특집 06

# 빅데이터 기반 CBF[1)]를 이용한 상품추천시스템 개발 및 챗봇 서비스 적용

채보병 · 유지호 · 유승호 ((주)유스바이오글로벌)

목 차



## 1. 서 론

코로나로 인한 '언택트'(Untact)시대가 되면서, 온라인 쇼핑몰 및 비대면 서비스를 제공하는 다양한 플랫폼들을 이용하여, 빅데이터 기반의 빠르고 효과적인 주문 및 배송방법에 대한 경쟁이 불붙고 있다[1]. 특히, 2021년 통계청의 자료에 따르면, 첫째. '온라인 쇼핑 거래액 26.4% 상승', 둘째, '모바일 쇼핑의 거래액은 29.4% 상승'으로 변화를 읽을 수 있고, '쿠팡', '네이버' 등의 배송 속도전에 'GS 리테일'까지 참여를 알리면서, 대기업 간의 배송 경쟁이 가속화 되고있는 상황이다[2][3]. 다만, 이러한 '온라인 소비 가속화' 및 '빅데이터 기반 배송전쟁'에 따른 거대 플랫폼과 공룡유통망의 결합 및 속도경쟁이 영세한 업체들에게 어려움이 될지, 기회가 될지는 아직은 미지수라고 할 수 있다[4]. 즉, 인지도가 없는 '스타트업'이나 '영세한 판매자'의 경우 이미 잘 알려진 쇼핑 플랫폼 내에 자사의 제품을 상위 카테고리에 노출시켜 소비자에게 판매를 유도하기는 어렵다. 이와 같은 이유로는 거대 플랫폼 등도 '광고비', '마케팅비'에 따라서 차등적으로 제품의 노출순위를 정하거나 추천하는 서비스 알고리즘을 도입하고 운영하고 있기 때문이다.

특히, 오프라인에서 인공지능(AI)과 빅데이터 기술을 활용해 상품들의 판매 시기, 시간, 지역 등을 분석 및 예측하여 소비자의 요구에 맞게 더 발빠르게 대처하도록 하는 각종 데이터 정보요소가 기업의 경쟁력이 되었다. 이에 발맞춰서 본고에서는 온라인에서 쇼핑채널의 형태로 소비자의 행동패턴, 상품의 선호도, 상품을 클릭한 횟수 등 다양한 데이터를 분석 및 활용하여 사용자 기반분석의 상품 추천 서비스를 제공할 수 있도록 연구하였다[5]. 이에 대한 적용 플랫폼으로는 국내에서 가장 많은 사람이 이용하는 카카오톡 기반 플랫폼이며, '챗봇 프로토타입' 안에서 소비자에게 '상품추천'을 하고, 판매자도 쉽게 서비스를 이용하는 방법

1) CBF(Contets-based-filtering) : 콘텐츠 기반 필터링(내용기반 필터링)이라 하며 아이템의 정보나 특성을 이용해 사용자의 요구 정보를 기반으로 문서를 추천하는 방법

을 제안하고자 한다.

## 2. 본 론

현재, '상품추천 서비스'는 '음악', '도서', '영상', '영화', '뉴스' 등 플랫폼 내 다양한 카테고리 분야에서 활용되고 있을 뿐만 아니라, 국외에서는 '아마존', '구글' 등이 대표적이고, 국내에서는 '네이버', '다음', '쿠팡' 등 쇼핑과 관련된 대부분의 플랫폼에서 사용하고 있는 서비스 알고리즘이다[6]. 다만, 국내에서는 플랫폼별로 조금씩 서로 다른 이름으로 불리워지고 있고, 현재의 '카카오톡 서비스'내에서는 먼저, '채널'을 추가할 경우에 누구나가 '알림톡' 형태로 광고 메시지를 발송할 수 있고, '카카오'에서도 '쇼핑 채널', '라이브 광고 채널' 등으로 쇼핑 채널의 형태에 따라서 '추천서비스'가 운영되고 있다. 특히, 후자인 '플랫폼사 주도'의 '탑다운 형식의 추천서비스'는 첫째, '할인경쟁 및 쿠폰발급'을 통한 마켓팅 형태, 둘째, '유통공룡업체나 인플루언서를 이용한 고객유치' 등으로 크게 요약될 수 있다. 즉, '채팅을 통한 의사소통 기반'이 바로 '카카오톡 서비스 특성'임에도, 등록한 소비자의 대화정보 등을 바탕으로 한, '선호상품을 추천'할 수 없어서, '추천 광고 메시지'가 어느 한순간 '스팸 메시지'가 되기도 한다.[7] 때문에, 이를 예방하면서도 소비자가 중심인 채널을 만들기 위해서는, 채널 이용자의 ①'연령', ②'채널 성격', ③채널내의 '이용자 패턴 정보' 등을 통한 맞춤형 정보 추천 등을 통한 '상품추천의 최적화'와 '서비스 이용 형태의 대화형식 변경'이 필요하다. 이에 실제로 '추천서비스'와 '카카오톡 챗봇'을 결합하여 챗봇 채널을 추가한 소비자에게 '추천서비스'를 제공할 수 있도록, 또한 카카오톡 채널에 등록된 사람에게 추천리스트를 '카카오i 오픈빌더'기반의 '챗봇' 내에서 정보를 보여주는 방식인 카드(케러셀)형식으로 보이도록 파일럿 서비스를 개발하여, 2.3.에서 '서비스 연동'과 '추천서비스 알고리즘 적용결과'에 대하여 함께 제시하였다.

### 2.1 추천시스템 알고리즘 동향

'추천 알고리즘'이 활발히 연구되면서 기술적으로 다양한 알고리즘들이 개발, 연구가 활발하게 진행되면서, 아래의 (그림 1)과 같이 '추천서비스'의 종류는 매우 다양하며, 여러 가지의 종류 중 일부를 수정하여 재구조화하여 도식화였다[8]. 해당 알고리즘 중에서는 상품정보만을 이용하는 '콘텐츠 기반 필터링'을 이용하였으며, 개념과 특징을 간단히 요약하여 동향을 최대한 알기 쉽도록 정리하였다.


(그림 1) 대표적인 추천시스템 알고리즘의 종류

#### 2.1.1 콘텐츠 기반 필터링(Contents Based Filtering, CBF)

콘텐츠 기반 필터링(이하 내용 기반 필터링)은 상품의 정보를 분석하여 아이템과 아이템 간, 아이템과 소비자 간 유사성 및 선호도를 분석하여, 소비자에게 유사한 상품을 추천을 해주는 방식이다[8]. 상품의 상세 정보, 타 소비자가 제공한 평가 상품 평가 점수 등 상품의 주요 정보를 토대로 상

품을 파악해서 유사도를 구하고, 소비자가 선호하는 상품 선택 시 가장 유사한 아이템을 추천해주는 방식이고, 소비자의 구매정보를 수집한다면 소비자의 프로파일 정보를 토대로 이전 구매한 상품과 유사한 상품을 추천한다.[13]

#### 2.2.2 CBF의 특징

콘텐츠 기반 필터링은 상품의 주요 정보 혹은 소비자의 과거 구매 이력과 같은 프로파일 정보를 활용하여 상품 및 아이템을 추천하고, 다른 소비자의 정보가 부족할 경우 이용하기 용이하다. 내용 기반 필터링은 소비자 개인의 평가에 기반을 둬서 (그림 1)에 협업 필터링으로 구성 시 발생하는 상품의 평가가 없을 때 추천리스트에 들지 못하는 first rater 문제가 발생하지 않고, 고객의 프로파일과 상품의 정보(내용)를 기반으로 유사도를 구해 소비자에게 추천하기 때문에 왜 해당 상품이 추천리스트에 추가되었는지 설명하기 용이하다.

반면, 상품별 특징(feature)을 선택할 수 없는 데이터(노래, 이미지 등)은 상품별 특징을 골라내기가 굉장히 어렵고, 고객의 과거 구매 이력이나 고객이 선호하는 상품과 유사한 데이터들만 추천리스트에 추가되어 시간에 따라 변화하는 고객의 취향을 반영할 수 없다. 즉, 추천리스트에 다양한 상품들이 포함되지 못하는 과도한 특수화(Over Specialization) 문제점이 발생한다.

### 2.2 추천서비스 개발과 적용

본 장은 추천서비스의 구현에 대해 데이터 수집부터 추천리스트 구현에 관한 내용과 흐름도를 위의 (그림 2)처럼 서술한다. 상품 데이터는 출산/육아 상품 중 8개 카테고리를 대상으로 데이터를 수집하였고, 상품의 특징을 수집해 상품들의 텍스트 기반 문서 유사도를 구하는 cosine similarity를 구현해서 소비자가 선호하는 상품과 유사한 상품들이 추천될 수 있도록 했다. 전체적인 서비스 구현 흐름도는 아래의 (그림 3)처럼 고객의 취향을 파악해서 알려주는 상품 추천서비스와 온라인 마케팅이 동시에 가능한 카카오톡 기반의 챗봇 서비스를 적용해서 카카오톡 내에서도 다양한 활용이 가능하다는 점을 제시한다.

(그림 2) 추천서비스 구현 흐름도

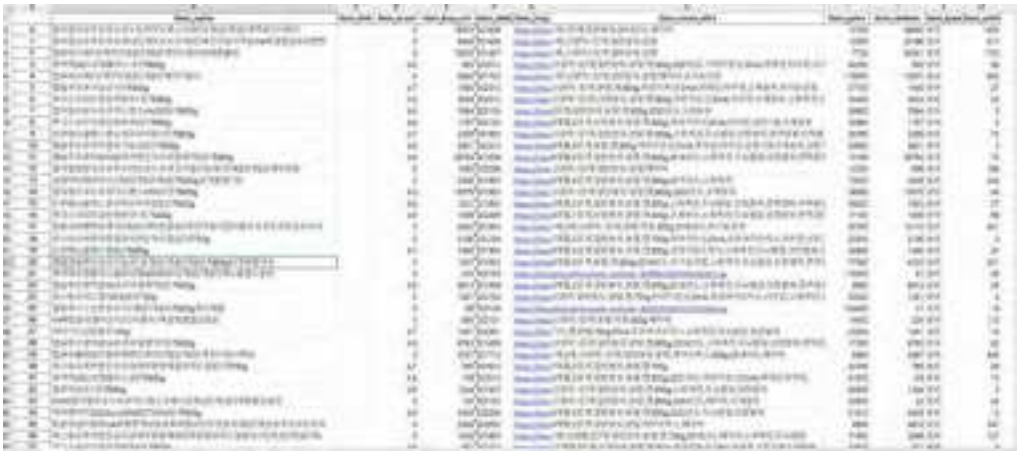

(그림 3) 챗봇 서비스 전체 흐름도

### 2.2.1 상품 데이터 수집

출산/육아 상품 8가지 범주 데이터를 Python의 Beautiful Soup, Selenium 패키지를 활용하여 '네이버 쇼핑'에서 웹 스크래핑을 진행해 총 13,840건을 수집하였다. 수집한 상품 특징은 왼쪽부터 [상품 이름, 링크, 구매횟수, 이미지, 상세특성, 가격, 리뷰 횟수, 상품 타입, 찜 횟수] 등을 수집하고, (그림 4-2)처럼 상품 특성을 활용할 수 있도록 전처리를 진행하였으며, 상품의 평점과 구매횟수, 리뷰 모두 없는 상품은 제외하여 (그림 4-2)처럼 총 9,623건의 데이터를 수집하였다.

### 2.2.2 상품 특성 벡터화

수집한 데이터의 특성을 벡터화 후에 문서 유사도를 구해야 하는데, 그 벡터화 과정에서 TF-IDF 방법을 이용한다[8].



(그림 4-1) 네이버 쇼핑 웹 스크래핑과 전처리 한 데이터 중 일부

(그림 4-2) 전처리 후 데이터 총 개수

왼쪽부터 [상품 이름, 링크, 평점, 구매횟수, 출시일, 이미지, 상세특성, 가격, 리뷰 횟수, 상품 타입, 찜 횟수] 순서

#### 2.2.2.1 TF-IDF

TF-IDF란 특정단어가 문서 내에서 등장하는 빈도(TF)와 그 단어가 문서 전체 집합에서 등장하는 빈도(IDF: 역문서 빈도)를 고려하여 벡터화를 하는 방법이다. 하나의 문서 단위로 벡터를 만드는데, 식 (1)과 같이 각 인덱스에 해당하는 단어가 문서에 등장하는 빈도와 문서 집합 전체에 등장하는 단어 빈도의 역수를 곱하여 구하게 된다.[9] TF는 Term Frequency의 약자로 문장 document(d) 안에서 특정 term(t)이 몇 번 등장 했는지를 나타낸다. 예를 들어 이유식이라는 단어가 4번 등장했다면 document의 TF 값은 4가 된다.

$$tf(t,d) = f_{t,d} \quad (1)$$

여기서 $f_{t,d}$ 값은 위에서 말한 문서 안에서 term이 나온 횟수 자체이다. 단, 때에 따라 tf 값을 나온 수치 그대로 사용하는 것이 아니라 boolean 빈도, 로그 스케일 빈도, 증가 빈도 등 변형해서 사용하기도 하며, 본 장에서는 기본적인 문서 내 단어 빈도를 구하였다.

IDF는 Inverse Document Frequency의 약자로 Document Frequency 값을 역수로 만든 값을 의미한다. 식 (2) DF는 한 문서만 고려하는 값이 아닌 여러 개의 문서가 있을 때 어떤 특정한 단어(term)가 얼마나 많은 문서에 등장하는지 확인할 수 있는 값이 된다.

$$df(t,D) = \frac{|\{d \in D : t \in d\}|}{|D|} \quad (2)$$

d는 개별 문서를 뜻하며 D는 전체 문서 집합을 뜻한다. df 값은 특정단어(t)가 들어가 있는 문서 수를 전체 문서 수로(D) 나눠준 값이다.[10]

$$idf(t,D) = \frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

idf(t,D) 값은 단순히 이 값을 역수로 만든 값이다. 역수로 만든 idf 값의 의미는 해당 문서에 등장한 특정단어가 타 문서에서는 잘 나오지 않은 단어인지 측정하는 척도이다. 식(3)

DF의 역수를 취하게 될 때 전체 문서의 수가 많으면 많을수록 idf의 값이 기하급수적으로 커져 보통 로그를 취해주고, 특정단어가 전체의 문서에서 존재하지 않으면 분모가 0되는 상황을 방지하고자 분모에 1을 더해주어 결과적으로 idf는 다음의 식(4)와 같은 식이 된다.

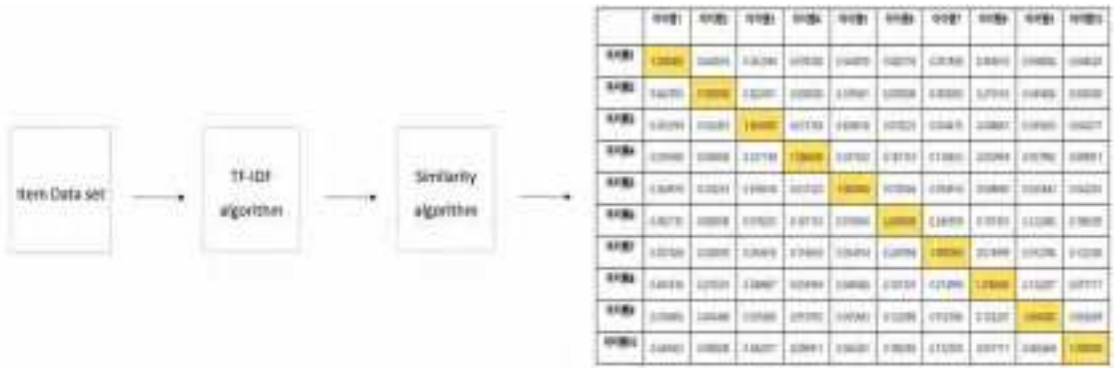
$$idf(t,D) = \log\left(\frac{|D|}{|1 + \{d \in D : t \in d\}|}\right) \quad (4)$$

최종적으로 식(5)와 같이 TF-IDF는 이런 의미를 나타낸다. 해당 문서에서 특정단어가 나온 횟수 값과 그리고 이 단어가 다른 문서에는 잘 나오지 않는 그런 단어인지를 같이 고려하여 해당 값의 계산은 다음과 같이 두 값을 곱해주기만 하면 된다. TF-IDF 값이 크면 '특정단어'가 '특정 문서'에서만 출현하는 빈도가 높다는 뜻이며, TF-IDF 값이 낮으면 '특정단어'는 여러 문서에서 여러 번 나타나거나 아니면, 특정 문서에 출현 빈도가 낮다고 볼 수 있다[8][9].

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (5)$$

#### 2.2.2.2 문서 유사도 계산

벡터화한 아이템-키워드 행렬을 통해 식(6)과 같이 아이템 간의 유사도를 계산한다. 사용할 수 있는 수식은 여러 가지 있으나, 주로 사용되고 있는 코사인 유사도(Cosine similarity)를 이용한




(그림 5) 코사인 유사도를 적용한 행렬 도식화

다.[10][12]

(그림 5)는 기존 아이템-키워드 행렬에서 코사인 유사도 적용을 도식화하였다. 예시로 사용자 A가 '아이템 1'을 선호하면 '아이템 1'과 가장 비슷한 '아이템 2'을 추천한다.

### 2.2.3 가중평점 계산

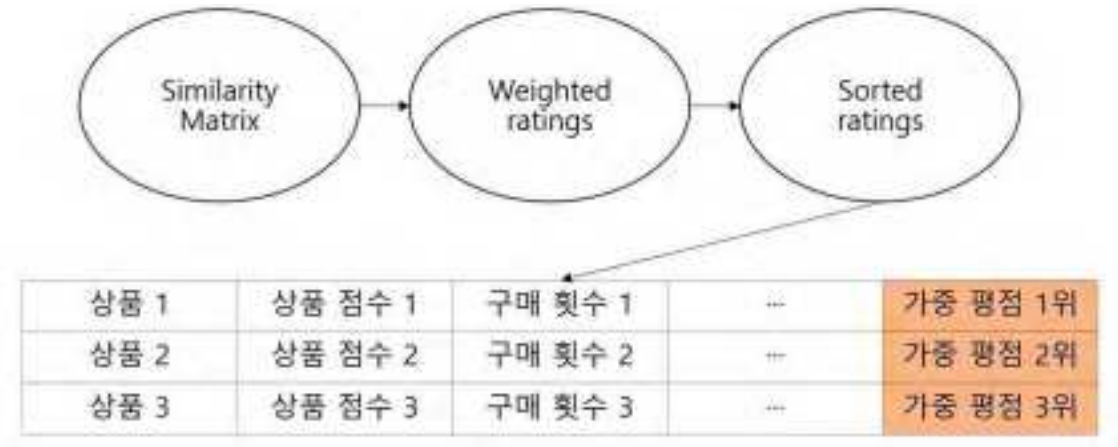
코사인 유사도를 적용 후 그대로 추천하게 된다면, 소비자와 취향이 동일한 아이템을 추천해줄 수 있으나, 소비자가 스릴러 장르의 영화를 좋아한다고 해서 스릴러 장르의 모든 영화를 좋아하지는 않듯, 소비자의 취향을 좀 더 반영하기 위해 아이템의 평점, 리뷰 개수 등을 이용하여 평점에 가중평점을 식(7)과 같이 계산하였다.

(그림 6)처럼 가중평점을 기준으로 상품을 추천하되, 구매횟수에 Min-Max Scaler를 적용하여 구매횟수가 가장 높은 상품에 가중평점 + 1, 가장 낮은 상품에 가중평점 + 0을 추가하였다[11]. 상품 평점은 좋지만, 구매횟수나 상품 리뷰 수가 적은 상품이 추천리스트에 들어가는 것을 방지하고자 구매횟수에 해당 수식을 식(8)과 같이 적용하여 가산하였다.

$$Similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (6)$$

$$\text{가중평점}\,(Weighted\ Rating) = \left(\frac{v}{(v+m)}\right) * R + \left(\frac{m}{(v+m)}\right) * C \quad (7)$$

v : 개별 영화에 평점을 투표한 횟수
m : 평점을 부여하기 위한 최소 투표 횟수
R : 개별 영화에 대한 평균 평점
C : 전체 영화에 대한 평균 평점




(그림 6) 가중평점 및 구매횟수를 이용해 상품별 점수를 계산

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (8)$$

## 2.3 결과

내용 기반 추천서비스를 구현과 사용되는 알고리즘에 대해 서술하였다. (그림 7)처럼 카카오 챗봇 서비스와 추천시스템을 연동하였는데, 사용자가 챗봇 시작하기를 입력하고 시작하기 버튼을 클릭 시 아이템 추천하기 위한 카테고리 3개를 설정하도록 요청한다. 소비자가 카테고리를 모두 입력하면 3개 카테고리에 대한 가중평점을 계산한 후 10개 상품을 추천하여, 내용 기반 필터링으로 추천서비스를 제공하는데, 타 소비자의 데이터를 수집해 소비패턴, 선호 상품 등을 학습시킬 수 있다면 개개인에게 맞는 상품을 더 집중적으로 추천할 수 있을 것이다. 챗봇의 특성상 소비자가 발화문(텍스트)을 입력했을 때 추천하는 리스트를 보여주거나 상품의 특징 및 브랜드 정보를 이용해서 사용자가 원하는 상품의 특징이나 이름을 검색하면 유사한 상품 중 가장 평이 좋은 상품을 추천해주는 등 기능 구현을 상대적으로 쉽고 빠르게 구현 가능하다는 장점이 있고, 소비자가 카카오톡 채널 이용법만 알고 있다면 적은 비용으로 높은 효율의 마케팅 및 서비스를 제공할 수 있어 확장성에 용이하다.

(그림 7) 추천서비스와 챗봇 연동 예시

## 3. 결 론

이로써, '내용 기반 필터링'(CBF) 기법을 이용한 챗봇 내에서 추천시스템 프로토타입을 구현하였다. 쿠팡, 아마존 등 대형 플랫폼들은 협업 필터링이나 '내용 기반 필터링'(CBF)을 조합 하는 등 2가지 이상 기법을 이용해 성능이 향상된 상품서

비스를 제공해주고 영세 플랫폼은 마케팅에 신경을 써도 인지도가 상대적으로 부족하여 다수의 소비자에게 서비스를 제공하기 어렵다. 따라서 본 논문에서 제안하고자 하는 바는 새로 플랫폼을 구축하지 않고 기능 추가 및 유지보수에 대한 편의, 서비스 구현 비용 절감, 응용 방안 확대 가능, 접근성 확보와 같은 장점이 있는 카카오톡을 이용하여 카카오톡 채널 내 사용 목적이 비슷한 소비자들에게 서비스를 제공하자는 것이다.

현재 사용되고 있는 온라인 쇼핑몰의 마케팅 챗봇 채널들은 대화할 수 없는 광고 채널이고 마케팅하는 상품마저 소비자 개인의 맞춰지는 것이 아닌 다수의 이용자에게 발송되기만 한다. 소비자와 챗봇의 대화를 통해 사용자가 선호하거나 비선호하는 상품들을 알아내고 소비자가 추천서비스를 이용하거나 사용자의 수집된 이력을 바탕으로 추천하는 물품이 있다면 대화형 서비스와 마케팅을 동시에 이용하는 방법을 개발하였다.

다만, 적용한 추천 알고리즘 특성상 서비스의 단점도 명확하다. 첫 번째는 내용 기반 필터링의 특징으로 상품의 특성을 이용하여 소비자가 선호하는 상품 중 특성이 비슷한 상품군들을 추천해주지만, 새로운 상품이거나 비인기상품의 경우는 추천리스트에 들어갈 수 없고, 두 번째는 시간에 따라 고객의 선호 상품이 변할 경우, 기존 소비자의 소비패턴으로 인해 즉각적인 피드백이 어렵고 특정 상품만 반복적으로 추천하는 과도한 특수화 문제가 발생해 협업 필터링과 비교하면 추천리스트의 다양성을 보존할 수 없다.

향후 소비자의 선호 상품 이력, 타 소비자의 상품 평가 등 소비자 데이터를 수집하고 협업 필터링 등 알고리즘을 추가적으로 학습을 하고, 기존 알고리즘과 융합시켜 하이브리드 필터링을 구성한다면 상품 내용과 유저들의 평가를 기반으로 추천시스템의 정확도 향상과 추천리스트 검증에 관해 연구를 진행할 것이다. 추천리스트의 성능 향상과 어느 정도의 신뢰수준인지 소비자에게 알려준다면 소비자도 모르는 선호 성향을 끌어내어 소비자들의 작은 고민을 덜어 내줄 수 있을 거라 기대한다.

## 참 고 문 헌


[1] 배영임, 신혜리 (2020). “코로나19, 언택트 사회를 가속화하다. 이슈&진단”, 1-26

[2] KOSIS(통계청, 2021년 3월 온라인쇼핑 동향 및 1/4분기 온라인 해외 직접 판매 및 구매 동향 ), 2021.05.25.

[3] 조선Biz.,“2시간 배송으로 네이버·쿠팡에 도전장...7월 출범하는 ‘GS리테일’ 조직통합이 관건“ https://biz.chosun.com/distribution/channel/2021/05/28/DMVXB67DABGVFC3W7K6CAGZTOI/?utm_source=naver&utm_medium=original&utm_campaign=biz

[4] 손해용, 김기환, “오프라인 소매업의 종말?…‘리테일 아포칼립스’ ” 가속화, 조선일보, 2021.02.08.(2021.05.25.)

[5] 스타트온, “지금은 '배송전쟁' 시대, 빅데이터 기반 새벽배송” https://brunch.co.kr/@startonkr/11

[6] 배은영, 유석종 (2018). 키워드 기반 추천시스템 데이터 셋 구축 및 분석. 한국정보기술학회논문지,16(6), 91-99.

[7] 조선Biz., "카카오 알림톡, 올해만 150억건...스팸 대책 필요" https://newsis.com/view/?id=NISX20181011_0000439558&cID=13001&pID=13000

[8] 손지은, 김성범, 김현중, 조성준, “추천시스템 기법 연구 동향 분석”, Journal of the Korean Institute of Industrial Engineers, Vol. 41, No. 2, pp. 185-208, April 2015.

[9] 이말례, 배환국 (2002). “TFIDF를 이용한 키

워드 추출 시스템 설계." 인지과학, 13(1), 1-11
[10] 곽란, 김성현, 이성우, 서봉원 (2018). "지능적 이슈 트래킹 시스템." 한국HCI학회 학술대회, 351-356
[11] minmaxscaler, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[12] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi "Semantic Cosine Similarity"
[13] Poonam B. Thorat, R. M. Goudar, Sunita Barve, "Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System", International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 4, January 2015

## 저 자 약 력

**채 보 병**
이메일 : bbc@youthbioglobal.com

- 2013년 3월~2015년 2월 상일 미디어 고등학교 졸업
- 2021년 3월~현재 ㈜유스바이오글로벌 연구원

**유 지 호**
이메일 : jiho@youthbioglobal.com

- 2009년 9월~2012년 2월 고려대학교 교육대학원 교육정보전공 석사
- 2014년 9월~2016년 2월 동국대학교 교육학과 HRD 박사 수료
- 2017년~현재 유스바이오글로벌 이사
- 관심분야 : 에듀테크, 챗봇

**유 승 호**
이메일 : ceo@youthbioglobal.com

- ㈜유스바이오글로벌 대표이사
- 서울대학교 의과대학 임상의과학과 총동창회장
- 한국의료제품임상연구회 부회장
- 동국대학교 겸임교수 (Medical Biotech Dept)
- 산자부 무역기술장벽 대응위원회 바이오 의료산업계 대표위원
- 식약처 민관국제협력단 IMDRF 임상평가팀장
- 관심분야 : 경영 / R&D