Structural Analysis on Social Network Constructed from Characters in Literature Texts

Gyeong-Mi Park
BK21 Center for U-Port IT Research and Education, Pusan National University, Korea
miya11@pusan.ac.kr

Sung-Hwan Kim and Hwan-Gue Cho
Dept. of Computer Engineering Pusan National University, Korea
Email: { sunghwan, hgcho } @pusan.ac.kr

***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 graphs, 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 [9], [11].

Some previous work proposed the open problems on the dynamic property of social network extracted from text. 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 Tosltoy.

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. 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 networks are 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

Manuscript received October 30, 2012; revised November 22, 2012; accepted November 25, 2012.

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2010-371-B00008). Corresponding should be addressed to Hwan-Gue Cho (hgcho@pusan.ac.kr).

doi:10.4304/jcp.8.9.2442-2447

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 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 replying 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 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 [6], [8], [14]. The main goal of these work are 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], [12]. 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 [3], [10], [12], [16]. Elsan et al proposed a method for extracting social networks from nineteenth-century British novels [5]. 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 gets richer and poor gets poorer by showing various statistics gathered in word analysis. The previous work focused on the static structure of social network such as degree distribution and connection form. But this paper also studies how the structure of social network varies along chapters by chapters.

III. PRELIMINARY

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 $< w_{i,j} >$, where $w_{i,1}$ is the first word of statement s_i. And finally all words $w_{i,j}$ consists of a sequence of letters $< l_{i,j,k} >$. $|S_i|$ denotes the number of words(is the statement s_i. In a similar way let $|T|$ 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 define the as followings.

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

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 also appeared in text with different form, we need to locate the wordrank().

Now we have to define the distance metric on statements and words. $\text{dist}_X(p, q)$ denote the distance

between two entities p and q over X metric space, where X could be the statement space or word space. Let me show a few examples for this distance function. In the following let $\mathrm{dist}_{\mathrm{state}}(\mathrm{w_{i,p}}, \mathrm{w_{j.q}})$ denote the number of statement which lies in between two different statements $\mathrm{s_i}$ and $\mathrm{s_j}$. That is defined $\mathrm{j} - \mathrm{i} + 1$ if $\mathrm{i} < j$. That means if two words locate in a same statement, then the statement distance between two words is zero. And we define. $\mathrm{dist}_{\mathrm{word}}(\mathrm{X}, \mathrm{Y}) = |\mathrm{wordrank(Y)} - \mathrm{wordrank(X)}|$

As you expect, if two characters X and Y are closed related in 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. 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 text space. And it is easy to know that workdrad($D_1 = 4$, wordrank(H_2) $= 8$.

TABLE I.
STATEMENT DISTANCE SAMPLES

#	*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.*

IV. CONSTRUCTING SOCIAL NETWORK FROM TEXT

A. Constructing Social Network from Literary Text

Now we will explain how to extract the social network from literary fiction. Let T be a text to be analyzed and $\mathrm{G_T(V, E)}$ be the corresponding graph constructed from T. All words representing characters are mapped to vertices of $\mathrm{G, V}$. Next we decide how to make each characters (vertices) connected by assigning edge relation. We give the edge connection as $(\mathrm{v_i}, \mathrm{v_j})$ if the weight between $\mathrm{v_i}$ and $\mathrm{v_j}$, weight$(\mathrm{v_i}, \mathrm{v_j})$, is less than a threshold constant $\mathrm{c^*}$, which is given by user to meet the own purpose. Since there are two distance measures based or statement or word entity, we applied $\mathrm{dist_{state}}()$ measure for weight function in this paper, which was decided by comparative work against $\mathrm{dist_{word}}()$. 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 $\mathrm{k} = dist_{state}(w_x, w_y)$, 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 $\mathrm{weight(X, Y)} > c_0$ then we give the connecting edge between Y and Y . In implementation, we do not consider any 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, α. If we increase c_0 or α then we get the more dense social network structure and we will get the sparse network if we lower c0. So generally let us $G_T(V, E, \alpha, c_0)$ represent the social network extracted from the literature T with two control parameters α, c_0.

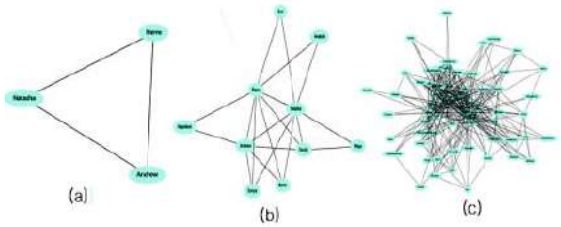

Figure 1. Three Social Networks extracted from fictions (a) Social network extracted from Crime and Punishment with the threshold constant ca = 23.56. (b) Another social network from \Crime and Punishment with more a loose threshold cb = 23.56.constant compared to case (a). (c) We can get the nearly complete social network of the fiction with a threshold cb = 23.5

B. Kernel Construction for Literary Social Network

Since all fictions consists 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,α,c_0).

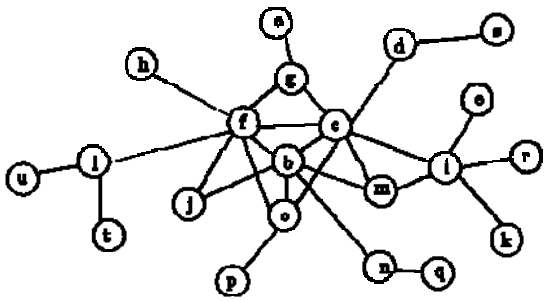

Figure 2. G_T (V,E,α,c_0), the original Social Network Graph with 21 characters, that is Kernel_0 (G,t)

We call the $1-degree\ preson(node)$ whose graph degree is 1. So there are 9 $1-deree\ person = \{u,t,h,a,p,q,s,e,r\}$ who can be considered some boundary persons in fiction since their interaction is quite low. 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 we denote *(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 the fol-lowing Figure, we show $Kernel_1(G,1)$.

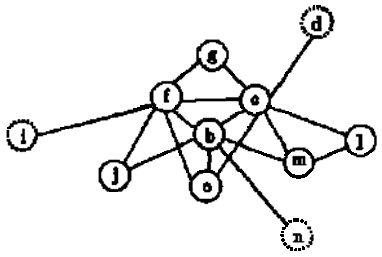

Figure 3. Kernel_1 (G,1) showing all degree 1 nodes(boundary nodes) are removed from Kernel_0 (G,t). The shaded nodes {i,n,d} will be removed if we construct Kernel_2 (G,1).

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 "The Earth", where more than 500 persons are listed with brief explanation on the role and relationship to other characters and protagonists.

TABEL II TEST DATA 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. Connectedness of Social Graph according to Influence Region

Let we have s_i statements from a text. As we explained, if a character c_x appears in s_w statement, then we will check other characters which appear in next statements. The number of statements of forward statements determines the graph connection. If we look ahead a smaller statement forward, the social connection should be sparse, even may be disconnected. Let the number of statements to be looked forward be the influence region for a character of s_i.

In this paper we experiment the connectedness effect by controlling the influence region. We have check three more statement forward in the previously depicted social graph, which means the influence region is 3. Now we will try to find the connectedness by controlling the influence region 0,1, to 10. Ideally speaking, all characters referred in a fiction should be connected unless that fiction consists of independent chapter such as omnibus essay. It is very interesting to find the minimum influence region which guarantees the connectedness of social graph extracted from a text.

In the following Table, we show the component size with respect to the influence region $\mathrm{d} = 0,1,2,3,4,5$, where $\mathrm{d} =$ implies we only considered the co-occurrence characters in a single statement. Experiment shows the minimal region depends on the languages and writing style and genre. It would be interesting topic the variations on influence region for graph connectedness and it implication.

TABEL III THE NUMBER OF COMPONENTS BY STATEMENTS DISTANCE

Title	d=0	d=1	d=2	d=3	d=4	d=5
War and Peace	22	9	5	4	2	1
Harry Potter	74	20	4	4	2	1
The Earth	67	33	23	21	18	14
Three Kingdoms	92	21	7	5	2	1

C. Connected Component Analysis on Edge Weight

We are interested in the connectedness of whole social graph with respect to the edge weight. If we only consider the "strong" edge, then we would get the lots of disconnected component since some strong local group would be connected without any link edges connecting two different social groups. So by controlling the edge weight we can get the locally related group. Also we have to focus on the number of connected components with respect to the edge weight control.

According to our experiment, the number of components are increasing fast initially with lower value of cut weight c_*, which enables us to estimate the "the link characters" who connects two different local group. We think this property would be one important feature to distinguish the plot style of any writers.

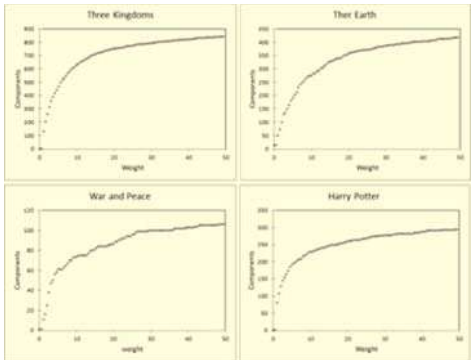

Figure 4. The number of components by edge weight

D. Variation on Maximum Component Size and Its implication

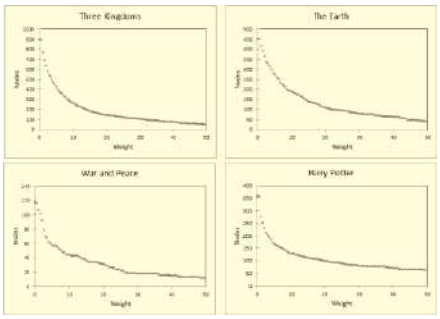

Figure 5. the number of node maximum component node according edge weight

We define the maximum component, the connected component whose number of nodes are maximal compared to other components. Generally we believe the maximal component would contain the main protagonists of the fiction. We tried to identify the maximal component of intermediate social network for a given cut value. We found that the maximal component of some fictions was not reduced with respect to the cut weight. That fact implies that the core groups of characters are highly related compared to the boundary persons along whole story. This variation of the size of maximal component would show the interesting characteristics of novel plot.

VI. CONCLUSION AND FUTURE WORK

A. Contribution Summary Experiment Summary

In this paper we considered how to extract the social network from literary text. We can conclude 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.

- Structural analysis on Social graph extracted from literary fiction gives interesting facts which help us to characterize the semantic structure of fictions. For example the variation on the maximal size of component would clearly shows the strength of core group of protagonists over whole story.
- The extracted literary social network shows quite similar structure in terms of lexical level to the narrative structure of fictions intrinsically.
- Our experiment clearly showed that this literary social network is one kind of complex system where most of properties are under power law.
- Also the number of connected components with respect to the cut value for edge weight implies that the social relationship among main characters is so higher than those of boundary persons. And the connected component will help us to find the local social group and the intermediate linking persons who connect the different local groups automatically.

B. 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 TextGraphs Workshop at ACL*, 2012, pp.4-12.
[2] 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.
[3] Y.-M. Choi. “Greek myth as a complex network,” *The Korean Physical Society*, vol. 49, no. 3 pp. 298-302, 2004.
[4] I. Dagan, L. Lee and P. C. Fernando, “Similarity-Based Models of Word Cooccurrence Probabilities,” *Machine Learning*, vol. 34 no.1-3, pp. 43-69, 1999.
[5] 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.
[6] C. L. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*, Empirical Press, 2004.
[7] 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.
[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] S.-R. Kim, “Complex network analysis in literature: Togi,” The *Korean Physical Society*, vol. 50, no. 4, pp. 267-271, 2004.
[10] R. Krestel and L. Chen, “Using co-occurence of tags and resources to identify spammers,” in *Proc. of Machine Learning and Principles and Practice of Knowledge Discovery*, 2008, pp. 38-45.
[11] 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.
[12] 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, 2004, pp. 1580-1585.
[13] 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
[14] R. Lambiotte., M. Ausloos and M. Thelwall. “Word statistics in Blogs and RSS feeds: Towards empirical

universal evidence," *Journal of Informetrics*, vol. 1, no. 4, pp.277-286, 2007.
[15] 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.
[16] J. Stiller, D. Nettle and R. I. Dunbar, "The small world of shakespeare's plays," *Human Nature*, vol. 14, no. 4 pp.397-408, 2003.
[17] Y. Fujii, T. Yoshimura and T. Ito. "Filtering harmful sentences based on three-word co-occurrence," in *Proc. of the 8th Annual Collaboration Electronic messaging Anti-Abuse and Spam Conference*, 2011, pp. 64-72.

Gyeong-Mi Park is at BK21 Center for U-Port IT Research and Education. She received the B.S. degree from 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

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.