Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Improving Accuracy of Movie Recommender System Using Word2Vec and Deep Neural Networks

Boo-Sik Kang*

Division of Service Management, MokWon University

ABSTRACT

Word2Vec is a most popular method in text mining area, recently. It converts words to vectors using association among words in sentences. Similar words are nearly located in the vector space. As deep learning technology has developed rapidly, deep neural network that adopts deep learning is being applied in many areas. Improving predictive accuracy of recommender algorithms is a major work in the area of recommender systems. This study proposed an integrated method for movie recommender systems using Word2Vec and deep neural networks. First, it constructs Corpus of users and movies. It generates sentences for constructing Corpus. It finds users that give same rating to a movie, generates a sentence with those users, and constructs the Corpus of users using the sentences. It finds movies that are given same rating from an user, generates a sentence with those movies, and constructs the Corpus of movies using the sentences along the same lines. Secondly, it calculates user vectors and movie vectors using Word2Vec. Thirdly, it learns a model of deep neural networks with inputs composed of the user vectors and the movie vectors. Lastly, it recommends movies to the target user through the learned deep neural networks. To validate, the proposed method was applied to filmtrust dataset. The experimental results of 10-fold cross validation showed that the proposed method improved accuracy greatly than conventional collaborative filtering method(uCF). Also, it showed that the proposed method could solve problems with comparatively high accuracy, recommendation problems to new customers and new products, but the conventional collaborative filtering methods had difficulties recommendation to those problems.

KEYWORDS: Word2vec, Collaborative filtering, Recommender systems, Deep neural networks, Deep learning

ARTICLE INFO: Received 21 August 2018, Revised 20 September 2018, Accepted 12 October 2018.

*Corresponding author is with the Division of Service Management, MokWon University, 88 Doanbuk-ro Seo-gu Daejeon, 35349, KOREA.

E-mail address: bookang@mokwon.ac.kr

1. 서 론

전자상거래가 증가할수록 상품추천시스템의 활용은 더욱 커져간다. 상품추천시스템에서 가장 많이 활용하고 있는 기법 중의 하나는 협업필터링 방식이다[1,2]. 상품추천시스템에서 상품추천의 정확도 개선은 중요한 주제로서, 협업 필터링 기법의 정확도 개선을 위한 많은 연구가 제안되고 있다[3,4]. 전통적인 사용자기반 협업필터링 방식은 추천 대상자와 성향이 유사한 이웃사용자들을 찾아내고, 이웃사용자들의 구매정보를 활용하여 추천 대상자가 구매하지 않은 상품들 중에서 가장 구매 가능성이 높게 예측되는 상품들을 추천한다[3,4].

최근 자연어 처리 분야중 하나인 텍스트 분석 분야에서 문서분류[5], 감성사전 구축[6], 한국어 신문기사 분류[7], 문서기반 검색방법[8], 감성분석에 의미론적 요소 반영[9]등 Word2Vec이 활발하게 사용되고 있다. Word2Vec은 단어를 벡터로 변환하는 기법으로, 문장 내부의 단어 간 연관성을 파악해 벡터로 변환하며, 유사한 단어는 벡터 공간에서 가까운 거리에 위치하게 된다[10]. Word2Vec은 텍스트 분석외의 영역에서도 응용할 수 있는 데, 협업필터링 기법에서도 활용하는 방안이 여러 연구에서 제안되고 있다[4,11,12,13]. 전통적인 협업필터링 방식은 사용자 평점의 상관 유사도를 사용하는데 반해, [4,11,12,13]의 연구에서는 사용자의 평점을 직접 사용하지 않고 Word2Vec을 이용하여 사용자나 상품의 특성을 다차원의 벡터로 변환하고, 변환된 벡터의 유사도를 구하여 추천한다.

최근 이미지 처리나 자연어 처리 영역에서 딥러닝 기법이 높은 성과를 나타내면서, 딥러닝 기법을 추천시스템에 적용하는 연구도 제안되고 있다[14,15]. [14,15]는 사용자 평점 정보가 없는 경우의 추천을 위해 사용자의 영화의 시청순서를 고려하여 순환신경망을 이용한 추천방안을 제안하고 있다.

본 논문은 사용자 평점정보가 있는 환경하에서 사용자간 유사도를 이용하여 추천하는 전통적인 협업필터링 방식 대신에 Word2Vec과 딥러닝 기법 중 심층 신경망을 결합하여 영화추천시스템의 예측 정확도를 개선하는 방안에 대해 제안하고, 실험분석을 통해 검증하고자 한다.

또한 전통적인 협업필터링 방식은 신규고객이나 신규상품에 대해 추천하기 어려운 문제점을 갖고 있다. 제안된 방안의 경우 신규상품이나 신규고객의 경우에도 추천가능 함을 보이고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 Word2Vec과 심층 신경망에 대해서 살펴본다. 제 3장에서는 본 연구에서 제안한 방안에 대해 설명한다. 제 4장에서는 제안한 방법을 filmtrust 데이터[16]에 적용하여 실험하고 분석한다. 제 5장에서는 결론을 기술한다.

2. Word2Vec과 심층 신경망

2.1 Word2Vec

Word2Vec은 2013년 구글에서 발표된 연구로, 자연언어, 음성, 이미지와 같이 사람과 관련된 영역에서 많이 사용되고 있으며, 모델 내부에서 심층 신경망을 이용해 문장내의 단어를 다차원의 벡터로 변환하여 주는 모델이다[10].

Word2Vec 기능은 파이썬에서 Gensim[17] 라이브러리를 주로 사용하며, Word2Vec을 사용하는 과정은 개략적으로 다음과 같다[4,10]. 먼저, 하나의 문서는 여러 문장으로 구분된다. 각 문장은 의미있는 단어들의 조합으로 구성된 말뭉치로 정리된다. 각 문장은 다양한 단어들로 구성되어 있어 문서를 문장-단어 매트릭스로 표현하면 매우 큰 사이즈의 희박행렬로 표현된다. 정리된 전체 말뭉치를 Word2Vec에 입력으로 넣고, 다차원의 출력 크기

M을 지정하면, 출력으로 각 단어 간 연관성을 파악하여 M 차원의 다차원 벡터를 구한다.

2.2 심층 신경망

심층 신경망은 신경망의 은닉계층이 2개 이상인 신경망이다. 심층 신경망의 경우 복잡한 문제를 모델링할 수 있으나 그래디언트 소실문제, 낮은 학습속도 등의 문제가 있었다[18]. 최근 딥러닝 기술 및 GPU 활용 기술이 발전하면서 심층 신경망이 가졌던 문제들이 해결되고 이미지 처리, 자연어 처리 등 많은 영역에서 활발하게 사용되고 있다[18]. 심층 신경망의 역전파 알고리즘이 가졌던 그래디언트 소실문제는 ReLU 활성화 함수와 드롭아웃 방법을 활용하여 개선하고, 늦은 학습속도는 GPU를 활용함으로서 개선하였다.

3. 제안 추천 알고리즘

본 연구는 영화추천의 정확도를 높이기 위해서 Word2Vec과 심층 신경망을 결합하는 방안에 대한 연구로, 이를 위한 절차는 <그림 1>과 같다.

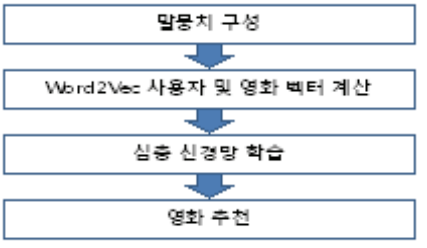

그림 1. 추천 알고리즘 절차
Figure 1. Procedure of the recommendation algorithm

말뭉치 구성단계에서는 사용자 구매 상품 및 평점 정보를 이용하여, 사용자 벡터를 구하기 위한 사용자로 구성된 문장과 영화 벡터를 구하기 위한 영화로 구성된 문장을 구성한다. 문장은 단어로 구성되는데, 여기에서는 사용자 ID, 영화 ID를 단어로 간주한다.

사용자로 구성된 문장을 구하는 과정은 다음과 같다. 같은 영화를 본 모든 사용자 ID와 사용자별 영화 평점을 찾는다. 이 사용자 ID를 단어로 간주하여 영화 당 문장을 구성한다. 선행연구[4]에 의하면 영화 당 문장을 확장하여 구성하는 방법의 정확도가 높았으므로 이 연구에서도 문장 확장 구성 방법을 사용한다. 예를 들어, 영화1을 본 모든 사용자ID를 찾고 같은 평점을 부여한 사용자들로 하나의 문장을 구성하였다. 즉 4점을 부여한 사용자들부터 0.5점을 부여한 사용자들까지 영화1에 대해 여러 개의 문장을 구성한다. 이 과정을 모든 영화에 대해 반복하여 사용자 전체 문장을 구한다. 각 문장 당 단어(사용자 ID)의 조합이 사용자 말뭉치가 된다. 영화에 대한 문장을 구하는 과정은 사용자별로 본 영화들을 찾고, 같은 평점을 부여한 영화별로 문장을 구성한다. 전체 사용자에 대해 반복하면 영화 전체 문장을 구할 수 있고, 각 문장 당 단어(영화 ID)의 조합이 영화 말뭉치가 된다.

Word2Vec을 이용한 사용자 벡터 계산단계에서는 전 단계에서 구성한 사용자 전체 문장의 말뭉치를 입력으로 넣어 사용자 벡터를 계산한다. 선행연구[4]의 결과를 이용하여, 사용자 벡터의 크기는 200으로 설정하였고, 윈도우 크기는 10으로 설정하였다. 200차원의 벡터는 단어의 연관성을 파악하여 구하는데, 각 문장 내 인접 단어 10개(윈도우 크기)까지의 연관성을 고려하여 구한다. 영화 벡터에 대해서도 동일하게 계산한다. 이 단계의 결과 사용자 ID별로 200차원의 벡터와 영화 ID별로 200차원의 벡터를 얻게 된다.

심층 신경망의 학습을 위한 구성은 <그림 2>와 같다.

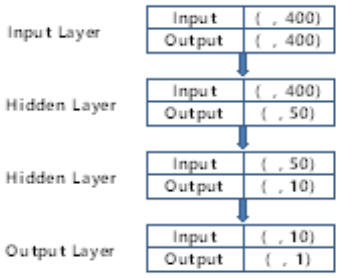

그림 2. 심층 신경망 모델구조
Figure 2. A model architecture of deep neural networks

{사용자ID, 영화ID, 평점}으로 구성된 원시 데이터를 {사용자 벡터, 영화 벡터, 평점}으로 먼저 변환한다. 입력 계층에서는 사용자 벡터(200차원)와 영화 벡터(200차원)로 구성된 400차원의 데이터를 입력으로 받아 변환없이 출력한다. 첫 번째 은닉계층은 400차원 데이터를 입력으로 받아 50차원의 데이터를 출력한다. 활성함수는 ReLU를 사용한다. 두 번째 은닉계층은 50차원 데이터를 입력으로 받아 10차원 데이터를 출력한다. 활성함수는 ReLU를 사용한다. 출력계층은 10차원 데이터를 입력으로 받아 1차원 데이터를 출력한다. 출력계층의 활성함수는 지정하지 않는다. 출력계층의 1차원 출력데이터는 사용자별 영화 평점 정보를 학습하게 된다.

사용자를 위한 영화별 평점 예측작업 후에 예측 평점이 높은 영화 순으로 추천 대상자에게 영화를 추천한다.

4. 실험분석

4.1 실험데이터

본 연구에서 제안된 방안을 평가하기 위해 LibRec[19]에서 제공하는 filmtrust 데이터를 사용하였다. filmtrust 데이터는 1508명의 사용자가 2071개 영화에 대해 0.5점에서 4점사이의 영화 평점을 부여한 데이터로 {userid, movieid, movieRating} 3개의 속성과 35,497개의 사례로 구성되어 있고 데이터 밀도는 1.14%이다[3,16]

4.2 실험 및 결과분석

본 연구에서 제안한 방식의 추천 정확도 평가를 위한 평가척도로 협업필터링 연구에서 일반적으로 많이 이용하는 실제값과 예측값의 절대값평균오차(Mean Absolute Error, MAE)를 사용하였다[1,3]. 전체 데이터는 학습데이터(90%)와 시험데이터(10%)로 구분되고, 학습데이터를 이용하여 영화추천 모델을 구축하고 시험데이터내의 영화 평점을 예측하였다. 이와 같은 과정을 10번 반복하는 10겹 교차검증시험을 실시하였다. 시험은 Keras 2.1.3과 Gensim 라이브러리의 Word2Vec을 사용하였다. Word2Vec의 하이퍼 파라미터인 벡터 크기 M과 윈도우 크기 W은 선행연구[4]를 이용하여 M=200, W=10으로 설정하여 실험하였다. 심층 신경망의 구성은 사전 실험에서 다양한 형태로 실험하였으며 비교적 성능이 우수한 <그림 2>의 모델을 최종 선택하여 실험하였다. 추가로 드롭아웃에 대해 사전 실험한 결과 드롭아웃은 0으로 설정하였다.

제안한 연구(wDNN)의 실험 결과는 선행연구[3,4]의 전통적인 사용자기반 협업필터링(uCF)의 결과와 비교하였다. 10-겹 교차검증을 실시한 결과는 <표 1>과 같다. uCF의 MAE 평균은 1.01894인 반면에 wDNN의 MAE 평균은 0.66909로, 제안된 방법(wDNN)이 전통적인 협업필터링 방식(uCF)에 비해 정확도를 크게 개선함을 알 수 있다.

예측 정확도 차이에 통계적 유의성 검정을 위해, 쌍체 t-검정을 실시하여, <표 2>의 결과를 얻었다. 쌍체 t-검정의 결과 통계적으로 유의한 결과를 얻었다.

표 1. 10-겹 교차검증시험의 MAE
Table 1. MAEs of 10-fold cross validation test

	uCF	wDNN
1	0.97775	0.66283
2	1.0287	0.67518
3	1.00804	0.66355
4	1.0217	0.66239
5	1.03705	0.675
6	1.00795	0.65444
7	1.01932	0.66949
8	0.9664	0.65885
9	1.03004	0.69103
10	1.09247	0.67813
mean	1.01894	0.66909

[4]에서는 전통적인 사용자 기반 협업필터링에서 사용자 유사도를 측정할 때, Word2Vec을 이용하여 사용자 벡터를 구한 후 이 사용자 벡터를 이용하여 사용자 유사도를 측정하고 이를 기반으로 하여 협업필터링 하는 방식([4]에서 제안한 방식은 wCF라 한다)이 추천 정확도를 개선함을 보였다. wCF의 10겹 교차검증결과 평균 MAE는 0.83502였다[4]. 본 연구에서 제안한 wDNN의 평균 MAE는 <표 1>에 보인 것처럼 0.66909로 [4]의 wCF방식보다도 본 연구에서 제안된 방식이 정확도를 크게 개선함을 알 수 있다.

표 2. uCF와 wDNN의 쌍체 t-검정 결과
Table 2. Results of pair-wise t-test of uCF and wDNN

	mean	var.	t-value	p-value
uCF	1,01892	0.00118	38.1	0.00000
wDNN	0.66909	0.00012		

또한 협업필터링 방식은 기본적으로 신규상품이나 신규고객에 대해 추천하기 어려운 문제점을 갖고 있다. 본 연구에서는 신규고객에 대해서는 사용자 벡터의 평균 벡터로 대체하고, 신규상품에 대해서는 영화 벡터의 평균 벡터로 대체하여 추가적으로 실험해 보았다. 실험결과는 신규고객이나 신규상품을 포함하여 10겹 상호 검증시험을 실시한 결과 평균 MAE는 0.67163으로 나타났다. 이 결과는 <표 1>에서 보인 것처럼 신규상품과 신규고객이 배제된 협업필터링 방식(uCF)의 추천 정확도인 1.01894보다 좋은 결과를 보였다.

결과적으로 Word2Vec과 심층 신경망을 결합한 영화추천 방식이 추천 정확도를 크게 개선할 수 있음을 실험결과가 보여 주었다. 또한 신규고객이나 신규상품의 경우 사용자 평균 벡터와 영화 평균 벡터를 이용하여 추천함으로서 전통적인 협업필터링 방식이 가지는 문제도 해결하면서도 비교적 높은 정확도를 나타냄을 알 수 있었다.

5. 결 론

추천 알고리즘의 예측 정확도를 높이는 것은 상품추천시스템 영역에서 주요 과제 중 하나이다[3,4]. 상품추천시스템에서 많이 활용되고 있는 방법으로 협업필터링 방식이 있다[1,2]. 전통적인 협업필터링 방식은 사용자간 유사도를 측정하여 유사도가 높은 이웃 사용자의 정보를 이용하여 추천한다. 본 연구에서는 영화추천의 예측 정확도를 높이기 위해서, 유사도를 이용한 추천방식 대신에 Word2Vec과 심층 신경망을 결합하는 방식에 대해 제안하였다. 본 제안의 추천절차는 Word2Vec을 이용하여 사용자 벡터와 영화 벡터를 구하고, 사용자 벡터와 영화 벡터를 심층 신경망으로 학습하여 추천한다.

제안한 방식(wDNN)을 선행연구[4]와 비교하였다. 제안된 방식은 전통적 사용자 기반 협업필터링 방식(uCF)에 비해 영화추천시스템의 정확도를 크게 개선하였다. 또한 Word2Vec을 활용하여 사용자 벡터를 구하고 사용자 벡터의 상관 유사도를 이용하는 사용자 기반 협업 필터링 방식(wCF)에 비해서도 정확도를 개선하였다. 또한 협업필터링

방식이 가지는 신규고객이나 신규상품에 대한 추천의 어려움에 대해서도, 본 연구에서 제안한 방식을 활용하면(신규고객에 대해서는 사용자 벡터의 평균벡터로, 신규상품에 대해서는 영화 벡터의 평균 벡터로 대체), 비교적 높은 정확도를 가지면서도 신규고객 및 신규상품 문제를 해결할 수 있음을 알 수 있었다.

본 연구에서 제안한 방식은 영화 데이터를 이용하여 검증하였으나 일반화를 위해서는 추가적으로 다양한 데이터에 대해 적용하여 실험할 필요가 있다. 다만 연구실험 결과는 추천시스템의 정확도 개선을 위해 Word2Vec과 심층 신경망이 효과적으로 활용될 수 있음을 보여 주고 있다. 현재 다양한 딥러닝 기법이 급속도로 발전하고 있어 이후 여러 딥러닝 기법을 결합하여 추천정확도를 더욱 높이는 것이 향후 과제로 남아 있다.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, *Item-based collaborative filtering recommendation algorithms,* Proceedings of the 10th International Conference on World Wide Web, pp. 285-295, 2001.

[2] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, *EigenTaste: a constant time collaborative filtering algorithm,* Information Retrieval, Vol. 4, No. 2, pp. 133-151, 2001.

[3] B. S. Kang, *The study of recommendation algorithm's predictive accuracy improvement using structural holes on trust-based social networks,* Journal of Knowledge Information Technology and Systems, Vol. 12, No. 1, pp. 209-217, 2017.

[4] B. S. Kang, *Improving predictive accuracy of user-based collaborative filtering using word2vec*, Journal of Knowledge Information Technology and Systems, Vol. 13, No. 1, pp. 169-176, 2018.

[5] J. M. Kim, and J. H. Lee, *Text document classification based on recurrent neural network using word2vec*, Journal of Korean Institute of Intelligent Systems, Vol. 27, No. 6, pp. 560-565, 2017.

[6] C. Heo, and S. Y. Ohn, *A novel method for constructing sentiment dictionaries using word2vec and label propagation*, The Journal of Korean Institute of Next Generation Computing, Vol. 13, No. 2, pp. 93-101, 2017.

[7] D. W. Kim, and M. W. Koo, *Categorization of korean news articles based on convolutional neural networks using doc2vec and word2vec,* Journal of Korean Institute of Information Scientists and Engineers, Vol. 44, No. 7, pp. 742-747, 2017.

[8] W. J. Kim, D. H. Kim, and H. W, Jang, *Semantic extension search for documents using the word2vec*, Journal of the Korean Contents Associations, Vol. 16, No. 10, pp. 687-692, 2016.

[9] J. Y. Ahn, J. H. Bae, N. K. Nam, and M. Song, *A study of 'emotion trigger' by text mining techniques*, Journal of Intelligence and Information Systems, Vol. 21, No. 2, pp. 69-92, 2015.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality,* Advances in Neural Information Processing Systems 26, pp. 3111-3119, 2013.

[11] O. Barkan, and N. Koenigstein, *Item2vec: neural item embedding for collaborative filtering,* 2016 IEEE 26th International

Workshop on Machine Learning for Signal Processing, pp. 1-6, Sep. 2016.

[12] V. Kuzmin, *Item2vec-based approach to a recommender system*, Bachelor's Thesis, University of Tartu, 2017.

[13] G. S. Jeon, S. G. Kong, and Y. S. Cho, *Word2vec based collaborative filtering for movie rating prediction*, Korea Software Congress 2017, pp. 844-846, 2017.

[14] R. Devooght, and H. Bersini, *Collaborative filtering with recurrent neural networks,* arXiv preprint arXiv:1608.07400, 2016.

[15] M. H. Kwon, S. E. Kong, and Y. S. Choi, *Improving recurrent neural network based recommendations by utilizing embedding matrix,* Journal of KIISE, Vol. 45, No. 7, pp. 659-666, 2018.

[16] G. Guo, J. Zhang, and N. Yorke-Smith, *A novel bayesian similarity measure for recommender systems*, Proceedings of the 23th International Joint Conference on Artificial Intelligence, pp. 2619-2625, 2013.

[17] Gensim, https://radimrehurek.com/gensim/. Sep. 2017.

[18] B. T. Zhang, *Deep hypernetwork models,* Communications of the Korean Institute of Information Scientists and Engineers, Vol. 33, No. 8, pp. 11-24, 2015.

[19] Librec, http://www.librec.net/datasets.html. Oct. 2016.

Word2Vec과 심층 신경망을 활용한 영화 추천 시스템의 정확도 개선

강부식

목원대학교 서비스경영학부

요 약

Word2Vec은 최근 텍스트 마이닝에서 가장 활발하게 사용되고 있는 방법 중 하나이다. 텍스트 문장내의 단어 간의 연관성을 파악하여 문장의 단어를 벡터로 변환한다. 유사한 단어일수록 벡터 공간에서 가까이 위치한다. 최근 딥러닝 기술이 발전하면서 많은 분야에서 딥러닝을 활용한 심층 신경망이 활용되고 있다. 추천 알고리즘의 정확도를 높이는 것은 상품추천시스템 영역에서 주요 과제 중 하나이다. 본 연구에서는 Word2Vec과 심층 신경망을 활용한 영화추천 방식에 대해 제안하였다. 먼저 사용자와 영화에 대한 말뭉치를 구성한다. 말뭉치를 구성하기 위해 문장을 생성한다. 동일한 영화에 대해 같은 평점을 부여한 사용자들을 찾고 이 사용자들로 문장을 구성하여 사용자 전체 말뭉치를 구성한다. 비슷하게 동일한 사용자가 같은 평점을 부여한 영화들을 찾고 이 영화들로 문장을 구성하여 영화 전체 말뭉치를 구성한다. 다음에, Word2Vec을 이용하여 사용자 벡터와 영화 벡터를 구한다. 이들 벡터를 심층 신경망의 입력으로 넣어 심층 신경망을 학습한다. 학습된 신경망을 이용하여 추천 대상자에게 영화를 추천하게 된다. 제안한 방안의 검증을 위해 filmtrust 데이터에 적용하여 실험하였다. 10-겹 교차검증을 실시한 결과 Word2Vec과 심층 신경망을 활용한 영화 추천방식이 전통적인 사용자기반 협업필터링 방식에 비해 정확도를 크게 개선함을 알 수 있었다. 또한 전통적인 협업필터링 방식이 신규상품이나 신규고객에게 추천의 어려움이 있는 문제를 본 연구에서 제안한 방식에서는 비교적 높은 정확도를 가지면서도 해결 가능함을 알 수 있었다.

Boo Sik Kang received the bachelor's degree in the Department of Industrial Engineering from the KyungHee University in 1985. He received the M.S. degree and the Ph.D. degree in the Department of Industrial Engineering from KAIST in 1989 and 2000, respectively. From 1989 to 2001, he was a researcher at Korea Telecom. He was a professor in the Division of Service

Management at MokWon University from 2001 to 2018. His current research interests include data mining, customer relationship management, service quality management. He is a life member of the KKITS.

E-mail address: bookang@mokwon.ac.kr