International Journal of Advances in Computer Science & Its Applications
Volume 6 : Issue 1 [ISSN 2250-3765]
Publication Date : 18 April, 2016

A Semantic approach for Text Clustering using WordNet based on Multi-Objective Genetic Algorithms

[Jung Song Lee, Han Hee Hahm, Jong Joo Lee, Soon Cheol Park]

Abstract— **In this paper, we propose a method of Multi-Objective Genetic Algorithms (MOGAs), NSGA-II and SPEA2, for document clustering with semantic similarity measures based on WordNet. The MOGAs showed a high performance compared to other clustering algorithms. The main problem in the application of MOGAs for document clustering in the Vector Space Model (VSM) is that it ignores relationships between important terms or words. The hierarchical structure of WordNet as thesaurus-based ontology is an effective technique, which is used in semantic similarity measure. We tested these algorithms on Reuter-21578 collection data sets and compared them with Genetic Algorithms (GA) in conjunction with the semantic similarity measures based on WordNet. Also, we used F-measure to evaluate the performance of these clustering algorithms. The experimental results show that the performance of MOGAs based on WordNet is superior to those of the other clustering algorithms in the same similarity environments.**

Keywords— **Document Clustering, Multi-Objective Genetic Algorithm, Semantic Similarity Measure, WordNet**

I. Introduction

Clustering is an unsupervised classification technique that partitions the input space into *K* regions [1]. The document clustering classifies documents by grouping documents with similar features. The document clustering, which is one part of a clustering, is important in the text mining field [2], [3].

Currently, Genetic Algorithms (GA), which is one of the artificial intelligence algorithms, is widely used in document clustering. GA is a randomized search and optimization technique guided by the principles of evolution and natural genetics, which can be used in complex and large landscapes. It provides near optimal solutions [4].

Document clustering based on GA can provide appropriate cluster solutions using the searching capability of GA. The performance of the document clustering based on GA is better than other clustering algorithms [5]. However, it slows down the performance of clustering, so it is not used to prevent a premature convergence. To effectively avoid a premature convergence, Fuzzy Logic based on GA (FLGA), which exerts several control parameters to manipulate the crossover probability and the mutation probability of GA, has been proposed [6]. When the iterations of the best fitness without improvement reach consecutive maximum generations, the diversity of the population is extended by increasing crossover and mutation probability. Generally, it can effectively avoid trapping into a local optimum and also accelerate the evolving speed. However, it depends on several control parameters to manipulate the crossover probability and the mutation probability, such as parameter dependence. Recently, to solve these problems (premature convergence, parameter dependence), document clustering using Multi-Objective Genetic Algorithms (MOGAs) has been proposed [7], [8]. MOGAs define the document clustering problem as a Multi-Objective Optimization Problems (MOP) having two cluster validity indices. It uses two of MOGAs, NSGA-II [9] and SPEA2 [10] to solve MOP. Document clustering using MOGAs shows a higher performance than the other clustering algorithms (*k*-means, conventional GA). When these algorithms are applied in document clustering, most of them use the Vector Space Model (VSM) to represent documents. That is, each unique word in the vocabulary represents one dimension in vector space. However, document clustering has certain limitations when VSM is used, because VSM makes matches simply via keywords. Thus, the relationships between important words which do not co-occur are literally ignored in VSM [11]. In this paper we introduce MOGAs with semantic similarity measures for document clustering. We use the broad-coverage taxonomy and hierarchical structure of WordNet as a thesaurus-based ontology to detect semantic relationships between documents.

In the next section is a brief review of MOGAs. The details of MOGAs for document clustering with semantic similarity measures are described in section 3. Experimental results are given in section 4. Conclusions and future work are given in section 5.

II. Multi-Objective Genetic Algorithms

A. *Multi-Objective Optimization Problems*

In the optimization problems, when there are several objective functions these problems are called Multi-Objective Optimization Problems (MOP). MOP has many solutions that optimize one objective function but does not

Jung Song Lee, Soon Cheol Park
Division of Electronics and Information Engineering, Chonbuk National University
567 Baekje-daero, Deokjin-gu Jeonju-si, Jeollabuk-do, Republic of Korea

Han Hee Hahm
Department of Archeology and Cultural Anthropology, Chonbuk National University
567 Baekje-daero, Deokjin-gu Jeonju-si, Jeollabuk-do, Republic of Korea

Jong Joo Lee
Department of Korean Language and Literature, Chonbuk National University
567 Baekje-daero, Deokjin-gu Jeonju-si, Jeollabuk-do, Republic of Korea

optimize other objective functions (e.g. Conflict among objectives). Therefore, it is almost impossible to simultaneously optimize all objective functions [12]. The Pareto based method is often used to solve MOP with this character. This method finds a set of solutions by the dominance relation between candidate solutions. It is called Pareto Optimal Solution Set [13].

B. Multi-Objective Genetic Algorithms

Various algorithms have been suggested in order to solve the MOP. They are dependent on the initial search space and various solutions cannot be found. GA solves this disadvantage. GA simultaneously searches different regions of objective function space and makes it possible to find several members of the Pareto optimal solution set for difficult problems in a single run. The reproduction process enables the combination of existing solutions to generate new solutions [14]. GA for solving MOP is often called Multi-Objective Genetic Algorithms (MOGAs). Variations of MOGAs have been used in many applications and their performances were tested in several studies, i.e. PESA-II, SPEA2, NSGA-II, etc., representing leading research in this category. In these methods, NSGA-II and SPEA2 are easy to implement and do not have parameters for diversity in a population [15]. So, we applied these algorithms to document clustering.

III. MOGAs for Document Clustering with Semantic Similarity Measures

A. Document Representation

In most existing document clustering algorithms, documents are represented using the Vector Space Model (VSM). The document vector that represents the character of the document is formed by the weights of the terms indexed in a document [16]. We extracted the indexed terms by using stop words and Porter's stemming, and calculated the term weight by Okapi's calculation [17]. In VSM we use cosine measure to compute the similarity between two documents [18]. In order to perform document clustering, an important process is document representation. The general approach uses VSM to represent documents. VSM has many drawbacks. Because each unique term in the vocabulary represents one dimension in feature space, VSM needs a large number of features to represent high dimensions, and it can easily cause the overflow problem. Significantly, if we represent all documents by this method, it ignores relationships between important terms. In the natural language of a document, if VSM is used with words with the same concept, the meaning of these words becomes ambiguous. So, the document clustering algorithms are needed for an appropriate similarity measure.

B. Semantic Similarity Measures based on WordNet

The semantic similarity measure, which is an important issue in the field of text-oriented research, such as Natural Language Processing and Data Mining, is a proximity measure between words. The similarity between two words is often represented by the similarity between the concepts related with the two words. The version used in this paper is the WordNet 2.0, which has 144684 words and 109377 synonym sets. WordNet is an online software package developed at Princeton by a group led by Miller that can make it possible to measure the semantic similarity and relatedness between a pair of concepts [19]. It organizes the lexicon by nouns, verbs, adjectives, and adverbs, named synsets. Synsets represent terms or concepts with similar meaning and are organized into senses, which are different meanings of the same term [20]. The semantic similarity of two concepts c_1 and c_2 can be calculated from the tree-like hierarchical structure of WordNet. Generally, it used to find the shortest path connecting these two concepts. Figure 1 shows a part of such a hierarchical semantic knowledge base.

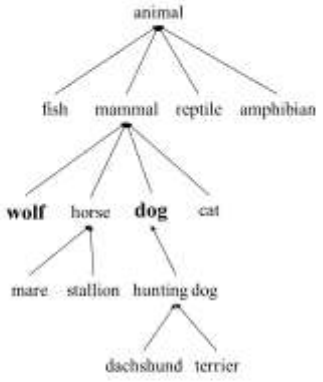

Figure 1. A part of hierarchical semantic knowledge base.

The shortest path between ‘‘wolf” and ‘‘dog” is ‘‘wolf-mammal-dog”. The minimal length of the path is 2. The synset ‘‘mammal” is called the subsumer for concepts ‘‘wolf” and ‘‘dog”.

In this paper we used the semantic similarity measure, which is based on the concepts inter-connected hyponymy (‘‘Is-A”) hierarchy given by Li et al., [21]. We used two strategies to calculate semantic similarity of two concepts, c_1 and c_2. These strategies use the shortest path length between two concepts and depth of the subsumer in the hierarchy. Strategy 1 is denoted by: $\boldsymbol{sim(c_1,c_2)=f_1(l)}$ and Strategy 2 is denoted by: $\boldsymbol{sim(c_1,c_2)=f_1(l)\bullet f_2(h)}$, where l is the shortest path between concept c_1 and c_2. h is the depth of subsumer in the hierarchy. Here, it is assumed that the impacts of parameters l and h on the similarity are independent of each other. Thus, f_1 and f_2 are independent functions. f_1 and f_2 defined by: $\boldsymbol{f_1(l) = e^{-\alpha l}}$ and $\boldsymbol{f_2(h) = (e^{\beta h}-e^{-\beta h})/(e^{\beta h}+e^{-\beta h})}$, where α is a real constant between 0 and 1 and $\beta > 0$ is a smoothing factor. The semantic similarity between these two words (w_1, w_2) is measured by: $\boldsymbol{sim(w_1,w_2)} = \mathbf{max\{sim(c_1,c_2)\}}$, where, w_1 is represented by a number of a concepts ($c_{1,1}$, $c_{1,2}$, ..., $c_{1,a}$) and w_2 is represented by a number of b concepts ($c_{2,1}$, $c_{2,2}$, ..., $c_{2,b}$). The semantic similarity between these two documents is then defined by:

$$sim(d_1, d_2) = (\sum_{i=1}^{m}\sum_{j=1}^{n} sim(w_{1,i}, w_{2,j}))/mn, \qquad (1)$$

where, m and n are the number of WordNet lexicon words included in documents d_1 and d_2 respectively. Unfortunately, some important words which transform to incomplete forms after stemming in a special document are not found in

WordNet and will not be considered as concepts for similarity evaluation.

C. Document Clustering based on MOP

The document clustering using GA with a single objective function can regarded as an optimization problem to optimize the cluster validity index. This view offers a chance to apply MOP on the clustering problem. Therefore, the document clustering was thought of as MOP optimizing two cluster validity indices through this trade off relation as: **arg max** ($\boldsymbol{F_{CH}}(\boldsymbol{C_i}) \wedge \boldsymbol{F_{DB}}(\boldsymbol{C_i})$), where *CH* and *DB* are represented as CH (Calinski and Harabasz) index [22] and DB (Davis and Bouldin) index [23] for the objective functions. C_i is a chromosome and C_i={CN_1, CN_2, … , CN_j, … , CN_m}. CN_j is the cluster number assigned to a document and 1<= CN_j <= *K*. *n* is the number of chromosomes in a population, *m* is the number of documents and *K* is the number of clusters.

D. Chromosome Encoding and Evolution Principles

The chromosome is encoded by a string of integers. Each chromosome in the population is initially encoded by a number of *m* genes with an integer value in the range 1~*K*. Where, *m* is the number of documents and *K* is the number of clusters. Thus, each gene represents a document, and the value of a gene represents a cluster number. MOGAs using the cluster validity indices as the objective functions require higher computational complexity. Therefore, we adopted the simple cluster validity indices CH index and DB index for document clustering using MOGAs. Multi-point crossover and uniform mutation were adopted in the evolution operators [24].

E. Document Clustering using NSGA-II and SPEA2

By applying the sharing technique, NSGA maintained the diversity of the Pareto Optimal Solution Set. The loss of the optimal solutions was caused in the evolutionary process because it has increased computational complexity and has not applied an elite preserve strategy. For this, NSGA-II, adding *Fast Non-dominated Sort* and *Crowding Distance Assignment Operation*, was proposed. Under the elite preserve strategy, SPEA stores the Pareto Optimal Solution Set separately. However, the diversity cannot be maintained because of the fitness an an assignment problem. So, an improved version of the SPEA, namely SPEA2, is proposed, which is the new *Fitness Assignment* and *Archive Truncation Method*. In MOGA, the Pareto Optimal Solution Set contains a large number of solutions. That is, document clustering using the MOGAs does not return a single cluster solution. The identification of promising solutions from the Pareto Optimal Solution Set has been investigated in several papers, named Decision Maker (DM) [25]. But, these techniques are very difficult. So, we manually selected one of the best cluster solutions in the Pareto Optimal Solution Set.

IV. Experiment Results

In this section, we implement our method of MOGAs, NSGA-II and SPEA2, for document clustering on the Reuters-21578 collection, which is one of the most widely adopted benchmark datasets in the text mining field, and compare and discuss the performance of MOGAs with conventional GA. Also, we used F-measure [26] to evaluate the performance of these clustering algorithms. The population number in our conventional GAs and MOGAs is 300. These algorithms are terminated when the number of generations reaches 1000 or when the iterations without improvement reach consecutive 20. In the current test data set 1, containing 200 documents from four topics (earn 50, gnp 50, cocoa 50, gas 50), data set 2, containing 200 documents from four topics (coffee 50, trade 50, crude 50, sugar 50) and data set 3, containing 300 documents from six topics (coffee 50, trade 50, crude 50, sugar 50, grain 50, ship 50) are selected. After being processed by word extraction, stop word removal, and Porter's stemming, there are 2654, 3436 and 4210 index terms, respectively. Also, the index term weight that is extracted by using the Okapi's calculation was determined. We implemented our experiments in two steps: First, the performance of the different document similarity measurements was compared. Second, the performance of our MOGAs for document clustering was compared with conventional GA in the same similarity measurement environments. The performances of the document similarity measurements were calculated by sim_{VSM}, S_1 (Strategy 1) and S_2 (Strategy 2) were compared in this study. The basic parameters are set as: α =0.25 in Strategy 1 and α =0.2, β =0.6 in Strategy 2. The performance of the proposed clustering algorithm is then illustrated with the different similarity measures (Table 1) for each data set.

TABLE I. PERFORMANCE OF THE PROPOSED MOGAS WITH THE DIFFERENT SIMILARITY MEASURES ON EACH DATA SET.

Algorithms	F-measure		
	Data set 1	*Data set 2*	*Data set 3*
NSGA-II (DB, CH) - sim_{VSM}	0.92	0.68	0.68
SPEA2 (DB, CH) - sim_{VSM}	0.90	0.63	0.63
NSGA-II (DB, CH) - S_1	0.94	0.74	0.75
SPEA2 (DB, CH) - S_1	0.93	0.72	0.74
NSGA-II (DB, CH) - S_2	0.96	0.80	0.79
SPEA2 (DB, CH) - S_2	0.94	0.78	0.77

From Table 1 we can see that with S_2, our MOGAs almost get the best F-measure on all data sets. Consequently, document clustering using MOGAs based on semantic similarity measures with S_1 and S_2 shows a 6% and 10% better performance, respectively, than with VSM.

We also compared MOGAs with conventional GA in the same semantic similarity environment. The comparison results are shown in Table 2.

TABLE II. PERFORMANCE OF THE PROPOSED MOGAS IN COMPARISON WITH GA ON EACH DATA SET.

Algorithms	F-measure		
	Data set 1	*Data set 2*	*Data set 3*
NSGA-II (DB, CH) - S_1	0.94	0.74	0.75
SPEA2 (DB, CH) - S_1	0.92	0.72	0.74
NSGA-II (DB, CH) - S_2	0.96	0.80	0.80
SPEA2 (DB, CH) - S_2	0.94	0.78	0.77
GA (DB) - S_1	0.70	0.58	0.52
GA (CH) - S_1	0.78	0.60	0.58
GA (DB) - S_2	0.76	0.62	0.57
GA (CH) - S_2	0.82	0.66	0.63

In summary, we can expect performance improvements of document clustering by using MOGAs with semantic similarity measures based on WordNet. Consequently, document clustering using MOGAs based on semantic similarity measures with S_1 and S_2 shows the performance about 20% and 19% better than GA(DB), respectively. Also, S_1 and S_2 demonstrate a respective performance of about 15% and 14% better than GA(CH).

Figure 2 shows the computational times until generation of each data set of MOGAs and GAs is 1000 minutes. We can see from Figure 2 that the computational time of MOGAs is greater than that for GAs.

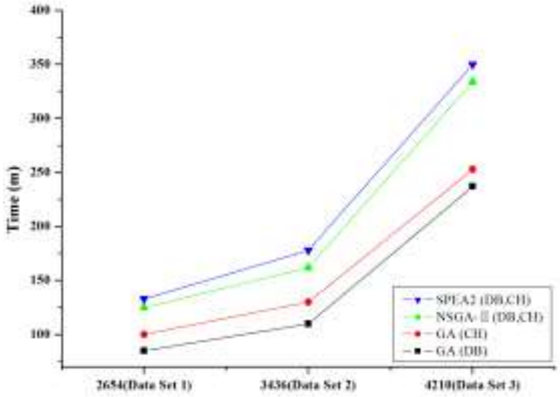

Figure 2. The computational times of MOGAs and GAs.

We consider the F-measure of these clustering algorithms, which MOGAs and GAs based on semantic similarity measures with S_1 and S_2, for the case of the increase of computational time until generation is 1,000. In order to make a fair comparison, we create the same initial population for MOGAs and GAs first. The CH index was selected because it obtained the best performance in GAs. Figures 3, 4 and 5 illustrate the results of the F-measure as an increase in computational time for each data set.

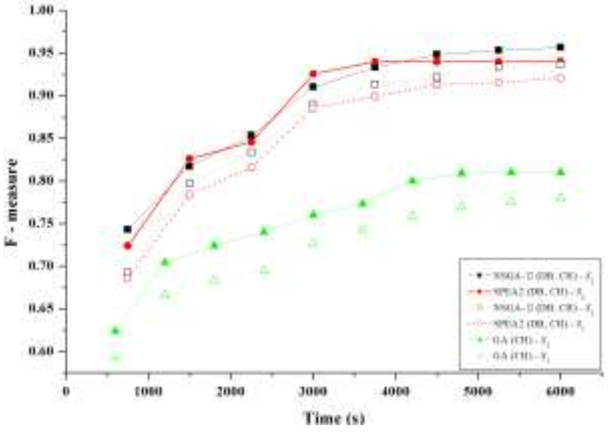

Figure 3. The best F-measure against computational times on data set 1.

We can see from Figure 3 that the F-measure increases with the raising of computational time. Although GAs converges rapidly for data set 1, the F-measure increases slowly which trap into a local optimum. The F-measure of MOGAs with S_2 increases rapidly and has better results than that of the GAs for data set 1.

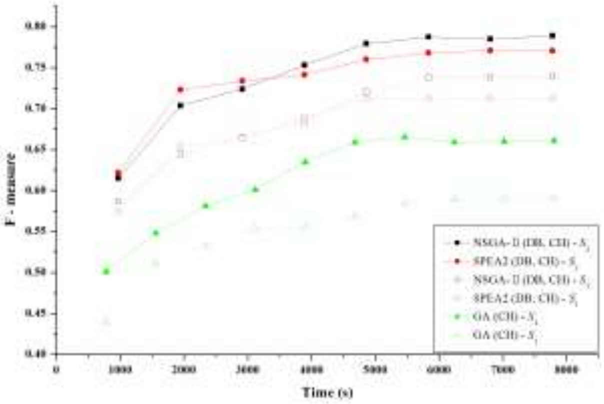

Figure 4. The best F-measure against computational times on data set 2.

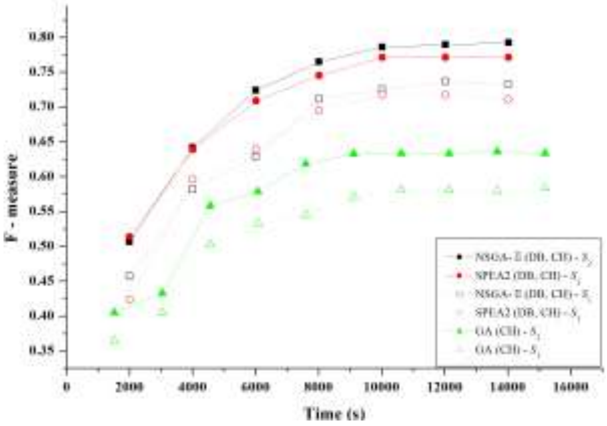

Figure 5. The best F-measure against computational times on data set 3.

In Figures 4 and 5, the GAs and the MOGAs are converged at almost the same computational time (about 5000, 12000 seconds respectively), but the F-measures of the MOGAs with S_2 are much better than that of the MOGAs with S_1 and GAs with S_1 and S_2.

This implies that the MOGAs with S_2 are more effective than the MOGAs with S_1 and GAs with S_1 and S_2 at finding better clustering solutions. For each data set, the average F-measure values of the MOGAs, NSGA-II and SPEA2 with S_2, are 0.95, 0.79 and 0.785 in each data set respectively.

V. Conclusions and Future Works

In this study, MOGAs, two of MOGA's algorithms, NSGA-II and SPEA2, with two of the semantic similarity measures are proposed for document clustering. The main problem in the field of document clustering is that the document is represented as a bag of words, while the conceptual similarity between each pair of documents is ignored. So, we applied WordNet, which is one of the thesaurus-based ontologies for document similarity measure of document clustering to the Reuters-21578 documents collection to demonstrate the effectiveness of our clustering algorithm, which we have demonstrated to be superior to other clustering algorithms. In our two experiments, we compared the performance of the different document similarity measurements and compared them with conventional GA in the same similarity measurement environments. WordNet improved the clustering performance of MOGAs. The results show that our MOGAs, in conjunction with the Strategy 2, use the shortest path length between two concepts and depth of the subsumer in the hierarchy and get the best clustering. Also, Strategy 2 was more suited for resolving ambiguity of words in the corpus than Strategy 1. In the future, we will apply the matrix factorization technique, which consists of singular value decomposition (SVD), non-negative matrix factorization (NMF), and parallel processing, to the algorithms. In addition, we will adopt Decision Maker (DM) to select one of the best cluster solutions in the Pareto optimal solution set.

Acknowledgment

This work was supported by the National Research Foundation of Korea Grant funded (NRF) by the Korean Government (NRF-2013S1A5B8A01072201 & NRF-2013R1A1A2063572) and the Brain Korea 21 PLUS Project, National Research Foundation of Korea.

References

[1] M. Charikar, V. Guruswami and A. Wirtha, “Clustering with qualitative information,” Journal of computer and system science, vol. 71, pp. 360-383, October, 2005.

[2] W. B. Croft, D. Metzler and T. Strohman, Search Engines Information Retrieval in Practice, Addison Wesley, 2009.

[3] P. Pantel and D. Lin, “Document clustering with committees,” 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Finland, 2002.

[4] U. Maulik and S. Bandyopadhyay, “Genetic algorithm-based clustering technique,” Pattern Recognition, vol. 33, no. 9, pp. 1455-1465, 2000.

[5] W. Song and S. C. Park, “Genetic algorithm for text clustering based on latent semantic indexing,” Computers and Mathematics with Applications, vol. 57, pp. 1901-1907, 2009.

[6] W. Song and S. C. Park, “Latent semantic analysis for vector space expansion and fuzzy logic-based genetic clustering,” Knowledge and Information Systems, vol. 22, pp. 347-369, 2010.

[7] J. S. Lee, L. C. Choi and S. C. Park, “Document clustering using multi-objective genetic algorithm with different feature selection methods,” 1st International Workshop on Semantic Interoperability, 2011.

[8] J. S. Lee, L. C. Choi and S. C. Park, “Multi-objective genetic algorithms, NSGA-II and SPEA2, for document clustering,” Communications in Computer and Information Science, vol. 257, pp. 219-227, 2011.

[9] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, “A fast elitist multiobjective genetic algorithm: NSGA- II,” IEEE transaction on evolutionary computation, vol. 6, no. 2, pp. 182-197, 2002.

[10] E. Zitzler, M. Laumanns and L. Thiele, “SPEA2: Improving the strength pareto evolutionary algorithm,” Proceedings of the EROGEN, 2002.

[11] S. Zelikovitz and H. Hirsh, “Using LSI for text classification in the presence of background text,” Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 113-118, 2001.

[12] J. L. Cohon and D. H. Marks, “A review and evaluation of multiobjective programming techniques,” Water Resources Research, vol. 11, no. 2, pp. 208-220, 1975.

[13] Y. Censor, “Pareto optimality in multiobjective problems,” Appl. Math. Optimiz, vol. 4, pp. 41-59, 1977.

[14] K Deb, Multi-Objective using Evolutionary Algorithms, John Wiley & Sons, England, 2001.

[15] K. Abdullah, W. C. David and E. S. Alice, “Multi-objective optimization using genetic algorithms : A tutorial. Reliability Engineering and System Safety,” vol. 91, pp. 992-1007, 2006.

[16] L.C. Choi, K.U. Choi, and S.C. Park, “An automatic semantic term-network construction system,” International Symposium on computer Science and its Applications, pp. 48-51, 2008.

[17] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

[18] Haoxiang Xia, Shuguang Wang, and Taketoshi Yoshida, “A modified ant-based text clustering algorithm with semantic similarity measure,” Journal of Systems Science and Systems Engineering, vol. 15, no. 4, pp. 474-492, 2006.

[19] G. .A. Miller, “WordNet: A lexical database for English,” Comn. ACM, vol. 38, no. 11, pp. 39-41, 1995.

[20] A. Hotho, S. Staab, and G. Stumme, “Wordnet improves text document clustering,” Proc of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference, 2003.

[21] Yuhua Li, Zuhai A. Bandar, and David Mclean, “An approach for measuring semantic similarity between words using multiple information sources,” IEEE Trans on Knowledge and Data Engineering, vol. 15, no. 4, 2003.

[22] T. Calinski and J. Harabasz, “A Dendrite Method for Cluster Analysis,” Communications in Statistics, vol. 3, no. 1, 1974.

[23] D. L. Davies, and D. W. Bouldin, “A Cluster Separation measure,” IEEE transactions on Pattern analysis and Machine Intelligence, vol. 1, no. 2, 1979.

[24] L. D. Davis, Handbook of Genetic Algorithms. Van Nostrand Reinhold, 1991.

[25] X. Blasco, J.M. Herrero, J. Sanchis and M. Martínez, “A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization,” Information Sciences, vol. 178, pp. 3908-3924, 2008.

[26] D. Fragoudis, D. Meretakis and S. Likothanassis, “Best terms:an efficient feature-selection algorithm for text categorization,” Knowl Inform Syst, vol. 8, pp. 16-33, 2005.

About Author (s):

Jung Song Lee received his B.S. degree in the Division of Electronics and Information Engineering, Jeonbuk National University, Jeonju, Jeonbuk, Korea, in 2011. He is currently a M.S. candidate in the Division of Computer Science and Engineering at Chonbuk National University. His research interests include Pattern Recognition, Information Retrieval, Evolutionary Computing, Artificial Intelligence, Data Mining and Knowledge Discovery.

Soon Cheol Park received his B.S. degree in Applied Physics from Inha University, Incheon, Korea, in 1979. He received his Ph.D. in Computer Science from Louisiana State University, Baton Rouge, Louisiana, USA, in 1991. He was a senior researcher in the Division of Computer Research at Electronics & Telecommunications Research Institute, Korea, from 1991 to 1993. He is currently a Professor in the Division of Electronics and Information Engineering, Jeonbuk