | English | 1,279,375 | 126,112 | 655 |

### B. Frequency of Characters

We have enumerated all occurrences of level-0 name characters over literature.

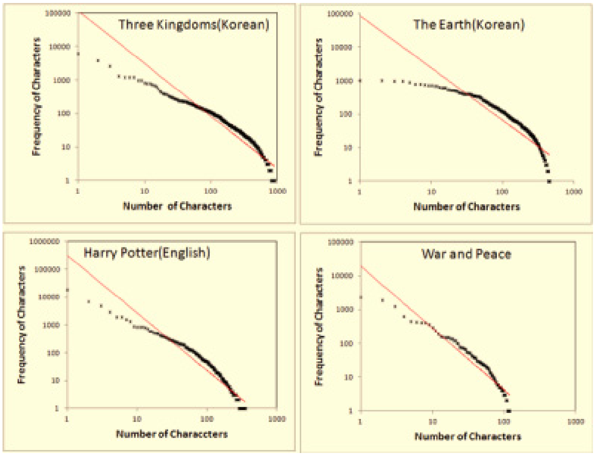

Fig. 5. Frequency of characters in log-log scale. all characters appeared in testing text shows a good approximation of power law distribution

### C. Distribution of Degree and Edge Weights:

Next we show the power law similar feature embedded in the edge weight sum of each node that is the summation of all weights of a designated node. Note that the weight sum of edge has nothing to do with the node degree. If one protagonist interacts only with a person very frequently, the edge weight sum gets higher, but the degree of that protagonist stays at most one. And if a person works frequently among some groups as a role of the bridge of community, saying a person has wide acquaintance, then the number of connected person will increase, that should be concluded as the high degree node.

TABLE III: Kernel of Social Network for War And Peace

| level | Nodes | Edges | Cut Value |
|---|---|---|---|
| 0 | 3 | 3 | 148.38 |
| 1 | 10 | 16 | 45.24 |
| 2 | 36 | 83 | 12.86 |
| 3 | 60 | 142 | 4.35 |

### D. Analysis on the Frequency of Person Pronouns

Since there so many name aliases in the text, it would be interesting to discover the distribution of person pronouncing between two level-0 names. As we expect the most frequent number is one. And there are some level-0 name pairs who keeps more than 13 person pronoun. So we computed the number distribution of person pronouns in between two adjacent level-0 name. The distribution is quite similar to Power Law distribution as was generally kept in complex system.

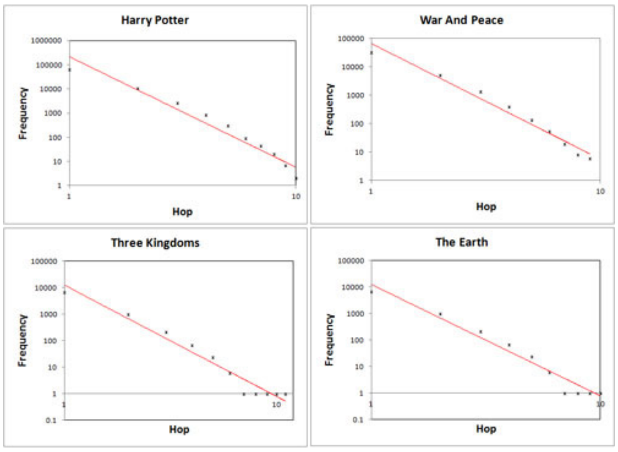

Fig. 6. Person Pronoun appeared in between two adjacently appeared characters

### E. Statement Distance Distribution of Protagonists

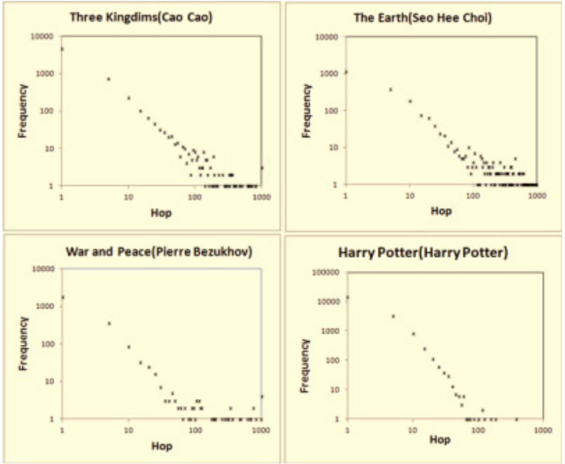

Fig. 7. Statement distance distribution between two adjacent characters

Next we tried to compute the distribution of statement distance for two adjacent characters in the sequential text level. As expected, the number of statements lying in between two adjacent characters shows the power law distribution.

### F. Contribution Summary Experiment Summary

In this paper we consider how to extract the social network from literary text. And we proved that Kernel() of the social network reveals the hierarchical structure of all characters according to the personal importance given by authors. So by controlling some control parameters we can easily the set of protagonist without understanding the story of fiction semantically. That result can be concluded that the simple lexical structure of literary such characters appearance clearly isomorphic to the semantic structure of the fiction. Another interesting fact we revealed is the topological structure of social network preserves the most important and crucial common features of complex system such as the general social(human) network and the huge biological networks including proteins and viruses. That is the power law distribution of relatedness among social identities. Let us summarize the result in the following.

1) We gave one algorithmic procedure to extract the social network from literary text, named literary social network.
2) The extracted literary social network shows quite similar structure in terms of lexical level to the narrative structure of fictions intrinsically.
3) New concept of Kernel() of the social graph help to identify the core characters in the mechanical and algorithmic way. By controlling that Kernel, we can classify all characters appeared according to the importance of each role in the fiction.
4) Our experiment clearly showed that this literary social network is one kind of complex system where most of properties are under power law. For example we showed that the number of interacted persons for each characters is the power law distribution. That means, the more referred persons would be expected to be referred in the future over the text level.
5) It is also interesting to investigate the common and different features between among fictions written by different languages such as Three Kingdoms(Korean), Crime and Punishment(English), Harry Potter(English, Korean).

### G. Future Work and Open Problems

Now we are trying to reveal the dynamic of social network in transition. For example though the fiction "The Romance of Three Kingdoms" has more than 1000 characters referred, the number of working characters are bounded for a fix size text window, which means the social network is so dynamic. If we measure the degree of dynamic development of fiction story, that would be interesting to classify fictions. Or we hope to characterize the topological structure of typical detective story and crime story or fantasy novel.

## References

[1] A. Hassan, A. Abu-Jbara, and D. Radev, “Extracting signed social networks from text,” in *Proc. of the Text Graphs Workshop at ACL*, 2012, pp.4-12.

[2] Y. Fujii, T. Yoshimura, and T. Ito, “Filtering harmful sentences based on three-word co-occurrence,” in *Proc. of Collaboration, Electronic messaging Anti-Abuse and Spam*, 2011, pp. 64-72.

[3] R. Krestel and L. Chen, “Using co-occurrence of tags and resources to identify spammers,” in *Proc. of ECML PKDD Discovery Challenge*, 2008, pp. 38-46.

[4] C. Monojit, C. Diptesh, and M. Animesh, “Global topology of word co-occurrence networks: beyond the two-regime power-law,” in *Proc. of Int. Conf. on Computational Linguistics*, 2010, pp. 162-170.

[5] S.-Y. Bong and K.-B. Hwang, “Keyphrase extraction in biomedical publications using mesh and intraphrase word co-occurrence information,” in *Proc. of the ACM DTMBIO*, 2011, pp.63-66.

[6] Y. Fujii, T. Yoshimura, and T. Ito. “Filtering harmful sentences based on three-word co-occurrence,” in *Proc. of 8th Annual Collaboration Electronic messaging Anti-Abuse and Spam Conference*, 2011, pp. 64-72.

[7] C. L. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*, Empirical Press, 2004.

[8] D.-H. Kim, G. Rodgers, B. Kahng, and D. Kim, “Modelling hierarchical and modular complex networks: division and independence,” *Physica A: Statistical Mechanics and its Applications*, vol. 351, no. 2-4, pp. 671-679, 2005.

[9] X. Luo and A. Zincir-Heywood, “Combining word based and word co-occurrence based sequence analysis for text categorization,” in *Proc. of the Machine Learning and Cybernetics*, vol. 3, pp. 1580-1585, Aug. 2004.

[10] Y.-M. Choi, “Greek myth as a complex network,” *The Korean Physical Society*, vol. 49, no. 3 pp. 298-302, 2004.

[11] S.-R. Kim, “Complex network analysis in literature: Togi”, *The Korean Physical Society*, vol. 50, no. 4, pp. 267-271, 2004.

[12] Y.-K. Lee, J.Ku, and H.Kim, “Analysis of network dynamics from the romance of the three kingdoms,” *The Journal of Korea Contents Association*, vol. 9, no. 4, pp. 364-371, 2009.

[13] J. Rydberg-Cox, “Social Networks and the language of greek tragedy,” *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, vol. 1, no. 3, pp. 1-11, 2011.

[14] D. K. Elson, D. Nicholas, and K. R. McKeown, “Extracting social networks from literary fiction,” in *Proc. of the Association for Computational Linguistics*, 2012, pp. 141-147.

**Gyeong-Mi Park** is at BK21 Center for U-Port IT Research and Education. She receive the Korean National Open University, Korea, and the M. S and Ph D. degrees from Pukyoung National University, Korea. Her research interests are computer vision and information retrieval.

**Sung-Hwan Kim** is a M.S student in Pusan National University. He received the B.S. degree from Pusan National University. His research interests are information retrieval and sequence processing.

**Hye-Ryeon Hwang** is a M.S student in Pusan National University. She received the B.S degree from Kyungsung University. Her research interests are bioinformatics and data visualization.

**Hwan-Gue Cho** is a Professor in Pusan National University. He received the B.S. degree from Seoul National University, Korea, and the M.S and Ph.D. degrees from Korea Advanced Institute of Science and Technology, Korea. His research interests are computer algorithms and bioinformatics.