옷 추천 시스템 데이터 셋 구축을 위한 텍스트 데이터 마이닝

이주상*, 정선태**, 차준엽*
*숭실대학교 대학원 정보통신공학과
**숭실대학교 스마트시스템소프트웨어학과
ibmhbm2@naver.com, cst@ssu.ac.kr

Text Data Mining to build a Dataset for Clothing Recommendation System

Ju-Sang Lee*, Sun-Tae Chung**,Jun-Yup Cha
*Dept. of Information and Telecommunication Eng., Graduate School, Soongsil University
**Dept. of Smart Systems Software, Soongsil University

요 약

추천시스템은 대량의 정보를 이용하여 특정 사용자가 선호할만한 상품의 리스트를 추천하는 것이다. 현재 추천시스템으로 유명한 Netflix, Amazon, Youtube 등은 기업내의 상품 및 사용자 데이터를 토대로 이루어 졌으나 스타트 업 및 소규모 기업이 추천 시스템을 구축하기 위해선 기반이 될 데이터셋 자체가 없으며 데이터 수집에도 한계가 있다. 본 논문에서는 옷 추천 시스템 구축을 위해 특정 기업만이 아닌 모든 의류매장들이 사용할 수 있는 데이터 셋 구축 방법에 대해 제안하며, 고객 데이터 셋 구축을 위한 텍스트 데이터 마이닝 처리 과정과 결과에 대해 기술한다.

1. 서론

추천 시스템은 여러 상품들중 특정 사용자가 가장 선호할만한 리스트만 찾아 추천하는 정보 필터링 시스템이다. 효과적인 추천 시스템 구축을 위해서는 추천 알고리즘의 성능과 충분한 양의 고객,상품 데이터가 필요하다. 알고리즘의 성능은 계산량이나 연산 속도뿐 아니라 빅데이터를 처리할 수 있는 환경인지를 고려해야한다. 예를 들어, 빅데이터 시스템이 구축되어 있다면 CF(Collaborative Filtering)나 DL(Deep Learning)을 사용할 수 있겠지만, 대량의 정보가 아니라면 오히려 성능이 떨어질수도 있기 때문이다[1]. 그리하여 전제조건이 되는 것이 고객 데이터와 상품 데이터이다. Netflix, Amazon, Youtube 등 추천 시스템을 적극 사용하고 있는 세계적인 대기업들은 그들만의 고객과 상품 데이터베이스를 기반으로 추천 알고리즘을 적용한다. 여기서 충분한 데이터가 없는 스타트업 및 중소기업은 추천 시스템 구축 이전에 콜드스타트 문제에 부딪치게 된다. 만약 추천 시스템을 위한 공공 데이터셋이 존재한다면 데이터가 없는 스타트업들이 추천시스템 구축의 한계를 극복하는 것은 물론이며 이미 추천시스템을 적극 사용중인 기업이라도 연구개발 목적으로 사용할 수 있을 것이다.

본 논문에서는 옷 추천 시스템을 위한 상품 및 고객 정보 데이터베이스 구축을 위해 Scrapy 기반의 웹 크롤러를 통해 Amazon,Yoox,Shopbop 등의 의류 페이지에서 상품 데이터와 리뷰 데이터를 수집하였고, 자연어 처리를 통한 리뷰 분석을 통해 리뷰 작성자의 정보와 상품에 대한 선호도를 알아내어 추천 시스템에 적합한 상품,고객 데이터셋을 구축하는 과정을 설명한다.

2. 제안 방법

옷 추천 시스템을 위한 데이터 셋 구축을 위해 소비자들이 상품 구매를 결정하기까지 어떤 요인들이 영향을 끼치는지 고민하고 추천 시스템에서 필요로 하는 데이터의 성격을 파악해야 한다. 소비자들의 의류 제품 구매 결정 요인에 대한 관련 연구에서는 제품의 속성이 구매 결정에 유의미한 영향을 미치며 다른 요

인들은 제품의 표현적, 물리적 속성, 사진 효과, 연령 등의 순으로 영향력이 보인다고 말한다.[2] 추천 알고리즘은 특정 사용자가 특정 속성들에 대해 얼마의 평점(선호도)를 줬는지 계산하여 아직 구매하지 않은 상품을 그 사용자가 얼마나 선호할지 결정하는 것이 기본이며, 요즘은 더 나아가 상품의 특정 항목이 선호에 어떤 관계가 있는지를 알고리즘적으로 알아내는 잠재 모델 기반 추천 알고리즘이 많이 사용된다.[3] 추천 알고리즘에 적용할 것을 고려하여 추려낸 옷 추천 속성들은 <그림 1>의 표와 같다.

고객 정보

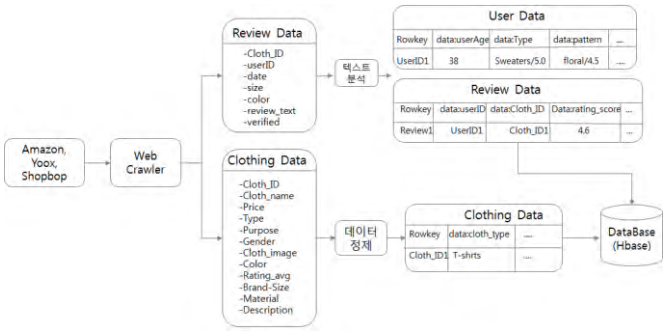
신체적 정보	사회적 정보	기타
나이	직업	구매 목적
성별	수입	선호 스타일
체형(사이즈)	지역	

상품 정보

디자인	기능	가치	기타
타입 (티셔츠, 드레스, 스웨터..)	목적 (사업, 파티, 캐주얼..)	가격	브랜드
색	재료	할인	매장정보
패턴	성별		세탁방법
특징			출시일
사이즈			

<그림1> 추천에 영향을 미치는 속성

상품의 각 속성들에 대해 기존 사용자들의 구매내역이 있고, 그 사용자들의 프로필이 존재한다면 신규 고객이 원하는 상품도 기존 사용자들의 데이터를 기반으로 예측할 수 있을것이다.



<그림2> 옷 추천 시스템 데이터 셋 구축 프로세스

<그림 2>는 옷 추천 시스템 데이터 셋 구축 과정이다. 웹 크롤러는 Amazon,Yoox,Shopbop 등의 의류 페이지에서 앞서 정리한 추천 알고리즘을 위한 속성들을 상품 데이터로 수집해 상품 데이터 테이블을 구축하고 각 상품에 대한 리뷰 데이터를 분석해 사용자의 나이, 상품 선호도 등을 추출해 사용자가 상품의 각 속성에 매긴 평점과 사용자의 프로필을 가지고 있는 사용자 데이터 테이블과 리뷰 데이터 테이블을 만든다. 상품, 사용자,리뷰 데이터 테이블의 관계는 사용자의 구매 기록, 상품에 대한 평점이 되면서 <표 1>과 같이 고객 D 의 상품 1 에 대한 선호도를 예측할 때 고객들의 상품 구매내역과 평점을 기반으로 고객들간의 유사성을 계산하여 고객 D 가 상품 1 에 대해 얼마 만큼의 선호도를 가질지 예측할수 있는 CF 기반 추천 알고리즘을 위한 기본 데이터 셋이 구성된다.

	상품 1	상품 2	상품 3	상품 4
고객 A	5	3	2	
고객 B	4	4		5
고객 C	2		4	3
고객 D	?	4	3	3

<표 1> 추천시스템의 평점 테이블 예

2-1. 데이터 수집

초기 데이터는 Scrapy[4]를 기반으로 작성한 웹 크롤러를 사용하여 Amazon[5],Yoox[6],Shopbop[7]등에서 300,000 개 가량의 상품 데이터와 리뷰 데이터를 수집하였다. 상품 데이터는 별도의 분석과정을 거치지 않으나 <표 2>와 같이 서로 다른 웹 페이지인 만큼 같은 데이터도 다르게 표기가 되었으며 일부 상품 정보 업로드가 글 작성자의 주관적인 생각으로 작성되기 때문에 ‘t shirt midi dress’ 와 같이 타입이 구체적으로 명시되지 않거나 같은 ‘blue’색상에 대해서도 ‘navy blue’,’cobalt blue’ 등 유사 데이터도 많이 존재한다. 향후 추천 시스템에서 쉽게 분석하고 검색이 용이하게 유사 데이터는 통합하고 구체적이지 못한 데이터는 하나의 형태로 교정하는등 정제를 거친후 데이터베이스에 저장한다.

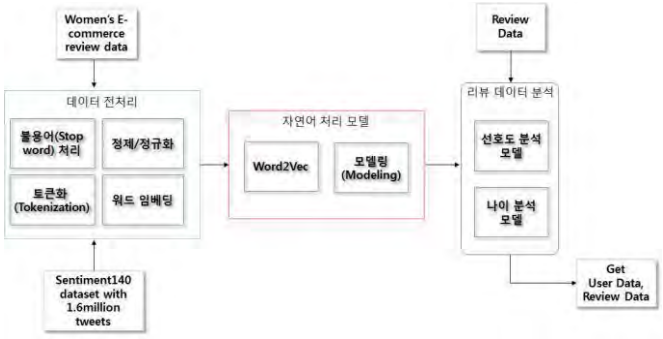
웹페이지 / 속성	Amazon	Yoox	Shopbop
사이즈	Small,Medium,Large	46,48,50	S,M,L
가격	$ 90.00	$90.00	US$90.00
타입	T-Shirts	t-shirt	Tops
색상	Navy Blue	Blue	Cobalt blue
옷 이름	High waist T shirt midi dress with pockets	PRADA	SUNDRESSES

<표 2> 수집한 직후의 데이터 형태

2-2. 리뷰 데이터 분석

추천을 위해 고객의 데이터는 그동안의 구매내역과 구매한 상품에 대한 평점(선호도), 상품에 대한 평점은 더 나아가 상품의 어떤 속성이 선호에 영향을 미치는지도 고려하기 위해 상품의 각 속성에 대한 선호도까지 고려할 수 있다. 또한 앞서 추천 알고리즘에 들어갈 속성으로 정의했던 나이,성별,체형등의 정보를 생각할 수 있다. 본 연구에서는 나이,성별,체형 세가지 속성들중 성별은 구매한 상품내역의 데이터에서 추출할수 있었고 체형은 고객이 직접 이미지를 업로드 하지 않으면 분석이 불가능했기에 제외하고 자연

어 처리를 기반으로 리뷰 작성자의 나이를 분석하고 상품에 대한 선호도를 분석하였다. 딥 러닝 기반 자연어 처리 분석 모델을 통해 리뷰 텍스트로부터 원하는 정보를 추출하며 <그림 3> 과 같은 과정을 거친다.



<그림 3> 고객 데이터 분석 자연어 처리 과정

본 연구에서는 23485 개의 의류 상품 리뷰,평점,카테고리,작성자 나이 등의 정보가 있는 Women's E-Commerce review data[8]과 1,600,000 개의 트위터 글과 트윗에 대한 작성자의 감정 지표가 있는 Sentiment 140 dataset with 1.6million tweets[9]를 훈련 데이터 셋으로 사용하였다.

2-2-1 데이터 전처리

분석할 용도에 맞게 텍스트를 사전 처리하는 작업을 한다. 데이터 전처리를 위해 딥 러닝을 위한 Python 라이브러리 Keras 와 자연어처리를 돕는 여러 툴을 제공하는 NLTK 라이브러리를 사용하였다.

불용어처리:'I','You','it'과 같이 문장내에 등장 빈도가 높으나 본 연구의 텍스트 분석 목적에 있어서 의미를 갖지 않는 단어들이 있다. NLTK 가 정의한 불용어를 사용하여 이러한 단어들을 제거한다.

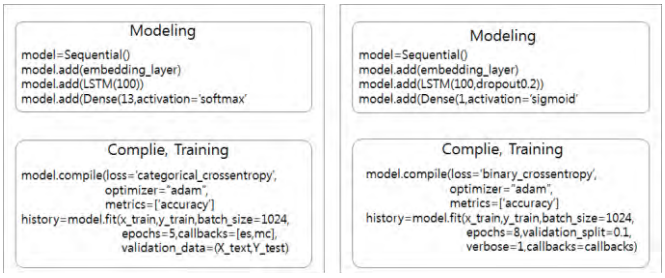
정제/정규화: 분석에 영향을 주지 않는 특수문자, 구두점등을 제거하고 결측,이상이 있는 데이터는 훈련에 잘못된 영향을 줄 수 있어 이러한 데이터들을 제거한다.

토큰화: 주어진 텍스트 데이터를 분석에 의미를 가지면서 가장 작은 단위로 나누는 것을 토큰화라고 한다. 여기서는 토큰의 기준을 단어로 정하였다.

워드임베딩: 자연어처리를 위해 필요한 과정으로 토큰화한 단어들에 실수를 부여하고 벡터화하는 것을 말한다.

2-2-2 자연어 처리 모델

컴퓨터가 분석할 수 있도록 하기 위해선 단어를 숫자화 시키는 워드임베딩 과정이 필요하다. Word2Vec 모델은 주어진 문장에서 모든 단어의 의미를 벡터화하여 단어간 유사도를 반영한다.[10] 벡터화된 데이터는 분석 모델에 임베딩 층으로 들어간다.

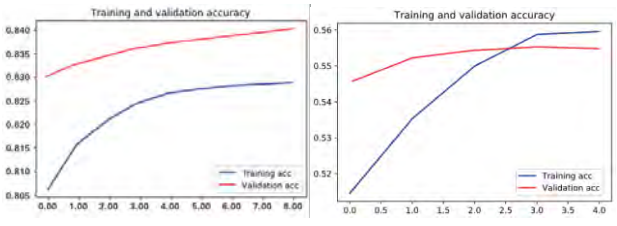


<그림 4> 나이분석,선호도 분석 모델과 컴파일

텍스트 분석 모델은 <그림 4>와 같이 구성된다. embedding_layer 는 앞서 구성한 벡터화된 데이터로 인공 신경망의 층의 하나로서 추가된다. Dense()는 전결합층을 추가하는 것이다. 첫번째 인자는 출력 뉴런의 수, 두번째 인자는 출력층에 사용되는 함수를 의미한다. 나이 분석의 경우 더 간결하고 확실한 결과를 출력하기 위해 나이를 5 세 단위로 레이블을 나눴다. 따라서 레이블 만큼의 수가 들어간다. 선호도 분석의 모델의 경우 0 에서 1 사이의 실수 결과를 출력하며 1 에 가까울수록 선호도가 높음을 의미한다. 출력되는 결과는 한가지이기 때문에 1 이 입력된다. 그리고 각각 다중 클래스 분류와 이진 분류 문제에서 주로 사용되는 Softmax 함수와 sigmoid 함수를 적용하였다. model.compile()과 model.fit()은 각각 모델링한 신경망을 컴파일하고 훈련시키는 과정이다. 데이터의 크기를 고려하여 과적합 방지를 위해 훈련회수를 각각 5, 8 로 진행하였다.

3. 결과

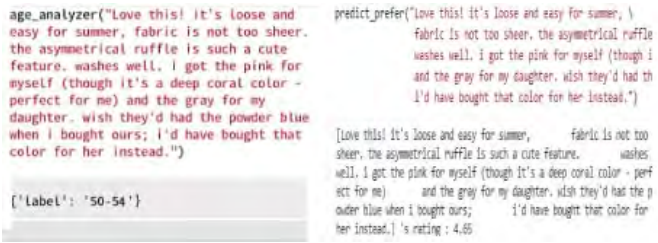
훈련용으로 사용한 데이터 셋을 8:2 비율로 나눠 8 은 훈련에 사용했으며 2 는 정확도 검증을 위한 테스트 데이터로 사용하였다. <그림 4>는 학습한 모델의 평가 결과를 나타낸다.



<그림 5> 선호도분석(좌) 나이분석(우)의 테스트 정확도

선호도 분석과 나이 분석의 평가 정확도는 각각 84%와 55%로 나타났다. 큰 차이의 정확도를 보인 원인은 데이터 셋의 규모 차이로 생각된다. 선호도 분석은 1,600,000 개의 충분한 양의 데이터 셋을 바탕으

로 84%의 정확도를 보인것으로 여겨지며, 나이 분석의 경우 23,000 개 가량의 데이터는 텍스트로부터 단어가 내포하는 연령대별 특징 및 유사도를 계산할만한 충분한 크기의 데이터가 아니었던 것으로 생각된다. <그림 6>은 구현한 모델들을 사용하여 텍스트를 입력했을때 모델의 출력값을 보여준다.



<그림 6> 나이 분석과 선호도 분석의 출력결과

4. 결론

본 논문에서는 옷 추천 시스템을 위해 데이터를 수집하여 상품 데이터 셋을 구축하고, 텍스트 분석을 통한 고객 프로필 데이터 셋 구축 방법을 추천 시스템을 위한 기반 데이터 셋 구축 방법을 제안했다. 텍스트를 통한 나이 분석 단계에서 충분한 정확도에 도달하지 못하여 고객 데이터를 온전히 구축하지 못하였다. 그러나 나이 분석 모델의 정확도가 적은 양의 데이터 셋으로 얻은 결과라는 점을 감안했을때, 향후 충분한 훈련 데이터 셋이 확보된다면 더 유의미한 결과를 도출해낼 수 있을 것으로 판단된다. 앞으로는 SNS 에서 업로드한 이미지를 기반으로 이미지 데이터 분석을 진행하여 추천 시스템을 위한 질 높은 데이터 수집을 계속 할 계획이다. 연구가 계속 성과를 보인다면 콜드스타트 문제에 직면해있는 스타트업도 활용 가능한 추천시스템 공공 데이터 셋 구축이 가능해질 것이라 생각한다.

참고문헌

[1] Sanjeevan Sivapalan, Alireza Sadeghian, Hossein Rahnama, Asad M. Madni, Recommender systems in e-commerce, 2014 World Autumation Congress(WAC), p179-184,2014

[2] 지혜경, 인터넷 쇼핑몰에서의 의류제품 구매결정요인, 한국의상디자인학회지, 14(2) p185-189

[3] 한국콘텐츠진흥원, <방송 트렌드 &인사이트> 2016년 4,5 월호(vol.05): 콘텐츠 추천 알고리즘의 진화

[4] Scrapy, https://docs.scrapy.org/en/latest/

[5] Amazon, Men's Fashion, Women's Fashion, https://www.amazon.com/ref=nav_logo

[6] Yoox, https://www.yoox.com/kr

[7] Shopbop, https://www.shopbop.com/

[8] Women's E-commerce Clothing Reviews, https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[9] Sentiment140 dataset with 1.6million tweets, https://www.kaggle.com/kazanova/sentiment140

[10] Justin Garten, Kenji Sagae, Volkan Ustun, Morteza Dehghani, Combining Distributed Vector Representations for Words, Associations for Computational Linguistics, Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Pages 95-101, 2015