The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 17, No. 5, pp.131-136, Oct. 31, 2017. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2017.17.5.131
JIIBC 2017-5-18

유전자 알고리즘 기반 용어 중의성 분석

Analysis of Term Ambiguity based on Genetic Algorithm

김정준*, 정성택**, 박정민***

Jeong-Joon Kim*, Sung-Taek Chung, Jeong-Min Park*****

요 약 최근 인터넷 미디어의 발달로 웹상에 수많은 문서자료들이 기하급수적으로 늘어나게 되었다. 이러한 자료들은 대부분 텍스트에 의해 그 내용이 무엇인지를 설명하고 있고 이에 따라 분류된다. 그러나 텍스트가 가지는 의미는 모호하게 해석되어질 여지가 많고 이를 정확히 해석하기 위해서는 다각도로 이를 살펴봐야 한다. 기존의 분류 방법에서는 단순히 텍스트의 출현만을 가지고 분류를 하였다. 따라서, 본 논문에서는 이를 유전자 알고리즘과 토픽추출을 기반으로 하여 용어 중의성을 분석하고 이를 단편화한 클러스터링 시스템을 구현하였다. 마지막으로 구현된 결과물을 토대로 기존의 방법과 비교하여 본 논문의 성능을 평가하였다.

Abstract Recently, with the development of Internet media, many document materials have become exponentially increasing on the web. These materials are described, and the information on what is the most by this text are classified according. However, the text has meant that many have room for ambiguous interpretation must look at it from various angles in order to interpret them correctly. In conventional classification methods it was simply a classification only have the appearance of the text. In this paper, we analyze it in terms genetic algorithm and local preserving based techniques and implemented a clustering system fragmentation them. Finally, the performance of this paper was evaluated based on the implementation results compared to traditional methods.

Key Words : Term Ambiguity, Genetic Algorithm, Documnet Classify, Documnet Clustering

Ⅰ. 서 론

최근에 인터넷미디어의 발전으로 다양한 매체들이 인터넷 상에 널리 퍼지게 되었으며 이러한 다양한 미디어 정보는 텍스트를 기반으로 정보의 내용을 표시하고 있으며, 이를 하나의 문서로 볼 수 있다. 이러한 다양한 문서를 분류 관리하기 위한 연구가 지속적으로 진행 중이다.[1,8,9]. 대표적인 문서 분류 방법으로 용어간의 상호 연관성을 측정하여 카테고리를 구성하고 이를 기반으로 분류하는 잠재 의미 분석(Latent Semantic Analysis) 기법이 대표적이다. 그러나 잠재 의미 분석 기법은 용어간의 상호연관성을 측정하기에 적합하지만 용어의 의미 내용을 파악하기는 힘들다. 즉, 용어에 의한 분류는 가능하지만 다의어에 대해서 판단해서 분류할 수가 없는 한계를 가진다[2,5,6]. 용어의 다중 의미를 파악하기 위해서는 용어를 단편화해서 생각할 필요성이 있다.

따라서, 본 논문에서는 기존의 잠재 의미 분석 기법에 용어의 중의성을 해석하기 위해 유전자 알고리즘의 방법

*정회원, 한국산업기술대학교 컴퓨터공학과
**정회원, 한국산업기술대학교 컴퓨터공학과
***정회원, 한국산업기술대학교 컴퓨터공학과
접수일자: 2017년 8월 11일, 수정완료: 2017년 9월 11일
게재확정일자: 2017년 10월 13일

Received: 11 August, 2017 / Revised: 11 September, 2017 / Accepted: 13 October, 2017
**Corresponding Author: unitaek@kpu.ac.kr
Dept. of Computer Eng., Korea Polytechnic University, Korea

론과 벡터 공간 모델을 응용하여 새로운 분류기법을 제안하고자 한다.

II. 관련 연구

1. 유전자 알고리즘

유전자 알고리즘은 전역 최적화 기법으로 자연세계의 진화과정에 기초한 계산 모델로써, 병렬적이고 전역적인 탐색 알고리즘을 다윈의 전자생존 이론에 맞추었다. 일정한 자료구조 안에 가능한 결과를 생성, 혼합, 변이 시켜 보다 좋은 해를 생성하는 기법이다[7].

유전자 알고리즘은 Selection, Cross Over, Mutation, Substitution의 4가지 주요 연산을 통해 최적해에 가까운 답을 찾아가는 방법으로 일정한 자료구조에 대하여 각 연산을 한 결과를 평가하고 보다 적합한 해를 남겨가며 이를 선택한다.

Selection 연산은 해의 후보가 되는 해들을 선택하는 연산이고, Cross Over 연산은 교배를 통해 새로운 해를 생성한다. Mutation 연산은 순서를 변형하거나 인자를 섞는 등의 변이 과정을 통해 적합한 해를 만들어내며, Substitution 연산은 새로운 해를 해집단에 추가하고, 열등한 해를 제외시키는 연산이다.

2. 잠재 의미 색인 기법

잠재 의미 색인 기법은 SVD(Singular Value Decomposition)이라는 선형 대수학 기법을 이용해 원본 행렬을 3개의 행렬로 분해하여 이를 이용하여 각각 행과 열의 상호관계를 판별하는 기법이다[3].

$$M = U\sum V^t \qquad (1)$$

식 1 에서 보는 바와 같이 원본 행렬 M은 U, $\sum$, V^t 3가지 행렬로 분해되며, 각각 행의 고유값, 차원의 고유값, 열의 고유값이라는 의미를 가지며, U*$\sum$ 행렬곱은 행의 상호연관성을 차원에 나타낼 수 있으며, $\sum * V^t$ 행렬곱은 열의 상호연관성을 차원에 나타내게 된다.

문서 분류 시스템에서는 TF-IDF(Term Frequency - Inverse Document Frequency)등을 행렬로 표현한 후 이를 잠재 의미 색인 기법을 통해 차원상에 문서와 용어의 위상을 표현함으로 분류해낼 수 있게 된다.

그러나 잠재 의미 색인 모델은 용어의 위상을 표현하기 좋으나 용어의 중의성을 파악할 수는 없는 문제점을 가지고 있다.

2. 벡터 공간 모델

벡터 공간 모델은 데이터를 벡터로 표현하는 대수적 모델로써, 특히 정보 검색에 쓰이고 있다[4].

만약 어떤 용어가 문서에 포함되면, 해당 단어는 0이 아닌 벡터 값을 가지게 되며, 이를 벡터 공간에 표현함으로 문서간의 유사성을 찾는 방법이다.

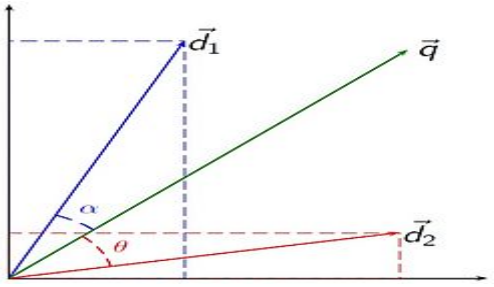

그림 1. 벡터 공간 모델
Fig. 1. Vector Space Model

그림 1 에서 보는 바와 같이 질의 q가 주어졌을 때, 문서 d1, d2를 벡터상에 표현하고 이를 통해 유사도 높은 문서를 찾음으로 문제를 해결한다.

그러나, 벡터 공간 모델은 길이가 긴 문서나 유사한 내용을 판별할 수 있는 방법이 없다는 단점을 가지고 있으며, 용어가 서로 독립적이라는 가정을 바탕으로 하기 때문에 현재에는 잘 쓰여지지 않는 기법이다.

III. GATA2

Genetic Algorithm based Ambiguity Analysis (GATA2)는 잠재 의미 색인의 단점인 중의성 해석을 위한 방법론으로 두 가지 기법을 활용한다. 첫 번째 유전자 알고리즘이고, 두 번째 벡터 공간 모델이다. 유전자 알고리즘으로 가능한 해에 대한 가능성을 넓히고, 해에 대한 검증을 통해 보다 좋은 성능을 입증하고자 하며, 유전자 알고리즘을 그대로 사용하는 것이 아닌 미미틱 기법과 담금질 기법을 이용한다. 미미틱 기법은 유전자 알고리즘의 전역탐색을 지역탐색으로 변형한 기법이며, 담금질 기법은 해집합을 사용하지 않고 하나의 해만을 사용한다

는 특징을 가지고 있다. 벡터 공간 모델을 이용하여 중의성 파악을 하고자 한다. 벡터 공간 모델은 잠재 의미 색인 자체가 벡터 공간에 위상을 가지는 시스템이므로 이를 그대로 적용하는 것이 아니라 중의성 파악을 위해 유전자 알고리즘의 Selection, Mutation 연산에 적용되어진다.

1. 유전자 알고리즘 설계

최초의 주어진 해는 잠재 의미 색인의 결과이며 이를 이용하여 유전자 알고리즘의 연산과 검증에 이용한다. 담금질 기법을 사용함으로 해들 간의 조합이 없음으로 Cross Over와 Substitution 연산은 없다.

Selection 연산은 현재 해에서 스칼라 값이 작은 용어를 제거한다. 이에 대한 효과는 자질 분류 효과 및 노이즈 제거 효과로 볼 수 있다.

Cross Over연산은 다음과 그림 2 과 같이 설명되어진다.

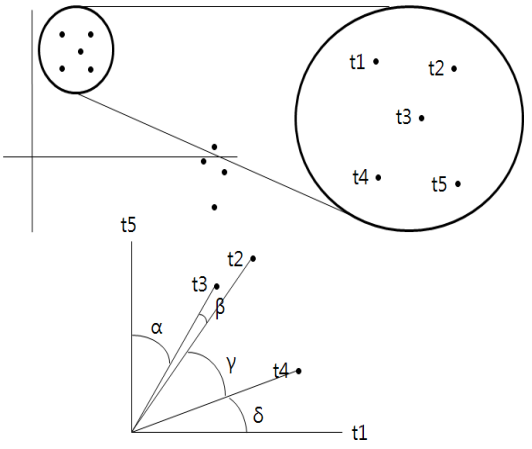

그림 2. Cross Over 연산
Fig. 2. Operation Cross Over

그림 2 에서 보는 바와 같이 잠재 의미 색인으로 분류되어진 클러스터에서 유사도가 차가 큰 용어를 선별하여 이를 벡터 공간 모델에 적용한다. 그림 2 와 같이 t1, t5를 기준으로 하였을 때, t2, t3, t4의 위상은 각각 t1, t5와의 코사인 유사도를 기준으로 판별한다. 이 때, 각각의 유사도를 판별 할 수 있게 되며, t2, t3는 상당히 작은 유사도로 식별되어 질 수 있다. 이러한 경우 t2, t3를 t'1로 동일한 의미로 간주하여 행렬을 재구성한다.

유전자 알고리즘은 원본행렬이 더 이상 변하지 않을 때까지 수행하여야 하지만 대규모 규모에 따라 1회 수행에 대한 연산량이 매우 많다. 따라서 많은 횟수를 수행하기에 무리가 있음으로 20회 수행하였다.

2. 벡터 지역 보호

잠재 의미 색인을 이용하여 클러스터를 구하더라도 용어의 의미를 파악할 수 는 없지만 각 용어가 가지는 의미벡터를 알 수 있다. 이 경우 동의어와 유의어가 겹쳐지는 경우라면 성능에 영향을 미치지 않고 오히려 좀 더 정확한 클러스터 결과가 나올 수 있다. 만약 동음이의어와 같이 복수의 의미를 가지는 경우의 분석에 대하여 클러스터별로 접근함으로써 지역 탐색을 통해 이 부분의 문제를 최소화 한다.

지역 탐색을 통한 방법은 아래와 같이 설명되어진다.

	d1	d2	d3	d4	d5	d6	D7
t1	3	3	4	7	-	-	-
t2	2	5	3	6	-	-	-
t3	1	-	-	2	4	3	6
t4	-	-	-	-	3	5	6

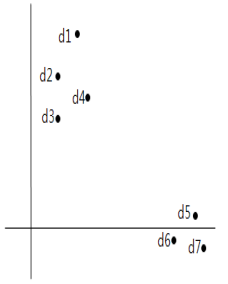

그림 3. 지역분석 예제 1
Fig. 3. Local Analysis Example1

그림 3 의 표와 같이 문서-용어 행렬이 있다고 가정하고 이 행렬이 오른쪽 그림과 같이 벡터 공간에 분포되어진다고 하였을 때, d1~d4의 클러스터와 d5~d7의 클러스터 2개의 행렬로 분해되어진다고 볼 수 있다. 이 때, 용어 t3은 두 개의 클러스터에 모두 등장한다. 만약 t3과 t4가 그림 2 의 β와 같이 축약되어진다고 가정하면 t3의 변경에 따라 d1~d4 클러스터도 영향을 받게 된다. 즉, 관련성이 적은 클러스터도 영향을 받는다고 볼 수 있다. LSI연산에 당연히 이러한 작은 변화도 영향을 받을 수 밖에 없다. 그러나 이를 최소화 하는 방법으로 지역적으로 변환하는 방법을 제안한다.

표 1. 지역분석 예제 2
Table 1. Local Analysis Example2

	d1	d2	d3	d4	d5	d7	D7
t1	3	3	4	7	-	-	-
t2	2	5	3	6	-	-	-
t3	1	-	-	2	-	-	-
t`1	-	-	-	-			

표 1 에서 보는 바와 같이 d5~d7 클러스터 내에 속해 있는 부분의 t3과 t4는 t`1로 변경되어 지고 d1~d4 클러스터내의 t3은 그대로 유지되어진다. 이는 d1~d4 클러스터에서 t3의 비중이 크다 작다를 확실히 분별할 수 없고 다른 클러스터와의 관계에 영향을 미칠지를 판별할 수가 없기에 이러한 방법을 사용하였으며, 만약 다른 클러스터와의 상호연관성에 영향을 주지 않으며, 낮은 비중을 차지하게 된다면 t1은 Selection 과정에서 삭제되게 되므로 이러한 방법을 사용하였다.

3. 시스템 설계도

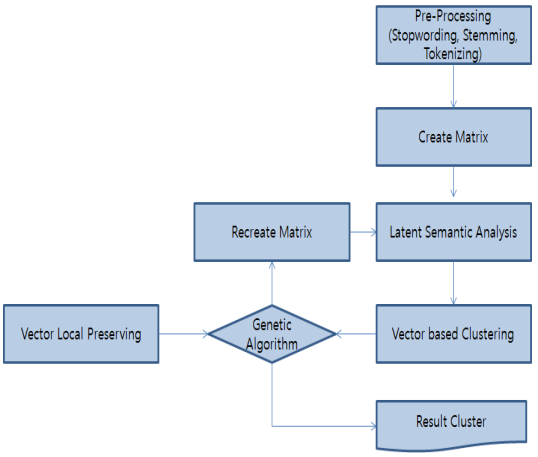

그림 4. 시스템 흐름도
Fig. 4. System Flow Chart

그림 4 에서 보는 바와 같이 본 논문의 GATA2는 먼저 전처리 과정으로 불용어 처리, 어근 추출, 토큰 추출 등을 한 후 이를 행렬화 한다. 만들어진 행렬을 잠재 의미 색인 기법으로 벡터 공간상에 표현한 후 이를 벡터 기반 클러스터링을 한다. 이 후 벡터 공간 보호를 기반으로 하는 유전자 알고리즘을 이용하여 결과를 검증하게 되는데 검증 결과를 만족하지 못하면 유전자 알고리즘으로 행렬을 변형하여 새로운 행렬을 가지고 벡터 공간에 잠재 의미 색인 기법으로 표현하고, 만족한 경우 클러스터링을 수행한다.

Ⅳ. 실험 및 결과

본 논문에서 개발한 GATA2의 성능을 검증하기 위해서 NewsGroup 20의 문서를 이용하였다. 위에서도 기술하였듯이 많은 데이터를 처리하기 위해서 처리해야 할 데이터량이 많음으로 인해 완전한 성능이 나올 수 있는 환경이 아니므로 유전자 알고리즘의 반복횟수는 20회로 제한하였으며, 검증에 사용된 평가 수치로 Precision,m, Recall, Accuracy, F-1 Measure가 사용되었다. Precision은 실제 해에 얼마나 결과가 정확하게 나왔음을 판별하는 정밀율이고, Recall은 실험 결과 도출되어진 결과의 재현율이다. Accuracy는 전체 결과가 얼마나 정확하게 나오는지를 판별하는 정확율로 사용되고, F-1 Measure는 Precision과 Recall의 조화 평균으로 사용된다.

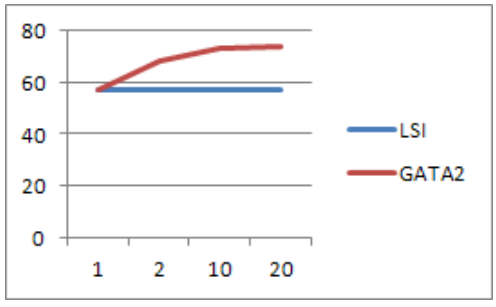

그림 5. Precision
Fig. 5. Precision

그림 5 에서 보는 바와 같이 Precision은 약 17%의 성능향상을 보였다.

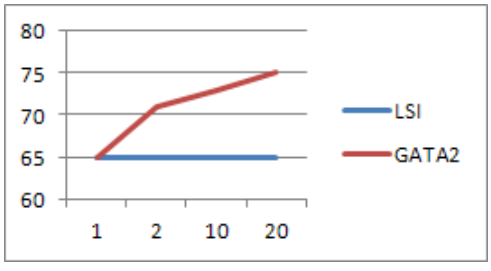

그림 6. Recall
Fig. 6. Recall

그림 6 에서 보는 바와 같이 Recall은 약 10%의 성능 향상을 보였다.

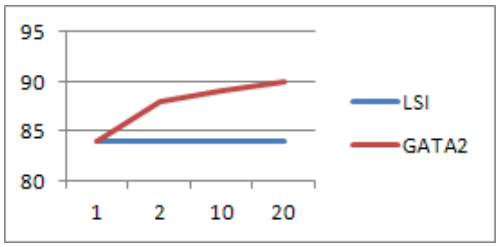

그림 7. Accuracy
Fig. 7. Accuracy

그림 7 에서 보는 바와 같이 Accuracy는 약 6%의 성능향상을 보였다.

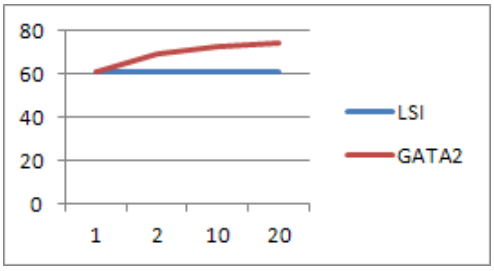

그림 8. F-1 Measure
Fig. 8. F-1 Measure

그림 8 에서 보는 바와 같이 Precision은 약 17%의 성능향상을 보였다.

위 성능 평가 항목의 결과들은 횟수가 반복됨에 따라 둔화되는 결과를 가져오는데 이는 답에 근접해 갈수록 성능저하가 나타남을 알 수가 있고, 따라서 일정 횟수 이상을 더 수행하더라도 보다 좋아지는 결과를 얻을 수 없음을 유추할 수 있다. 결과적으로 적정한 횟수를 산정함이 효율적인 결론을 도출 할 수 있다는 것을 알 수 있으나 본 실험의 20회의 결론으로는 효율적인 구간을 찾지는 못하였다.

V. 결 론

최근 다양한 인터넷 미디어로부터 수많은 정보가 생성되었고, 이를 관리하기 위한 기술의 필요성이 증대되었다. 이러한 연구는 정보 검색 시스템의 등장과 함께 지속적으로 연구되어 왔다.

문서 데이터를 분류하기 위한 연구는 여러 가지 방향으로 지속되어 연구되어 왔지만 용어의 의미를 파악하는 문제는 난제로 남았다. 즉, 다의어, 동음이의어 처리에 있어서 어려움이 있었고 이로 인하여 정확하지 못한 결과를 보이는 경우가 많았다.

따라서, 이러한 문제를 해결하기 위해서 본 논문에서는 유전자 알고리즘과 벡터 지역 보호를 통해 이를 시도하였다. 본 시스템은 유전자 알고리즘을 통해 검증과 개선이라는 방법론을 취하며, 자동적인 노이즈 필터링과 자질 선택을 할 수 있도록 하였다. 또한, 벡터 지역 보호를 이용하여 정밀성과 상호연관성을 최대한 보존할 수 있도록 하였다.

마지막으로, 본 논문에서 개발한 GATA2는 NewsGroup 20 문서를 분류하여 기존의 LSI와 비교하여 전체적으로 우수함을 입증하였다.

그러나, 본 논문에서 제안된 GATA2는 수행시간이 너무 길고, 이를 반복하여 적용해야하기에 성능은 우수할 수 있으나 실용화하기에는 무리가 있다고 보여 지며, 이를 개선하기 위한 방법론을 필요로 하고 있다.

References

[1] Chang, J. Y., “Topical Clustering Techniques of Twitter Documents Using Korean Wikipedia”, Journal of IIBC, 2014, Vol.14, No.5, pp.173-178.

[2] Deng C, Xiaofei H, Jiwei H., “Document Clustering Using Locality Preserving Indexing,” Jounal of IEEE Transaction on Knowledge and Engineering, 2005, Vol.17, No.12, pp.1624-1637.

[3] Landeuer, T.K., Foltz, P.W., and Laham, D., “Introduction to Latent Semantic Analysis,” Journal of Discourse Processes, 1998, Vol.25, No.2-3, pp.259-284.

[4] Salton G., Wong A., Yang C.S., ″A Vector Space Model for Automatic Indexing,″ Journal of Communications of the ACM, 1975 Vol.18, No.11, pp.613-620,

[5] Taiping Z., Yuan Y.T., Bin F., Yong X., “Document Clustering in Correlation Similarity

Measure Space," Journal of IEEE Transaction on Knowledge and Data Engineereing, 2012, Vol.24, No.6, pp.391-407.

[6] Teng, G., Xia, Y., Camria, E., Jin, P., and Zheng, T.F., "Document representation with statistical word senses in cross-lingual document clustering," Journal of Pattern Recognition and Artificial Intelligence, Vol.29, No.2, 2015, 1559003(26pages).

[7] Uysal, A.K., and Gunal, S., "Text Classification Using Genetic Algorithm Oriented Latent Semantic Features," Journal of Expert Systems with Applications, Vol.41, No.13, 2014, pp.5938-5947.

[8] Park D., C., Ronnel R. Atole "A Novel Multi-focus Image Fusion Scheme using Nested Genetic Algorithms with 'Gifted Genes'," Journal of IIBC, 2009, Vol. 9, No. 1, pp.75-87.

[9] Im S., J & Hwang H., J., "Design and Development of Simulation Framework for Processing Window Query in Wireless Spatial Data Broadcasting Environment," Journal of IIBC, 2014 Vol. 14, No. 5, pp.173-178.

저자 소개

김 정 준(정회원)

- 2003년 2월 : 건국대학교 컴퓨터공학과 학사
- 2005년 2월 : 건국대학교 컴퓨터공학과 석사
- 2010년 8월 : 건국대학교 컴퓨터공학과 박사
- 2010년 9월 ~ 2012년 8월: 건국대학교 컴퓨터공학과 강의교수
- 2012년 9월 ~ 2016년 2월: 건국대학교 컴퓨터공학과 조교수
- 2016년 3월 ~ 현재: 한국산업기술대학교 컴퓨터공학과 조교수

<주관심분야 : Database Systems, BigData, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.>

정 성 택(정회원)

- 1992년 2월: 한국과학기술원 전기 및 전자공학과 학사
- 1995년 2월: 한국과학기술원 정보 및 통신공학과 석사
- 2000년 2월: 한국과학기술원 전기 및 전자공학과 박사
- 1998년 1월 ~ 2000년 4월: (주)메디슨 MRI사업부 선임연구원
- 2000년 5월 ~ 2004년 2월: (주)메디너스 MRI연구소 연구소장
- 2004년 3월 ~ 현재: 한국산업기술대학교 컴퓨터공학과 교수

<주관심분야 : 영상처리, 헬스케어, 기능성 게임, HCI, etc.>

박 정 민(정회원)

- 2003년 2월: 한국산업기술대학교 컴퓨터공학과 학사
- 2005년 2월: 성균관대학교 컴퓨터공학과 석사
- 2009년 2월: 성균관대학교 컴퓨터공학과 박사
- 2011년 2월 ~ 2012년 6월: 성균관대학교 연구교수
- 2012년 7월 ~ 2014년 2월: 한국전자통신연구소 선임연구원
- 2014년 3월 ~ 현재: 한국산업기술대학교 컴퓨터공학과 조교수

<주관심분야: 사이버 물리 시스템, 자율 컴퓨팅, 소프트웨어 공학>

※ 이 성과는 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2017R1A2B4011243).