JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 1, pp. 265-274, January 2021
https://doi.org/10.9708/jksci.2021.26.01.265

Improvement of a Product Recommendation Model using Customers' Search Patterns and Product Details

Yunju Lee*, Jaejun Lee*, Hyunchul Ahn*

*Master's Candidate, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Master's Candidate, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

In this paper, we propose a novel recommendation model based on Doc2vec using search keywords and product details. Until now, a lot of prior studies on recommender systems have proposed collaborative filtering (CF) as the main algorithm for recommendation, which uses only structured input data such as customers' purchase history or ratings. However, the use of unstructured data like online customer review in CF may lead to better recommendation. Under this background, we propose to use search keyword data and product detail information, which are seldom used in previous studies, for product recommendation. The proposed model makes recommendation by using CF which simultaneously considers ratings, search keywords and detailed information of the products purchased by customers. To extract quantitative patterns from these unstructured data, Doc2vec is applied. As a result of the experiment, the proposed model was found to outperform the conventional recommendation model. In addition, it was confirmed that search keywords and product details had a significant effect on recommendation. This study has academic significance in that it tries to apply the customers' online behavior information to the recommendation system and that it mitigates the cold start problem, which is one of the critical limitations of CF.

▸ **Key words**: Recommender Systems, Search Keywords, Product Details, Collaborative Filtering, Doc2vec

[요 약]

본 논문에서는 검색 키워드와 상품 상세정보를 활용한 Doc2vec 기반의 새로운 추천 모형을 제안한다. 지금까지 추천 시스템에 관한 많은 기존 연구에서는 고객의 구매 이력이나 평점 같은 정형 데이터만을 사용하는 협업 필터링(CF) 알고리즘에 기반한 추천 모델이 제안되었다. 그러나 CF에서 온라인 고객 리뷰와 같은 비정형 데이터를 사용하면, 보다 나은 추천결과를 도출할 수 있다. 이에 본 연구에서는 기존 연구에서 거의 활용되지 않았던 검색 키워드 정보와 상품 상세정보를 제품 추천에 활용할 것을 제안한다. 본 연구의 제안 모형은 고객이 구매한 상품에 대한 평점, 검색어, 상품 상세정보를 종합적으로 고려한 CF 알고리즘을 이용해 추천결과를 생성한다. 이 때 비정형 데이터로부터 정량적인 패턴을 추출하기 위한 방법으로는 Doc2vec이 적용된다. 실험 결과 제안 모형이 기존 추천 모형보다 더 나은 성능을 보이는 것을 알 수 있었고, 검색어 및 상품 상세정보가 추천에 유의한 영향을 미치는 것을 확인하였다. 본 연구는 고객의 온라인 행동 정보를 추천시스템에 적용하였다는 점과 전통적인 CF의 한계 중 하나인 콜드 스타트 문제를 완화하였다는 점에서 학술적 의의가 있다.

▸ **주제어**: 추천시스템, 검색어, 상품 상세정보, 협업필터링, Doc2vec

• First Author: Yunju Lee, Corresponding Author: Hyunchul Ahn
*Yunju Lee (mmlas0ui@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Jaejun Lee (jrack707@naver.com), Graduate School of Business IT, Kookmin University
*Hyunchul Ahn (hcahn@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2020. 12. 22, Revised: 2021. 01. 20, Accepted: 2021. 01. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

오늘날 인터넷의 발달로 온라인 쇼핑이 일반화되면서, 고객은 다양한 상품정보를 쉽게 얻을 수 있게 되었다. 그러나 그 정보량이 너무 많기 때문에 고객은 원하는 상품과 필요한 정보를 찾기 위해 많은 시간을 들여야 했고, 이윽고 정보가 너무 많은 나머지 의사결정이 어려운 정보 과부하(Information overload) 문제에 빠지게 되었다.

이러한 문제를 예방하고자 학계 및 산업계에서는 고객이 구매한 상품이나 호감을 보인 상품 등을 참고하여 고객이 선호할 만한 상품을 제안하는 추천시스템에 주목하게 되었다[1]. 그런데 추천시스템의 성능이 좋지 못하면 오히려 고객의 불만 요소가 될 수 있기 때문에 추천 성능의 정확도 개선은 매우 중요한 문제가 되었다[2].

추천시스템 알고리즘 중 협업필터링(Collaborative Filtering; CF)은 추천 성능이 우수한 기법으로 알려져, 이를 활용한 연구가 많이 진행되고 있다. 협업필터링은 상품 간 혹은 고객 간 유사성을 계산한 후 서로 유사한 상품 및 고객의 정보를 기반으로 추천한다. 그런데 고객이 구매하는 상품의 수보다 쇼핑몰에서 판매되는 상품의 수가 훨씬 많기 때문에, 모든 상품에 대해 고객의 선호도를 파악하기가 쉽지 않다. 또한 신규 고객의 경우 구매 기록이 없기 때문에 추천이 어려운 콜드 스타트(Cold start) 문제가 발생하기도 한다.

이러한 한계를 극복하여 추천 성능을 개선하고자 많은 연구가 진행되고 있다. 추천 성능을 높인 연구들은 주로 추천 알고리즘을 개선하거나 새로운 데이터를 추가하여 추천 성능을 높인 경우가 많은데, 새로운 데이터를 활용하여 추천 성능을 높인 연구의 경우 정형 데이터만 활용한 연구가 대부분이었다. 그중 자주 사용되는 정형 데이터는 평점 데이터로, 평점은 고객의 선호를 직접적으로 나타내고 수리적 표현이 쉽다는 장점이 있어 현재까지도 많은 연구에서 사용되고 있다. 그러나 최근 평점에서 얻어진 고객 선호도 설명력에 의문이 제기되고 있고[3-5], 고객의 비정형 데이터를 사용하여 추천 성능을 개선한 연구가 등장하고 있어 비정형 데이터를 고려하는 연구가 더욱 필요할 것으로 보인다.

오늘날 비정형 데이터를 활용하여 추천 성능을 높인 연구는 주로 고객의 리뷰 데이터를 사용하였는데, 리뷰는 고객의 구매 결정에 중요한 요소 중 하나이고, 평점보다 사용자의 선호도를 상세히 나타내기 때문이다[5-6]. 그러나 고객이 구매 후 리뷰를 작성하지 않으면 데이터 수집에 어려움이 있어, 비인기 상품인 롱테일 상품의 경우 데이터 희소성(Sparsity)의 위험이 있다.

온라인에서 얻을 수 있고, 고객의 선호를 나타내는 비정형 데이터는 고객의 리뷰 외에도 무수히 많다. 특히 온라인 쇼핑몰의 경우 고객은 구매 의사결정 단계를 거쳐 최종 단계인 상품 구매에 이르게 되는데, 이 과정에서 고객은 관심 있는 상품의 추가 정보를 검색하고, 탐색하는 등 여러 행동을 통해 자신의 관심 및 선호를 드러내며, 일종의 기록 데이터를 생성하게 된다[7-8]. 본 연구는 이러한 고객의 온라인 행동에 주목하고, 특히 고객이 관심 및 선호도를 드러내는 행동인 검색 행동과 상세정보 보기 행동을 활용한 추천 시스템을 제안하고자 한다. 본 연구의 제안모델은 구매 여부만을 고려한 추천시스템에 비해 고객의 다양한 정보를 고려하였다는 점에서 콜드 스타트 및 데이터 희소성과 같은 한계를 극복하고, 추천 성능을 향상하고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안모델의 이론적 배경인 추천시스템, 협업필터링의 기본 개념과 비정형 데이터를 활용한 선행 연구에 대해서 살펴보고, 다음으로 본 연구가 활용하는 고객의 온라인 행동 정보에 대하여 살펴본다. 3장에서는 본 연구의 제안모델을 소개하고, 4장에서는 실증 분석에 사용된 데이터에 대한 소개와 실험, 실험 결과에 대하여 제시한다. 마지막으로 5장에서는 본 연구의 실험 결과에 따른 결론, 연구 성과 및 의의를 설명하고 본 연구의 한계점 및 향후 연구 방향을 제시한다.

II. Preliminaries

1. Related works

1.1 Recommender systems

고객은 물건을 구매하는 과정에서 수많은 상품을 보고, 그중에 어떤 물건을 선택할지 선택의 갈림길에 서게 된다. 이러한 상황에서 고객은 입소문이나 전문가의 조언과 같이 다른 사람들의 추천을 참고하여 의사결정을 내린다. 그런데 지인이나 전문가를 통한 추천은 상품 정보 수집에 한계가 있고, 새로운 상품에 대한 추천이 어렵다는 단점이 있다[1].

인터넷의 급속한 성장에 따라 온라인 쇼핑이 일반화되면서 고객은 다양한 상품과 정보를 쉽게 얻을 수 있게 되었다. 그런데 온라인에서 공유되는 정보량이 매우 많다 보니 고객은 원하는 상품을 찾기 위해 많은 정보를 처리해야 했다[9]. 이는 곧 정보가 너무 많은 나머지 의사결정이 어려운 정보 과부하(Information overload) 문제를 야기시켰고, 기업은 이러한 정보 과부하 문제를 해결하고자 고객별 정보를 선별하여 제공하는 개인화 추천시스템에 관심을 두게 되었다.

개인화 추천시스템은 새로운 상품 추천과 효율적인 정보 탐색이 가능한 시스템이다. 추천시스템은 고객의 행동 이력 데이터를 활용하여 고객의 선호에 맞는 상품을 추천하는데, 이러한 추천 방식은 알고리즘이 고객의 구매 패턴 및 선호도를 파악하기 때문에 새로운 상품이 등장하더라도 선호도 예측을 통해 추천이 가능하였다. 또한, 각 고객이 선호할만한 상품을 선별하여 제공하므로 정보 과부하 문제를 예방하였다[10]. 이에 따라 추천시스템 연구는 고객의 의사결정을 돕고 고객의 정보 탐색 및 수집의 부담을 덜어주었다는 점에서 많은 관심을 받았다.

추천시스템의 접근 방식으로는 크게 내용 기반 추천(Content-based Recommendation)과 협업필터링(Collaborative Filtering)으로 나누어진다. 내용 기반 추천시스템은 아이템의 속성을 분석하여 기존에 사용자가 선호했던 아이템과 유사한 것을 추천한다. 반면에 협업필터링은 고객의 상품 선호 데이터에 기반하여 목표 고객과 유사한 고객이 선호한 상품을 해당 고객에게 추천하는 방식이다. 협업필터링은 산업계 및 학계에서 우수한 성능을 가진 추천시스템으로 평가되고 있으며, 많은 연구가 고객의 구매 기록 및 평점 데이터를 활용하여 협업필터링 관련 연구를 진행하였다[12-14].

협업필터링은 기억 기반(Memory-Based), 모델 기반(Model-Based), 하이브리드(Hybrid) 세 가지 방식으로 구분되는데, 앞서 기술한 기본적인 협업필터링은 기억 기반 방식에 속한다. 기억 기반 협업필터링은 다시 아이템 기반, 사용자 기반 협업필터링으로 구분되며 아이템 기반 협업필터링은 아이템 간의 유사성을 분석하여 추천하는 방식으로, 상품 구매 이력에 따라 지속적인 추천이 가능하지만, 사용자 간 유사성이 고려되지 않기 때문에 선호도가 상이한 사용자의 경우 추천 성능이 낮아질 수 있다. 반면에 사용자 기반 협업필터링은 사용자 간의 유사성을 분석하여 추천 대상자와 가장 유사한 사용자를 찾아내고, 유사한 사용자의 구매 정보를 바탕으로 상품을 추천한다[11][15].

그러나 협업필터링에는 몇 가지 문제점이 존재한다. 먼저 협업필터링은 상품 구매 이력, 고객의 평점 등을 기반으로 추천하기 때문에 구매 이력이 없거나 많지 않은 신규 고객의 경우, 데이터가 희소하여 선호도를 예측하기 어려운 콜드 스타트(Cold Start) 문제가 발생한다. 또한, 온라인 쇼핑몰에서 판매하는 상품은 점차 다양해지고, 온라인 쇼핑이 일반화됨에 따라 이용 고객의 수 또한 급증하게 되어 추천 알고리즘 구현에 점차 많은 계산능력이 필요하게 되는 확장성(Scalability) 문제가 존재한다[16].

이러한 문제들을 해결하기 위해 많은 연구가 모델 기반 협업필터링 등 추천 알고리즘을 개선하거나, 고객의 선호도를 나타내는 새로운 데이터를 추가하는 방법으로 추천 성능을 향상하고 있다. 그러나 알고리즘을 개선하여 추천 성능을 높이는 연구에 비해 새로운 데이터를 활용한 연구는 상대적으로 부족한 실정이며, 고객 구매 이력 데이터를 제외하고 주로 사용된 새로운 데이터는 고객의 평점 데이터가 대부분이었다. 그런데 몇몇 선행 연구를 통해 평점 데이터만으로는 고객의 선호를 예측하기에 역부족이라는 문제가 지적되었으며, 정형 데이터뿐만이 아닌 비정형 데이터를 고려한 추천시스템 연구의 필요성이 대두되고 있다[4-5][17-18].

1.2 Prior studies that used unstructured data for recommender systems

위와 같은 흐름에 따라, 최근에는 비정형 데이터를 활용한 추천시스템 연구가 점차 증가하고 있다. 이러한 연구들은 Table 1에 제시된 바와 같이, 비정형 데이터를 통해 정형 데이터만으로는 얻을 수 없는 고객의 숨겨진 선호도를 추출하거나, 비정형 데이터 간 유사성을 계산하여 추천에 활용하였고, 또는 비정형 데이터가 가진 속성에 주목하는 등 다양한 시도가 있었다.

Table 1. Summary of the prior studies on recommender system using unstructured data

Purpose	Data	Ref.
To identify customer preferences	Customer review data	[4], [6], [19]
To calculate the similarity between users	Customer review data	[20]
	Customer search keywords data	[21]
To identify customers' attributes	Customer review data	[22]

이러한 연구들은 모두 비정형 데이터를 활용하여 추천 성능을 높인 결과를 보였는데, 먼저 비정형 데이터로 고객의 숨겨진 선호도를 추출한 연구들은 Table 1에 제시된 3개의 연구로 요약되며, 다음과 같다. Ma et al.(2017)는 고객 리뷰 데이터를 오피니언 마이닝 기법으로 분석하여 고객의 선호도를 추출한 후 협업필터링 알고리즘에 반영하였다. 그 결과, 협업필터링의 한계인 데이터 희소성 문제를 해결하였으며, 평점 데이터만 고려한 협업필터링 기

법보다 추천 성능을 향상하는 결과를 보였다[19]. 현지연 외(2019)의 연구 또한 고객 리뷰를 활용하여 고객의 선호도를 추출하였고, 추천시스템의 정확도를 개선하였다[4]. 해당 연구는 도메인 맞춤 감성 사전으로 리뷰가 가진 감성을 수치화한 후 평점과 결합하여 추천하는 모델을 제안하였는데, 제안모델이 평점만 고려한 일반적인 추천시스템보다 고객의 선호도를 더욱 정확하게 나타내어 더 높은 예측 정확도를 보여주었다. 윤소영과 윤성대(2020)의 연구에서도 고객 리뷰를 감성 분석하여 고객 선호를 도출한 후, 아이템 기반 협업필터링에 활용하였다[6]. 그 결과 정형 데이터만 사용한 전통적인 협업필터링 보다 추천 성능이 개선되는 결과를 보였다.

한편 비정형 데이터 간 유사성을 추출하여 추천시스템에 적용한 연구도 등장하였는데, 조승연 외(2015)의 연구는 토픽 모델링 기법을 기반으로 고객 리뷰를 활용한 추천시스템을 제안하면서 고객 리뷰간 유사성이 상품 추천에 유의미한 영향을 미친다는 점을 발표하였다[20]. 또한 조성은과 임희석(2019)은 고객 검색 키워드 데이터를 활용하여 키워드 간 유사성을 찾고 키워드 기반의 추천 서비스를 제안하였다[21].

마지막으로 비정형 데이터가 가진 속성에 주목한 연구도 있었는데, 이륜경 외(2019)는 토픽 모델링 기법을 활용하여 고객 리뷰가 포함하고 있는 다중 속성에 따른 추천시스템을 제안하였다[22].

이처럼 비정형 데이터를 활용한 추천 연구들은 비정형 데이터에 내포된 고객의 선호도 정보를 추출하고, 비정형 데이터 간 유사성이나, 속성 등을 파악하여 추천 알고리즘에 반영하였다. 그 결과, 협업필터링의 고질적 문제점인 데이터 희소성 및 콜드 스타트 문제를 해결하였고 추천 성능 또한 향상되어 비정형 데이터 기반 추천시스템의 활용 가능성을 보여주었다. 그런데 비정형 데이터를 사용한 추천 연구는 대부분 리뷰 데이터를 기반하여 진행되었으며, 리뷰 데이터가 아닌 비정형 데이터를 활용한 시도는 비교적 많지 않았다. 온라인 쇼핑몰과 같이 추천시스템을 활발히 사용하고, 다양한 비정형 데이터를 얻을 수 있는 산업의 경우 고객의 리뷰만으로 추천을 하는 것은 추천 성능의 한계가 있을 것으로 보이며, 이에 따라 다양한 비정형 데이터를 고려한 추천 연구가 필요할 것으로 보인다.

1.3 Online behavior of customers

고객은 상품을 구매하기 전 다양한 탐색 활동을 거쳐 구매 결정에 이르게 된다. 이러한 구매 결정 과정은 오프라인으로 구매 활동을 하는 고객이나 온라인을 통해 구매하는 고객 모두에게 크게 다르지 않지만, 온라인을 통해 구매하는 고객은 오프라인을 통한 구매 결정보다 정보 탐색 과정에서 검색, 조회 등의 행동으로 더 많은 정보를 찾는 경향이 있다[7]. 이와 같은 온라인 구매 환경을 반영한 마케팅 모델 AISAS는 고객의 구매 단계를 주의(Attention), 흥미(Interest), 검색(Search), 구매(Action), 공유(Share)의 5단계로 정의하였다[23].

그중 검색과 공유 단계는 오프라인에서 온라인으로 변화한 구매 환경에 따라 새롭게 등장한 단계로, 검색 단계는 고객이 스스로 관심이 있는 상품을 검색하여 정보를 수집하는 단계이며, 공유 단계는 고객이 구매한 상품에 대한 경험을 직접 공유하는 단계이다. 선행 연구에 의하면 이 두 단계에서 고객의 활동이 활발할수록 구매율이 높아지는 것으로 나타났으며, 이에 따라 마케팅의 관점도 검색과 공유 단계 중심으로 변화되고 있음을 알 수 있다[24]. 특히 검색 단계에서 고객은 상품 검색, 상품정보 조회 등의 행동을 하면서 구매 결정에 필요한 정보를 수집하고, 온라인 발자국을 남기게 되는데 이러한 행동 정보는 고객의 선호를 반영한다는 점에서 추천에 유용한 정보가 된다[25]. 이에 따라 본 연구는 고객의 구매 결정 과정 중 검색 단계에서 이루어지는 행동에 주목하여 고객의 온라인 행동 정보를 활용한 추천시스템을 제안하고자 한다.

현재까지 진행된 추천시스템 연구 중 비정형 데이터를 활용한 연구 대부분은 고객의 리뷰 데이터를 활용하여 진행되었다. 리뷰는 평점과 비슷한 성격의 데이터로, 고객의 평점에서 드러나지 않은 상세한 선호도를 파악할 수 있다는 점에서 많은 선행 연구에 사용되었다. 그러나 고객은 구매한 모든 상품에 대해 평점이나 리뷰를 남기지 않으며, 판매량이 적은 상품의 경우 리뷰 데이터가 부족하여 추천이 어려운 경우가 있어 콜드 스타트 문제가 발생할 수 있다[26].

그러나 고객의 온라인 행동 정보는 쇼핑몰에 방문한 고객의 행동에 따라 남겨진 데이터로써 평점 및 리뷰 데이터가 없는 고객에게도 추천이 가능하다. 특히 본 연구에서 주목하는 고객의 검색 행동은 고객이 흥미를 가진 상품의 정보를 직접 수집하는 행동으로, 고객은 검색 행동을 통해 검색 이력 데이터를 남기게 된다. 검색 이력 데이터는 분석을 통해 검색패턴을 파악할 수 있어 고객의 관심 상품 및 선호 상품 예측이 가능하다. 또한 고객은 상품의 상세정보를 조회하여 상품정보를 수집하는데, 이때 상세정보는 판매자가 제공하는 상품의 카탈로그로써 고객이 필요로 하는 상품의 정보가 상세히 포함되어 있다. 고객별 구매한 상품의 상세정보 데이터를 분석한다면, 각 고객이 선호하는 상품의 특성을 파악하여 고객의 선호를 예측할 수 있다.

그러나 선행 연구 중 고객의 온라인 행동 정보를 활용한 추천 연구는 많지 않으며, 정형 데이터인 고객의 세션 정보를 활용한 연구는 몇 차례 진행되었지만[27], 리뷰 데이터를 제외한 비정형 데이터를 활용한 추천 연구는 다소 부족한 실정이다. 이에 따라 본 연구는 고객의 선호를 예측할 수 있는 정보로 고객의 검색패턴과 고객이 구매한 상품의 상세정보를 활용하고자 한다.

III. The Proposed Scheme

1. Model Design

본 연구는 제안 모형과 함께 전통적 연구 모형을 구축하여 비교하고자 하였다. 본 연구의 전통적 연구 모형은 구매 이력 기반의 협업필터링 알고리즘을 적용한 추천 모형으로 설계하였으며, 제안 모형은 고객의 검색패턴 및 비정형 상품 상세정보를 기반한 추천 모형으로 설계하였다. 연구의 절차는 크게 7단계로 구성되고, 전체적인 흐름은 Fig. 1과 같다.

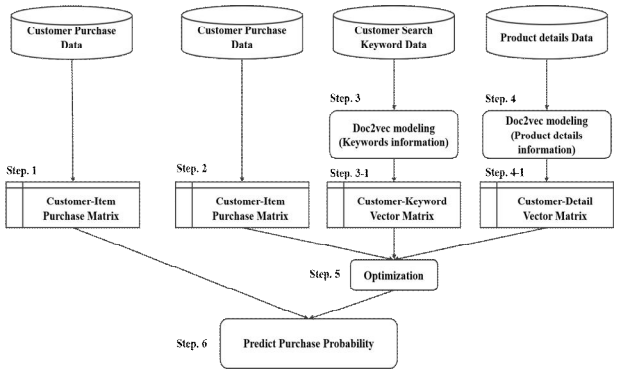


Fig. 1. Overview of the model process architecture

2. The Traditional Model

본 연구의 전통적 연구 모형에는 고객의 구매 이력 데이터가 사용되며, 가장 일반적인 협업필터링 방법론으로 알려진 사용자 기반 협업필터링 알고리즘으로 설계되었다.

(Step 1) 먼저, 전체 상품 중 가장 많이 판매된 상위 N개 품목을 선정하여, 이 개의 항목 중 중복 구매가 최소 m회 이상 이루어진 고객의 데이터를 사용한다. 그다음, 고객의 상품 구매 데이터를 이용하여 고객-상품 매트릭스를 구축하였다. 고객-상품 매트릭스는 고객이 상품을 구매하였는지, 구매하지 않았는지를 나타내는 이진 데이터로써 이진 데이터에 적합한 자카드(Jaccard) 유사도(1)를 활용한다. 자카드 유사도는 0과 1 사이의 값을 갖게 되며 1이면 두 고객 사이의 구매가 일치함을, 0이면 전혀 다름을 의미한다.

$$\mathrm{sim}(x,y) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}} \quad (1)$$

M_{11}:고객 x, y가 모두 구매한 경우
M_{10}:고객 x만 구매한 경우
M_{01}:고객 y만 구매한 경우
M_{00}:고객 x, y가 모두 구매하지 않은 경우

3. The Proposed Model

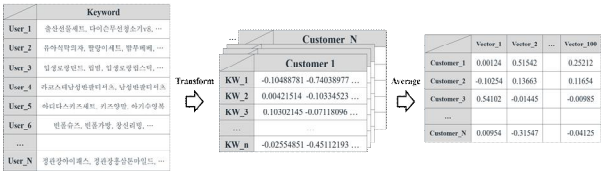
제안 모형은 고객의 구매 이력 데이터, 고객이 검색한 키워드 데이터, 고객이 구매한 상품의 상세정보 데이터를 모두 활용하여 각각의 고객-상품 매트릭스를 구축하는데, 이때 전통적 연구 모형과의 추천 성능 비교를 위해 같은 고객의 데이터를 사용한다.

(Step 2) 먼저 고객의 구매 이력 데이터를 기반한 고객-상품 매트릭스는 전통적 연구 모형과 같은 방법으로 구축한다.

(Step 3) 고객의 검색 키워드 데이터를 기반한 고객-키워드 매트릭스는 앞서 전통적 연구 모형에서 기술하였던 것과 같이 구매 이력 기반 매트릭스와 동일한 고객의 검색 키워드 데이터를 사용한다. 그다음 각 고객의 모든 검색 키워드 뭉치를 하나의 문서로 간주하여 Doc2vec을 적용한다[28].

Doc2vec은 document-to-vector의 줄임말로, 문장 내부의 단어 간 연관성을 파악하여 벡터로 변환하는 기법인 Word2vec의 확장된 방법으로, 단어와 함께 해당 문서의 위치 좌표를 학습하는 기법이다. Doc2vec은 Word2vec과 달리 단어가 아닌 문서를 비교하여 유사한 문서를 찾아 이를 고차원 공간에서 가까운 거리에 위치하게 한다[29-30]. 본 연구는 Doc2vec을 활용하여 고객의 검색 키워드 정보 및 구매 상품 상세정보를 기반한 고객별 위치정보를 구하고자 하였다.

(Step 4) 상품 상세정보 데이터를 기반한 고객-상세정보 매트릭스 역시 동일 고객이 구매한 상품의 상세정보 데이터를 사용하며, 검색 키워드 매트릭스 구축과 같은 방법으로 개별 고객이 구매한 모든 상품의 상세정보 뭉치를 하나의 문서로 간주하여 Doc2vec을 적용한다. 이는 Fig. 2를 통해 설명된다.



(a) Process of vectorizing search keyword data

(b) Process of vectorizing detailed information of products

Fig. 2. Vector representation of unstructured data using Doc2vec

Fig. 2의 (a)는 고객의 검색 키워드 데이터, (b)는 고객이 구매한 상품의 상세정보 데이터가 Doc2vec으로 전환되는 과정을 설명한 그림이다. 먼저 (a)는 고객마다 검색 키워드를 모두 합하여 데이터를 문서 형태로 전처리하고, 각 고객이 검색한 키워드 문서를 Doc2vec을 사용하여 벡터화한다. 그 후 만들어진 키워드 벡터를 평균하여 고객-키워드 매트릭스를 구축한다. 그다음 (b) 또한 마찬가지로, 각 고객이 구매한 상품의 상세정보 문서를 Doc2vec을 사용하여 벡터로 변환한 후 평균하여 고객-상세정보 매트릭스를 구축한다.

(Step 3-1, 4-1) 다음으로 Doc2vec을 이용하여 고객-키워드 매트릭스, 고객-상세정보 매트릭스를 구축하고, 이후 각 고객의 벡터값을 코사인(Cosine) 유사도(2)를 사용하여 전체 고객의 유사도를 구한다. 코사인 유사도는 벡터 간 코사인 각도 값을 이용하여 유사도를 측정하는 방법으로, 두 벡터의 방향이 같으면 1의 값을 가지며 1에 가까운 값일수록 유사도가 높다고 판단한다.

$$sim(x,y) = \frac{X \cdot Y}{\| X \| \cdot \| Y \|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \quad (2)$$

(Step 5) 앞에서 구축한 세 개의 유사도 매트릭스를 최적의 가중치로 결합하여 하나의 매트릭스로 구축한다. 이때 최적화의 대표적인 방법인 격자 탐색(Grid Search) 방법을 활용하여 가장 좋은 결합 비율을 찾고, 하나의 매트릭스로 결합한다.

(Step 6) , 전통적 연구 모형과 제안 모형에서 계산된 유사도를 기반하여 고객의 추천 상품 구매 확률을 각각 예측하고, 비교한다. 이때 상품 i에 대한 사용자 x의 구매 여부인 는 다음 산식(3)을 적용하여 예측한다.

$$\widehat{P_{x,i}} = \overline{P_x} + \frac{\sum_{y \in N} (P_{y,i} - \overline{P_y}) \cdot \quad S(x,y)}{\sum_{y \in N} |S(X,Y)|} \quad (3)$$

위 식에서 $\overline{P_x}$는 사용자 x의 평균 구매 예측 확률이고, S(x,y)는 추천 대상 사용자 x와 이웃 사용자 y사이의 유사도를 나타낸다. 그리고 N은 구매 집단을, y는 각각의 이웃을 나타내는 인덱스를 의미한다.

추천 성과 측정을 위한 지표는 재현율(Recall)과 정확률(Precision)을 동일한 가중치로 결합한 F1-measure(4)를 사용한다. 재현율(5)은 추천 대상인 고객이 실제 구매한 상품 중에서 추천 알고리즘에 의해 추천된 상품의 비율로 정의되며, 정확률(6)은 추천 알고리즘에 의해 추천된 상품 중에서 고객이 실제 구매한 상품의 비율로 정의된다.

$$F1 - measure = \frac{2 \times recall \times precision}{recall + precision} \quad (4)$$

$$recall = \frac{\text{고객이 실제 구매한 상품} \cap \text{추천한 상품}}{\text{고객이 실제 구매한 상품}} \quad (5)$$

$$precision = \frac{\text{고객이 실제 구매한 상품} \cap \text{추천한 상품}}{\text{추천한 상품}} \quad (6)$$

IV. Result

1. Data collection

본 연구는 롯데 멤버스, L.POINT 주최의 제5회 L.POINT Big Data Competition에서 제공된 데이터와 롯데쇼핑의 온라인 유통 채널에서 제공되는 상품 상세정보 데이터를 크롤링하여 사용하였다[31]. L.POINT Big Data Competition은 롯데그룹의 통합 멤버십 서비스인 L.POINT에서 제공하는 실제 데이터를 바탕으로 빅데이터를 분석하고 주제에 맞는 콘텐츠를 개발하는 국내 대표 빅데이터 공모전이다. 롯데 멤버스는 대한민국 국민의 60% 이상인 약 3,600만 명의 회원을 갖추었으며 롯데 그룹의 50여 개 그룹 및 외부 제휴사가 결합한 통합 멤버십 브랜드로써 많은 양의 라이프 스타일 데이터를 보유하고 있다. 본 연구에서는 그중 상품 구매 이력, 검색 이력 데이터를 활용하였다.

또한 고객이 구매한 상품의 상세정보 데이터를 수집하고자 롯데쇼핑의 온라인 유통 채널인 롯데온(Lotte On)에

서 고객 구매 이력 데이터를 기반으로 고객이 구매한 상품의 상세정보를 크롤링하였다. 상세정보 데이터는 상품 판매자가 올린 상품 카탈로그이며, 대부분 이미지 형식으로 이루어져 있다. 본 연구는 2020년 9월 22일부터 10월 5일까지 Python을 사용하여 총 2,716개 상품의 상세정보 이미지 35,299개를 수집하였다.

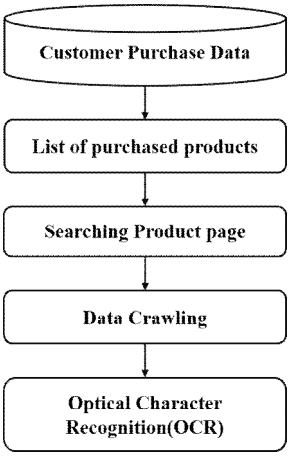


Fig. 3. Crawling Process

전체적인 크롤링 프로세스는 Fig. 3과 같다. 먼저 대표적 검색 엔진인 구글(Google)에서 고객이 구매한 상품의 상품명을 그대로 검색하되, 롯데온 사이트 혹은 롯데 계열사 사이트가 검색될 수 있도록 검색어와 관련이 있는 사이트 검색이 가능한 검색 연산자 'related:'를 사용하였다.

검색된 상품 페이지에 접속한 후 상품 상세정보 탭에 있는 상세정보 이미지 데이터를 크롤링하였으며, 수집된 데이터는 광학적 문자 인식(Optical Character Recognition, OCR) 기법으로 이미지 내 텍스트 데이터를 추출하여 사용하였다. 광학적 문자인식 기법은 컴퓨터가 광학적 메커니즘을 통해 인간과 같이 문자를 인식할 수 있는 기법으로, 이미지, PDF 파일 등 다양한 유형의 문서를 편집 및 검색 가능한 데이터로 변환할 수 있는 기술이다 [32]. 광학적 문자 인식 기법을 통해 추출된 상세정보 텍스트 데이터는 고객별로 취합하여 문서로 만들었다.

2. Result

본 연구가 제안하는 모형은 고객의 구매 이력 데이터뿐만 아니라 비정형 데이터인 고객의 검색패턴과 상품 상세정보를 활용한 추천 알고리즘으로, 구매 이력을 기반한 전통적인 추천시스템과 추천 성능을 비교하기 위하여 상위 N개 상품을 추천하고자 하였다. 또한 추천 성과 측정을 위해 F1-measure를 계산하여 비교하였다.

상위 N개 상품 추천 방법은 추천 대상 고객의 선호도가 가장 높을 것이라 예상되는 상품 N개를 추천하는 방식으로, 본 연구에서는 상위 3개, 5개, 7개를 선정하였으며, 각 추천 개수에 따른 F1-measure를 계산하였다. 실험 결과는 다음 Table 2와 같다.

Table 2. Comparison of F1-measures

Top N	Combination weights			F1-measure	
	Structured data	Unstructured data		Proposed model	Conv. model
	Purchase history	Search keywords	Product details		
Top 3	0.32	0.43	0.25	0.2098	0.1987
Top 5	0.11	0.04	0.85	0.2167	0.1791
Top 7	0.04	0.70	0.26	0.2013	0.1759

Table 2를 보면, 모든 경우에서 제안 모형의 성능이 전통적 연구 모형보다 좋음을 알 수 있다. 전통적 연구 모형과 비교하여 가장 성능이 개선된 경우는 Top 5로, 또한 모든 경우에서 가장 높은 성능을 보였다. 이때 결합한 매트릭스의 비율은 상세정보가 가장 높았으며, 비정형 데이터 기반 매트릭스의 비율이 89%를 차지하였다. 다음으로 성능이 개선된 Top 7의 경우, 결합한 매트릭스의 비율은 키워드가 가장 높았으며, 비정형 데이터 기반 매트릭스의 비율이 96%를 차지하였다. 마지막으로 Top3 의 경우에도 제안 모형이 전통적 연구 모형보다 높은 f1-measure값을 기록하였으며 추천 성능이 개선됨을 보였다. 이때 비정형 데이터 기반 매트릭스 비율은 68%로, 위의 결과와 함께 분석한 결과 제안 모형의 모든 경우에서 비정형 데이터 기반 매트릭스 비율이 정형 데이터 기반 매트릭스보다 상대적으로 높음을 알 수 있다.

이는 정형 데이터만 고려한 추천시스템보다 비정형 데이터를 함께 고려한 추천시스템의 성능이 모든 경우에서 향상되었음을 알 수 있다. 또한 추천 상품 수가 증가할수록 제안 모형의 매트릭스 결합 비율이 정형 데이터보다 비정형 데이터가 높은 것으로 보아, 더 많은 상품을 추천하기 위해선 구매 이력 정보보다는 고객의 온라인 행동 정보가 더 유익함을 알 수 있었다. 그런데 Top 5, Top 7의 경우 각각의 키워드 정보와 상품 상세정보가 서로 상호보완적이지 못한 관계를 보여 키워드 정보와 상세정보는 각각 배타적으로 유익한 영향을 미침을 알 수 있다.

Ⅴ. Conclusions

본 연구는 고객의 구매 이력만을 고려한 추천시스템의 성능 한계를 해결하기 위해, 고객의 검색패턴 및 비정형 상품 상세정보를 활용한 추천시스템을 제안하였다. 추천시스템의 성능을 개선하기 위한 연구는 활발히 진행되고 있지만 대부분 새로운 알고리즘 개발과 관련한 연구이거나, 구매 여부, 평점과 같은 정형 데이터만을 고려한 연구가 많았다. 그러나 최근 몇몇 선행 연구로부터 평점 데이터의 고객 선호도 설명력에 대한 의문이 제기되었고, 리뷰 데이터로 인한 희소성(Sparsity)이 문제가 되면서 이러한 한계를 보완하기 위해 더욱 다양한 비정형 데이터를 활용한 추천 연구가 필요할 것으로 보인다.

이에 본 연구는 고객의 온라인 행동에 주목하여 온라인 행동 정보를 활용하고자 하였다. 구체적으로, 본 연구의 제안모델은 고객의 구매 이력 데이터와 함께 비정형 데이터인 고객 검색 키워드 데이터, 상품 상세정보 데이터를 결합하여 고객의 정형·비정형 데이터를 함께 고려한 추천 시스템을 제안하였다. 결과적으로, 제안 모형은 전통적 추천 모형보다 높은 추천 성능을 보였으며 추천 상품의 수가 증가할수록 구매 이력 데이터보다 고객의 온라인 행동 정보가 더 유익한 영향을 미치는 것으로 확인되었다.

본 연구가 갖는 의의는 다음과 같다. 본 연구는 고객의 비정형 데이터인 온라인 행동 정보를 활용하여 추천시스템 모형에 적용하였다는 점에서 의의가 있다. 그간 추천시스템 연구 중 성능 개선을 위해 새로운 데이터를 활용한 경우는 대부분 평점이나 구매 여부와 같은 정형 데이터를 사용하였다. 최근 비정형 데이터를 활용한 추천시스템 연구가 활발하게 진행되고 있지만, 이러한 경우도 대다수가 리뷰 데이터를 사용한 연구였다. 온라인 쇼핑몰의 경우 고객의 리뷰 데이터 외에 고객의 구매 활동과 관련된 여러 비정형 데이터가 존재하지만 이러한 데이터에 주목하여 진행된 연구는 많지 않았다. 그러나 본 연구는 고객의 온라인 행동 정보에 주목하고 고객의 검색 키워드 이력과 고객이 구매한 상품의 상세정보를 Doc2vec 기법을 사용하여 추천 모형에 활용함으로써 기존 연구와 차별성을 갖고, 학술적인 의의를 지닌다고 할 수 있다.

또한, 본 연구는 제안 모형을 통해 전통적인 협업필터링의 한계를 극복하였다. 일반적으로 협업필터링은 구매 이력이나 평점 등의 데이터를 기반으로 추천하기 때문에, 구매하지 않았거나 구매 기록이 적은 신규 고객과 구매 상품에 대해 평점 및 리뷰를 남기지 않은 고객에게는 추천이 어렵다는 한계가 있다. 그러나 본 연구에서 활용하는 데이터는 고객의 검색 키워드와 고객이 구매한 상품의 상세정보로, 온라인 쇼핑몰에서 구매 단계로 이어지지 않고 검색 단계까지 행동한 고객의 경우 해당 고객의 검색 이력만으로도 상품 추천이 가능하다는 장점이 있다. 또한 온라인 쇼핑몰 이용 고객 중 검색 단계 없이 바로 구매 단계로 이어진 고객에게는 구매한 상품의 상세정보를 통해 상품의 특성을 추출하여 추천시스템에 반영이 가능하다.

반면 본 연구는 다음과 같은 한계점을 가진다. 본 연구가 사용한 데이터는 고객 간 중복 거래량이 많지 않았으며, 2018년도에 공개된 데이터를 사용하였기 때문에 상품 상세정보 수집 때 판매가 종료된 상품은 상세 페이지가 사라져 데이터 수집이 어려웠다. 연구에 사용된 데이터는 이러한 상품은 제외하고 수집하였기에 향후 연구에서는 더욱 확장된 크기의 최신 데이터를 사용하여야 할 것이다. 또한 본 연구의 결과에서 상품 5개를 추천한 Top 5, 7개를 추천한 Top 7의 경우 결합한 매트릭스의 비정형 데이터 기반 매트릭스의 비율은 높으나 Top 5의 경우 상세정보 기반 매트릭스가, Top 7의 경우 키워드 기반 매트릭스의 비율이 훨씬 높게 반영되었다. 이러한 점으로 보아 두 데이터가 서로 배타적임을 알 수 있었는데, 향후 연구에서는 이러한 점에 주목하여 온라인 쇼핑몰의 여러 비정형 데이터를 활용한 확장 연구를 진행할 필요가 있다.

REFERENCES

[1] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative filtering recommender systems," Foundations and Trends® in Human–Computer Interaction, Vol. 4, No. 2, pp. 81-173, May 2011. https://doi.org/10.1561/1100000009

[2] M. G. Kim, and K. J. Kim, "Recommender Systems using Structural Hole and Collaborative Filtering," Journal of Intelligence and Information Systems, Vol. 20, No. 4, pp. 107-120, Dec. 2014. http://dx.doi.org/10.13088/jiis.2014.20.4.107

[3] J. S. Lee, J. Y. Kim, and B. W. Kang, "A Study on Improvement of Collaborative Filtering Based on Implicit User Feedback Using RFM Multidimensional Analysis," Journal of Intelligence and Information Systems, Vol. 25, No. 1, pp. 139-161, Mar. 2019. http://dx.doi.org/10.13088/jiis.2019.25.1.139

[4] J. Y. Hyun, S. Y. Ryu, and S. Y. Lee, "How to improve the accuracy of recommendation systems : Combining ratings and review texts sentiment scores," Journal of Intelligence and Information Systems, Vol. 25, No. 1, pp. 219-239, Mar. 2019. http://dx.doi.org/10.13088/jiis.2019.25.1.219

[5] M. Srifi, A. Oussous, A. A. Lahcen, and S. Mouline,

"Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey," Information, Vol. 11, No. 6, id.317, Jun. 2020. https://doi.org/10.3390/info11060317

[6] S. Y. Yun, and S. D, Yoon, "Item-Based Collaborative Filtering Recommendation Technique Using Product Review Sentiment Analysis," Journal of the Korea Institute of Information and Communication Engineering, Vol. 24, No. 8, pp. 970-977, Oct. 2020. http://doi.org/10.6109/jkiice.2020.24.8.970

[7] C. Katawetawaraks, and C. Wang, "Online shopper behavior: Influences of online shopping decision," Asian Journal of Business Research, Vol. 1, No. 2, Oct. 2011. http://doi.org/10.14707/ajbr.110012

[8] C. Wu, and M. Yan, "Session-aware information embedding for e-commerce product recommendation," In Proceedings of the 2017 ACM on conference on information and knowledge management, pp. 2379-2382, Nov. 2017. https://doi.org/10.1145/3132847.3133163

[9] J. K. Kim, D. H. Ahn, and Y. H. Cho, "A Presonalized Recommender System, WebCF-PT: A Collaborative Filtering using Web Mining and Product Taxonomy," Asia Pacific Journal of Information Systems, Vol. 15, No. 1, pp. 63-79, May 2005. G704-000077.2005.15.1.001

[10] H. C. Ahn, "Improvement of a Context-aware Recommender System through Users Emotional State Prediction," Journal of Information Technology Applications & Management, Vol. 21, No. 4, pp. 203-223, Dec. 2014. G704-000870.2014.21.4.012

[11] M. J. Pazzani, and D. Billsus, "Content-based recommendation systems," In The adaptive web, pp. 325-341, Springer, Berlin, Heidelberg, 2007. http://doi.org/10.1.1.130.8327

[12] M. J. Ku, and H. C. Ahn, "A Hybrid Recommender System based on Collaborative Filtering with Selective Use of Overall and Multicriteria Ratings," Journal of Intelligence and Information Systems, Vol. 24, No. 2, pp. 85-109, Jun. 2018. http://dx.doi.org/10.13088/jiis.2018.24.2.085

[13] J. H. Kim, H. J. Jo, and B. M. Kim. "Game Recommendation System Based on User Ratings," Journal of the Korea Industrial Information Systems Research, Vol. 23, No. 6, pp. 9-19, Dec. 2018. http://dx.doi.org/10.9723/jksiis.2018.23.6.009

[14] G. Guo, J. Zhang, D. Thalmann, and N. Yorke-Smith, "Leveraging prior ratings for recommender systems in e-commerce," Electronic Commerce Research and Applications, Vol. 13, No. 6, pp. 440-455, Nov. 2014. http://doi.org/10.1016/j.elerap.2014.10.003

[15] J. E. Son, S. B. Kim, H. J. Kim, and S. Z. Cho, "Review and Analysis of Recommender Systems," Journal of the Korean Institute of Industrial Engineers, Vol. 41, No. 2, pp. 185-208, Apr. 2015. http://dx.doi.org/10.7232/JKIIE.2015.41.2.185

[16] X. Su, and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in artificial intelligence, Oct. 2009. http://doi.org/10.1155/2009/421425

[17] S. I. Choi, Y. J. Hyun, and N. G. Kim, "Improving Performance of Recommendation Systems Using Topic Modeling," Journal of Intelligence and Information Systems, Vol. 21, No. 3, pp. 101-116, Sep. 2015. http://dx.doi.org/10.13088/jiis.2015.21.3.101

[18] J. H. Su, W. Y. Chang, and V. S. Tseng, "Effective social content-based collaborative filtering for music recommendation," Intelligent Data Analysis, Vol. 21, No. S1, S195-S216, Feb. 2017 https://doi.org/10.3233/IDA-170878

[19] Y. Ma, G. Chen, and Q. Wei, "Finding users preferences from large-scale online reviews for personalized recommendation," Electronic Commerce Research, Vol. 17, No. 1, pp. 3-29, Mar. 2017. https://doi.org/10.1007/s10660-016-9240-9

[20] S. Y. Cho, J. E. Choi, K. H. Lee, and H. W. Kim, "An Online Review Mining Approach to a Recommendation System," Information Systems Review, Vol. 17, No. 3, pp. 95-111, Dec. 2015. http://dx.doi.org/10.14329/isr.2015.17.3.095

[21] S. E. Cho, and H. S. Lim, "A Study on Product Recommendation System Based on User Search keyword," Journal of Digital Contents Society, Vol. 20, No. 2, pp. 315-320, Feb. 2019.. http://dx.doi.org/10.9728/dcs.2019.20.2.315

[22] R. K. Lee, N. H. Chung, and T. H. Hong, "Developing the online reviews based recommender models for multi-attributes using deep learning," The Journal of Information Systems, Vol. 28, No. 1, pp. 97-114, Mar. 2019. http://dx.doi.org/10.5859/KAIS.2019.28.1.97

[23] H. G. Chae, S. U. Park, and J. Y. Kang, "Applying RSS Marketing on Internet Shopping Malls Based on AISAS Model," The Journal of Society for e-Business Studies, Vol. 13, No. 3, pp. 21-49, Oct. 2008. G704-000612.2008.13.3.007

[24] H. K. Koo, "Purchasing Behavior of AISAS Process According to the Generations for Fashion Products," Korea Science & Art Forum, Vol. 9, pp. 15-26, Dec. 2011. G704-SER000015040.2011.9..008

[25] C. Wu, and M. Yan, "Session-aware information embedding for e-commerce product recommendation," In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2379-2382, Nov. 2017. https://doi.org/10.1145/3132847.3133163

[26] J. W. Jeong, W. S. Hwang, H. J. Lee, and S. W. Kim, "Recommendation Exploiting Search-Keywords in Online Shopping," KIISE a collection of academic presentation papers 39(2C), pp. 95-97, Korea, Nov. 2012. I410-ECN-0101-2014-569-0007478306

[27] J. Ding, G. Yu, Y. Li, X. He, and D. Jin, "Improving Implicit Recommender Systems with Auxiliary Data," ACM Transactions on Information Systems (TOIS), Vol. 38, No. 1, pp. 1-27, Feb. 2020. http://doi.org/10.1145/3372338

[28] V. T. Phi, L. Chen, and Y. Hirate, "Distributed representation based recommender systems in e-commerce," *In DEIM Forum*,

2016.

[29] Q. Le, and T. Mikolov, "Distributed representations of sentences and documents," In International Conference on Machine Learning, pp. 1188-1196, Jan. 2014.

[30] 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, Vol. 26, pp. 3111-3119, Dec. 2013. https://dl.acm.org/doi/10.5555/2999792.2999959

[31] Lotte Members, L.pay|L.POINT, 제5회 L.POINT Big Data Competition

[32] R. Mithe, S. Indalkar, and N. Divekar, "Optical character recognition," International Journal of Recent Technology and Engineering (IJRTE), Vol. 2, No. 1, pp. 72-75, May 2013. http://doi.org/10.3403/00116062

Authors

Yunju Lee received the B.A. degree from Kangwon National University, Korea, in 2019. She is currently pursuing the M.S. degree in Business IT Graduate School, Kookmin University, Korea.

Yunju Lee is interested in business analytics, deep learning and text mining.

Jaejun Lee received the B.S. degrees in Mathematics from Dongguk University, Korea, in 2020. He is currently a M.S. candidate in the Graduate School of Business IT at Kookmin University, Korea.

Jaejun Lee is interested in business analytics, data mining and credit assessment.

Hyunchul Ahn received the B.E. in Industrial Management, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 1999, 2002 and 2006, respectively. Dr. Ahn joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2009. He is currently a professor in the Graduate School of Business IT at Kookmin University. He is interested in intelligent IS and IS adoption.