한국정보전자통신기술학회논문지(jkiiect)'15-12, Vol.8 No.6
논문 15-08-06-544 http://dx.doi.org/10.17661/jkiiect.2015.8.6.544

오피니언 마이닝 기반 SNS 감성 정보 분석 전략 설계

정은희, 이병관*

A Design of SNS Emotional Information Analysis Strategy based on Opinion Mining

Eun-Hee Jeong, Byung-Kwan Lee*

요 약 현재, SNS으로 소통되는 의견들이 증가하고 있기 때문에 SNS 메시지로부터 의미 있는 정보를 유추해내는 오피니언 마이닝(Opinion mining) 기술이 중요해지고 있다. 본 논문은 반의어와 부사의 위치에 따라 가중치를 다르게 설정하여 SNS의 감성 정보를 정확하게 추출하는 오피니언 마이닝 기반 SNS 감성 정보 분석 전략(SEIAS, SNS Emotional Information Analysis Strategy)을 제안한다. 제안하는 SEIAS(SNS Emotional Information Analysis Strategy)는 첫째, 오피니언 마이닝 분석에 필요한 감성사전을 구축하고, 둘째, SNS 데이터를 실시간으로 수집하고, 수집된 SNS 데이터와 감성사전를 비교하여 SNS 데이터의 의견값을 산출한다. 특히, 데이터의 의견값을 산출할 때, 반의어, 부사의 위치에 따라 가중값을 다르게 설정함으로써 기존의 SO-PMI와 비교하였을 때 오피니언 분석결과의 정확도를 향상시켰다.

Abstract The opinion mining technique which analogize significant information from SNS message is increasingly important because opinions communicated through SNS are increasing. This paper propose SEIAS(SNS Emotional Information Analysis Strategy) based on opinion mining that analogize emotional information from SNS setting a different weight according to position of antonym and adverb. Firstly, the proposed SEIAS constructs a emotion dictionary for opinion mining analysis, Secondly, it collects SNS data on real time, compare it with emotion dictionary and calculates opinion value of SNS data. Specially, it increases the precision of opinion analysis result compared to the existing SO-PMI because it sets up the different value according to the position of antonym and adverb when it calculates the opinion value of data.

Key Words : emotional information analysis strategy, opinion mining, antonym, adverb, different weight, SEIAS

1. 서론

최근 스마트 디바이스의 발달로 인해 언제 어디서나 소셜 네트워크 서비스에 접속할 수 있는 환경이 조성되었고, 그로인해 블로그, 트위터, 페이스 북과 같은 SNS 시장이 급속히 성장하게 되었다. 특히, 스마트폰을 활용하여 맛집이나 영화, 제품에 대한 다양한 리뷰들이 트위터, 인스턴트 메시지 모드 등과 같은 SNS에 실시간으로 상품에 대한 의견을 표현함으로써 다른 구매자들에게 영향력을 행사할 수 있게 되었으며, 그 영향력은 점차 증가하고 있다[1-2]. 따라서 대량의 SNS 의견들로부터 사용자가 원하는 정보를 빠르게 분석해주고, 의미있는 정보를 지능적으로 유추해내는 오피니언 마이닝(Opinion mining) 기술의 중요성은 어느 때보다 커지고 있는 실정이다[1-3].

This paper presents results from a study being carried out of the academic-industrial collaborative technology development projects supported by the Small Business Administration in 2014.(No.C0250089)
*Corresponding Author : Department of Computer Engineering, Catholic Kwandong University (bklee@cku.ac.kr)
Received December 01, 2015 Revised December 10, 2012 Accepted December 19, 2015

본 논문에서는 오피니언 마이닝 분석에 필요한 감성사전을 구축하고, SNS으로 작성되는 실시간 리뷰들을 수집하여 감성사전과 비교하여 SNS 데이터의 의견값을 산출한다. 이때, 반의어와 부사의 위치에 따라 다르게 가중치를 설정하여 좀 더 정확한 의견값을 산출할 수 있는 오피니언 마이닝 기반 SNS 감성정보 분석 전략인 SEIAS(SNS Emotional Information Analysis Strategy)을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 오피니언 마이닝에 대한 관련연구를 설명하고, 3장에서는 본 논문에서 제안하고 있는 SEIAS를 설명한다. 그리고 4장에서는 SEIAS의 분석 결과를 설명하고 5장에서 결론을 맺는다.

2. 관련연구

2.1 오피니언 마이닝

오피니언 마이닝은 기본적으로 문서가 긍정, 부정, 또는 중립 중 어떤 견해를 갖고 있는지 판별하는 일련의 과정이라 볼 수 있으며, 분석은 각 문서 최소 단위인 어휘의 감성 극성에 기반하여 이루어진다. 즉, 주요 어휘의 감성 극성이 미리 정의된 감성 사전을 구축한 후, 새로 주어진 문서에 출현한 어휘의 감성 극성에 따라 문서 전체의 감성 극성을 분류하게 된다. 즉 오피니언 마이닝은 미리 구축된 감성 사전을 사용하는 반면, 텍스트 마이닝을 입력 데이터에 대한 학습을 통해 다른 데이터를 예측한다는 차이점을 갖는다[3, 4].

2.2 오피니언 도출의 판별 기법

PMI(Point-wise Mutual Information) 기법은 주어진 두 단어가 긍정인지 또는 부정인지를 판별하기 위한 극성 분류 방법으로 다른 분류기법보다 사용 및 계산이 단순하고, 정확한 결과를 예측할 수 있으므로 오피니언 마이닝 분야에서 가장 널리 이용되는 극성 분류 기법이다[3,5,6].

$$PMI(word_1, word_2) = \log_2\left(\frac{P(word_1, word_2)}{P(word_1)P(word_2)}\right) \quad (1)$$

여기서 $P(word_1, word_2)$는 두 단어 $word_1$과 $word_2$가 같은 서술어 안에서 나올 확률을 말한다. 두 단어가 나올 확률이 독립적이면 결과값이 0이 나올 것이고, 양수이면 확률이 높아 비슷한 의미의 극성을 가지는 것을 의미한다. 음수이면 확률이 낮아 다른 의미의 극성을 가지는 것을 의미한다. 그리고 $P(word_1)$는 식(2)와 같다[2,3,6].

$$P(word_1) = \frac{1}{N} hit(word_1) \quad (2)$$

여기서 $hit(word_1)$는 단어가 포함된 문서의 개수이고, N은 전체 문서의 개수이다.

SO-PMI(Semantic Orientation from PMI)는 미리 긍정 어휘들과 부정 어휘들을 정의해 놓고 어휘의 극성을 분류하는 방법으로 SO-PMI값이 양수가 나오면 긍정이고, SO-PMI값이 음수가 나오면 부정으로 분류한다[2,6].

$$SO-PMI(word) = \sum_{pw \in PW} PMI(word, pw) - \sum_{nw \in NW} PMI(word, nw) \quad (3)$$

여기서 pw는 긍정 어휘 집합이고, nw는 부정 어휘 집합을 말한다. 어휘가 긍정적일수록 pw의 PMI값이 크고, nw의 PMI값이 작아 SO-PMI값이 크게 나올 것이다. 반대로 어휘가 부정적일수록 SO-PMI값이 작게 나올 것이다. 즉, 두 어휘의 의미 극성을 상대적으로 비교할 수 있다[7].

3. SEIAS 설계

본 논문에서는 반의어와 부사의 위치를 활용한 오피니언 마이닝 기반 SNS 감성 정보 분석 전략인 SEIAS(SNS Emotional Information Analysis Strategy)을 제안한다. 제안하는 SEIAS는

첫째, SNS에서 사용하는 일반단어, 함축단어에

대한 사전을 생성하고,

둘째, 주어진 키워드를 포함하고 있는 SNS를 수집하여 키워드와 관련된 특징 명사를 추출하고

셋째, 수집한 SNS를 단어(토큰)로 분리하고 일반단어사전과 함축단어사전을 이용하여 단어들의 긍정, 부정, 그리고 중립으로 분류하여 SNS의 문장에 대한 의견값을 산출한다. 이때, 반의어와 부사의 위치에 따라 다르게 설정된 가중치를 이용한다.

그림 1은 본 논문에서 제안하는 SEIAS에 대한 전반적인 절차를 설명하고 있다.

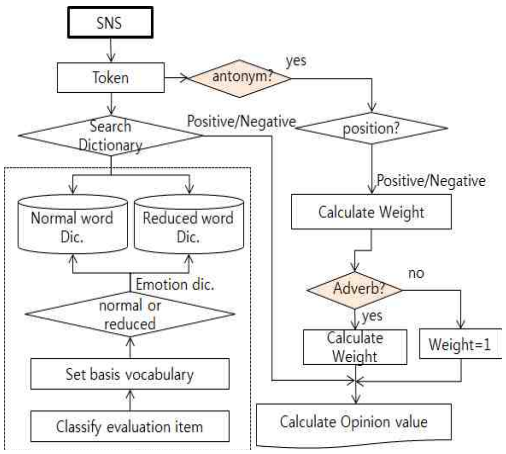

그림 1. SEIAS의 절차

Fig.1. The procedure of SEIAS

3.1 단어 사전 생성

감성정보 분석에서 가장 중요한 것은 단어 사전이다. 본 논문에서는 SNS를 통해 수집된 문장에 대한 감성 정보 분석을 위해 우선적으로 긍정 및 부정으로 명확하게 분류할 수 있는 일반 단어사전와 SNS에서 보편적으로 사용되는 함축어에 대한 단어를 등록한 함축 단어사전을 생성하도록 설계한다. 그리고 SNS의 토큰 분석 시에 새롭게 등장하는 단어를 긍정 및 부정으로 분류하여 일반 단어사전과 함축 단어 사전에 등록하도록 설계한다.

표 1은 본 논문에서 설계한 일반단어 사전과 함축단어 사전 구조를 설명한 것이다. 일반사전은 등록되는 단어에 대한 분류를 구분하도록 설계하여 단어 관리의 편리성을 도모하였고, 함축단어 사전의 경우에는 등록되는 함축어에 대한 설명을 추가하여 차후에 단어의 의미를 활용할 수 있도록 설계하였다. 또한 각 사전에 빈도(frequency) 필드를 추가하여 SNS에서 많이 사용되는 단어들을 파악할 수 있도록 하였다.

표 1. 단어 사전 구조

Table 1. Word dictionary structure

(a) 일반단어 사전

(a) normal word dictionary

Field name	Description
no	index number
classification	Classify
word	word
status	positive, negative, neutrality
frequency	frequency

(b) 함축단어 사전

(b) implication word dictionary

Field name	Description
no	Index number
word	Word
meaning	Meaning according to position
frequency	frequency

3.2 SNS 감성 정보 판단

SNS 문장의 의견값을 산출하는 SO-PMI는 긍정단어와 부정단어에 대한 횟수만으로 문장 전체의 의견값을 산출하는 단점을 가지고 있다. SEIAS는 이러한 SO-PMI가 갖는 단점을 보완하고 의견값을 좀 더 정확하게 산출하기 위해 SNS 문장을 토큰으로 분할하고, 긍정단어, 부정단어, 반의어에 대한 가중치, 그리고 부사에 대한 가중치를 부여하여 SNS 문장의 의견값을 좀 더 정확하게 산출하도록 설계하였다.

3.2.1 특징 명사 추출

본 논문에서는 그림 2와 같이 SNS로부터 특징명사를 추출하도록 설계한다.

특징명사를 추출하는 과정은 다음과 같다.

첫째, 선정한 주제에 맞는 SNS 데이터를 수집한다.

둘째, 수집한 SNS 데이터를 토큰으로 분리한 후에 명사를 추출한다.

셋째, 추출한 명사에 대한 누적 빈도수를 산출한다.

넷째, 수집된 SNS로부터 추출한 명사에 대한 누적빈도수가 임계치보다 큰 명사만을 특징 명사 테이블에 저장한다.

예를 들어, 삼척관광이라는 키워드를 입력하여 SNS 데이터를 수집하여 명사를 추출할 경우, 그림 2(b)와 같이 추출명사에 대한 출현 빈도수를 카운트한다. 그리고 출현 빈도수가 임계치보다 큰 추출명사만을 분류하여 특징명사 테이블에 저장하도록 설계한다. 차후에 이 특징 명사와 사용자의 특성 정보와의 피어슨 상관계수를 계산하여 사용자 맞춤형 감성정보를 제공하는데 활용하고자 한다.

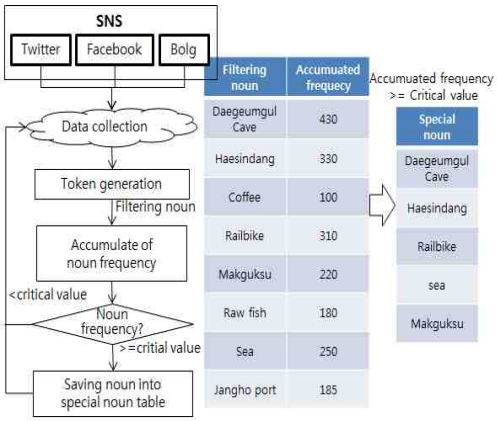

그림 2. 특징명사 추출 처리 절차

Fig. 2. The filtering procedure of special norm

3.2.2 감성 정보 측정

본 논문에서 제안하는 SNS로부터 감성 정보 산출하는 절차는 그림 3과 같으며 단계별 절차는 다음과 같다.

첫째, SNS로부터 수집한 데이터를 토큰으로 분리한다.

둘째, 데이터의 서술어를 탐지하고, 서술어가 존재하면 감성사전을 검색하여 서술어의 의견값을 식(4)와 같이 산출한다.

$$v.value = \begin{cases} +1, & \text{if 긍정} \\ 0, & \text{if 중립} \\ -1, & \text{if 부정} \end{cases} \quad (4)$$

예를 들어, “커피가 맛있다”일 경우 “맛있다”는 긍정(+1)이므로 문장의 의견값은 긍정(+1)이 된다.

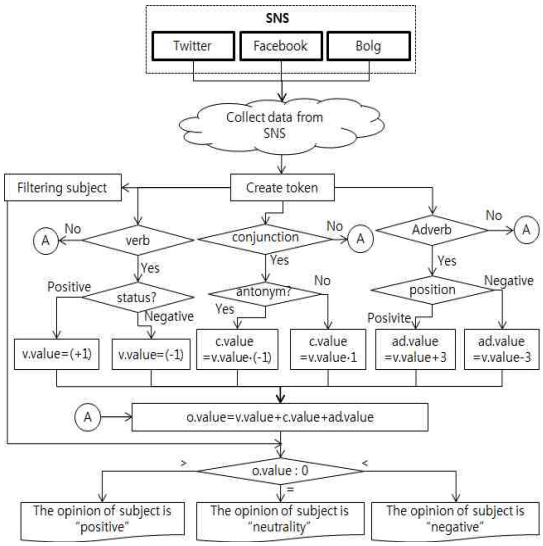

그림 3. 감성정보 산출

Fig. 3. the evaluation of emotion information

셋째, 데이터에 접속사가 존재하는지를 확인한다. 만약 데이터에 접속사가 존재하면, 이 접속사가 등위접속이면 앞의 서술어가 같은 의견값을 설정하고, 반의어이면 앞의 서술어와 다른 의견값을 설정한다.

$$c.value = \begin{cases} v.value \times (+1), & \text{if 서술어}_{\text{이전}} = \text{긍정 and 등위접속사} \\ v.value \times (-1), & \text{if 서술어}_{\text{이전}} = \text{긍정 and 반의어} \\ v.value \times (+1), & \text{if 서술어}_{\text{이전}} = \text{부정 and 반의어} \end{cases} \quad (5)$$

넷째, 데이터에 부사가 존재하는지를 확인한다. 만약 부사가 존재하면, 서술어의 의견값에 따라

의견값을 설정한다. 즉, 서술어의 의견값이 긍정이면 극정값을 설정하고, 서술어의 의견값이 부정이면 부정값을 설정한다.

$$ad.value = \begin{cases} v.value + 3, & \text{if } 부사_{이후} = 긍정 \\ v.value - 3, & \text{if } 부사_{이후} = 부정 \end{cases} \quad (6)$$

다섯째, 이렇게 계산된 의견값들을 식 (7)과 같이 주제어에 대한 의견값인 o.value를 산출한다. 그리고 o.value가 0보다 크면 긍정적, o.value가 0이면 중립, o.vlaue이 0보다 작으면 부정적으로 판단한다.

$$o.value = v.value + c.value + ad.value$$

$$o.value = \begin{cases} > 0, 긍정 \\ = 0, 중립 \\ < 0, 부정 \end{cases} \quad (7)$$

여섯째, 이렇게 산출된 의견값을 식 (8)을 이용해 특정 키워드에 대한 SNS의 데이터들의 감성평가에 따라 특정 키워드에 대한 전반적인 평가를 도출한다.

$$Opinion_{keyword} = \frac{\sum_{i=1}^{n} o.value_i}{n} \begin{cases} > 0, & 긍정 \\ = 0, & 중립 \\ < 0, & 부정 \end{cases} \quad (8)$$

4. 분석

본 논문에서 제안하는 SEIAS는 오피니언 마이닝에 필요한 사전을 구축하고, 제안하는 감성정보 판단 절차에 에 따라 문장의 의견값을 산출하였다.

표 2는 본 논문에서 설계한 감성 사전에 대한 긍정/부정 어휘를 설명한 것이다.

표 2. 긍정/부정 어휘(음식)
Table 2. Positive/Negative vocabulary(Food)

Type	Positive	Negative
price	cheap	expensive
	suitable	unsuitable
taste	delicious(good)	bad
	salty	bland
smell	good	bad
quantity	many(much)	few
	suitable	unsuitable
quality	satisfy	dissatisfy
	good	bad
	recommend	do not recommend

실험데이터는 국내 블로그 서비스 업체인 네이버, 다음, 티스토리, 트위터의 글 중에서 키워드에 대한 평가가 있는 글만을 수집하였다. 그리고 글 내에서 노출 빈도수가 주제어를 선별하여 특징 명사로 분류하였다.

표 3은 삼척이라는 키워드를 입력하여 수집한 네이버 블로그, 다음 블로그, 티스토리, twitter에서 노출 빈도수가 높은 주제어들을 추출한 결과이다.

표 3. 추출된 특징 명사들(상위 5개)
Table 3. Extracted special nouns(Top 5)

Number	Special nouns
1	Daegeumgul Cave
2	Haesindang Park
3	Railbike
4	Sea
5	Makguksu

SEIAS 실험에서는 평가의 정확도를 향상시키기 위해, 실험 데이터를 다음과 같은 CASE로 구분하여 진행하였다.

CASE 1) 반의어가 포함되지 않은 경우
- Test 1 : 긍정 단어만 있는 경우
- Test 2 : 부정 단어만 있는 경우

CASE 2) 반의어를 포함하고 있는 경우
- Test 1 : 긍정단어+반의어+부정 단어
- Test 2 : 부정단어+반의어+긍정 단어

CASE 3) 부사를 포함하고 있는 경우
- Test 1 : 부사+긍정단어
- Test 2 : 부사+부정단어
- Test 3 : 부사+긍정단어+부정단어

Test 4 : 긍정단어+부사+부정단어

CASE 4) 다섯째, 부사와 반의어가 모두 포함되어 있으며, 부사가 반의어 앞에 포함되어 있는 경우

Test 1 : 부사+긍정단어+반의어+부정단어

Test 2 : 부사+부정단어+반의어+긍정단어

표 4는 모의실험을 위해 작성된 샘플을 이용하여 본 논문에서 제안한 SEIAS와 기존의 SO-PMI의 평가 기법의 결과를 비교한 결과이다.

표 4. 감성 분석 결과(샘플)
Table 4. The result of emotion analysis(sample)

Case	Case 1		Case 2		Case3	
	Test1	Test2	Test1	Test2	Test1	Test2
positive	1	0	1	1	1	0
negative	0	1	1	1	0	1
antonym	0	0	1	1	0	0
adverb.	0	0	0	0	1	1
SO-PMI	1(S)	-1(S)	0(F)	0(F)	1(S)	-1(S)
SEIAS	1(S)	-1(S)	2(S)	-2(S)	3(S)	-3(S)

Case	Case3		Case4	
	Test3	Test4	Test1	Test2
positive	1	1	1	1
negative	1	1	1	1
antonym	0	0	1	1
adverb.	1	1	1	1
SO-PMI	0(F)	0(F)	0(F)	0(F)
SEIAS	3(S)	-3(S)	1(S)	-1(S)

*S:Success, F:Fail

표 4에서 설명하고 있듯이, 테스트에서 긍정단어의 수와 부정단어의 수가 같을 경우에, SO-PMI는 부사와 반의어를 고려하지 않았기 때문에 의견값이 정확하게 산출되지 않았지만, 제안하는 SEIAS는 부사와 반의어를 고려하였기 때문에 정확하게 의견값을 산출하였다. 따라서 제안하는 SEIAS가 기존의 SO-PMI와는 달리 정확도가 개선된 것을 알 수 있다.

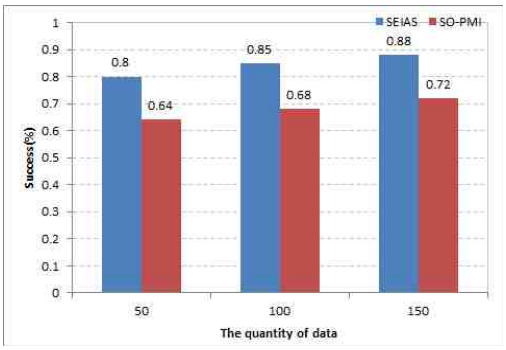

그림 4. 감성평과 결과
Fig. 4. The result of emotion evaluation

본 논문에서 제안하는 기법의 정확도를 측정하기 위해 네이버 영화 평가에 대한 데이터를 50, 100, 150개 단위로 수집하여 평가한 결과이다. 그 결과 그림 4와 같이 데이터의 수가 증가할수록 분석결과의 정확도가 증가되었으며, 평균적으로 SEIAS는 정확도가 84%로 측정되었지만, SO-PMI는 정확도가 68%로 측정되었다. 본 제안하는 SEIAS가 부사와 반의어를 반영하여 의견값을 산출하였기 때문에 기존의 SO-PMI보다 정확도가 약 16% 증가하였다.

5. 결론

본 논문에서는 반의어와 부사의 위치에 따라 가중치를 다르게 설정하여 SNS의 감성 정보를 좀 더 정확하게 산출할 수 있는 SEIAS를 제안한다. 제안하는 SEIAS에서는

첫째, SNS 문장내의 단어들을 긍정 및 부정으로 판단할 수 있는 일반 단어 사전을 생성하였다.

둘째, SNS에서 빈번하게 사용되는 함축단어에 대한 함축 단어 사전을 생성하였다.

셋째, 감성분석에 반의어와 부사에 대한 가중치를 위치에 따라 설정하고, 이 가중치를 이용하여 의견값을 산출함으로써 기존의 SO-PMI보다 정확도를 개선하였다.

향후, 일반단어 사전과 함축단어 사전을 확장하고, 좀 더 다양한 SNS 문장을 이용하여 실험한다면, 좀 더 정확한 감성정보 분류 결과의 정확도를

향상시킬 수 있을 것이다.

REFERENCES

[1] Gyung-Mi Park, Ho-Gun Park, Hyoung-Gon Kim, Hee-Dong Ko, “A Study of Opinion mining on SNS,” Communication of KIISE, vol.29, no.11, pp.54-60, 2011.

[2] Jung-Yeol Seo, Chan Koh, “Big Data Analysis by Sensitivity Analysis,” Journal of the Society of Convergence Knowledge, vol.2, no.1, pp.15-21, 2014.01.

[3] Eun-Hee Jeong, Byung-Kwan Lee, Yusrina Tifani, “A Emotion Information Analysis of SNS using Opinion Mining”, the Korea Institute of Information, Electronics, and Communication Technology, vol.8, no.2, pp.199-201, 2015.

[4] Seung-Woo Kim, Nam-Gyu Kim, “A Study of Emotional Dictionary Application Effect of Opinion Classifier,” J Intell Inform Syst, vol.20, no.1, pp.133-148, 2014.33.

[5] P. Turney, M. Littman, “Measuring praise and criticism:inference of semantic orientation from association,” Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp.417-424, 2002.

[6] Ji-Hoon Seo, Hye-Jin Jo, Jin-Tak Choi, “Design for Opinion Dictionary of Emotion Applying Rules for Antonym of Korean Grammar”, Journal of KIIT, vol.13, no.2, pp.109-117, 2015.

[7] Sang-Il Song, Dong-Joo Lee, Sang-Goo Lee, “Identifying Sentiment Polarity of Korean Vocabulary Using PMI,” Korea Computer Congress 2010, vol.37, no.1, pp.260-265, 2010.

저자약력

정 은 희(Eun-Hee Jeong) [종신회원]

- 1998년 2월 : 관동대학교 컴퓨터공학과 (공학석사)
- 2003년 2월 : 관동대학교 컴퓨터공학과 (공학박사)
- 2003년 9월 ~ 현재 : 강원대학교 지역경제학과 교수

<관심분야> 전자상거래 보안, 빅데이터, 헬스케어, IoT

이 병 관(Byung-Kwan Lee) [종신회원]

- 1986년 2월 : 중앙대학교 전자계산공학과 (공학석사)
- 1990년 2월 : 중앙대학교 전자계산공학과 (공학박사)
- 1988년 3월 ~ 현재 : 가톨릭관동대학교 컴퓨터공학과 교수

<관심분야> 네트워크 보안, 빅데이터, 데이터 마이닝, IoT