ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 December: 21(4): 93~110 http://dx.doi.org/10.13088/jiis.2015.21.4.093

빅데이터 기반 추천시스템 구현을 위한 다중 프로파일 앙상블 기법

김민정
국민대학교 데이터사이언스학과 석사과정
(snseej12@naver.com)

조윤호
국민대학교 경영대학 경영학부 교수
(www4u@kookmin.ac.kr)

기존의 협업필터링 추천시스템 연구는 상품에 대한 고객의 평점(rating)이나 구매 여부 데이터로부터 하나의 프로파일을 생성하고 이를 기반으로 추천 성능을 향상시킬 수 있는 새로운 알고리즘을 개발하는 위주로 진행되어 왔다. 그러나 빅데이터 환경이 도래하면서 기업이 수집할 수 있는 고객 데이터가 풍부해지고 다양해짐에 따라, 보다 정확하게 고객의 선호도나 행태를 파악하는 것이 가능하게 되었고 이러한 데이터, 즉 퍼스널 빅데이터(personal big data)를 추천시스템에 활용하는 연구의 필요성이 대두되고 있다. 본 연구에서는 마케팅의 시장세분화 이론에 근거하여 퍼스널 빅데이터로부터 고객의 선호도나 행태를 다양한 관점에서 표현할 수 있는 5종의 다중 프로파일(multimodal profile)을 개발하고, 이를 활용하여 협업필터링 추천시스템의 성능을 개선하고자 한다. 제안하는 5종의 다중 프로파일은 프로파일 통합 유사도, 개별 프로파일 유사도 평균, 개별 프로파일 유사도 가중 평균이라는 세 가지 앙상블 기법을 통해 협업필터링의 이웃(neighborhood) 탐색과정에 적용된다. 실제 퍼스널 빅데이터에 본 연구에서 제안하는 방법론을 적용한 결과, 단일 프로파일을 사용하는 협업필터링 알고리즘보다 추천 성능이 상당히 개선되었으며 앙상블 방법 중에서는 개별 프로파일 유사도 가중 평균 기법이 가장 높은 추천 성능을 보여주었다. 본 연구는 빅데이터 환경에서 추천시스템을 개발하고자 할 때, 어떠한 성격의 데이터로부터 고객의 특성을 규명하는 프로파일을 만들고 이를 어떻게 결합하여 사용하는 것이 효과적인 지 처음으로 제안하였다는 점에서 그 의의가 있다.

주제어 : 빅데이터, 추천시스템, 협업필터링, 다중 프로파일, 앙상블 기법

논문접수일 : 2015년 11월 25일 논문수정일 : 2015년 12월 14일 게재확정일 : 2015년 12월 14일
교신저자 : 조윤호

1. 서론

협업필터링(Collaborative Filtering) 기반 추천시스템은 성능이 가장 우수한 방법 중 하나로 알려져 있으며, 영화, 음악, 유통 등 다양한 분야에서 적용되고 있다. 기존의 연구 동향을 살펴보면, 협업필터링 추천시스템은 주로 상품에 대한 사용자의 평점(rating)이나 구매여부 데이터로부터 하나의 프로파일을 생성하여 새로운 알고리즘을 개발하는 위주로 연구되어왔다(Billsus and Pazzani, 1998; Herlocker et al., 2004; Lee and Park, 2007; Bar et al., 2013; Lee and Kim, 2013).

그러나 다양한 채널이 등장하고 이로 인해 정보의 생산과 보유가 증가하면서 디지털 데이터가 기하급수적으로 증가하는 추세이다(Bok and Yoo, 2014). 특히 2011년에는 전세계의 디지털 정보량이 1.8ZB에 달하여 제타바이트 시대에 진입하면서 대규모 데이터가 이슈로 떠오르며, 빅데이터(big data)라는 용어가 등장하였다(Ward and Barker, 2013). 이러한 빅데이터 환경이 도래

하면서 기업에서도 수집할 수 있는 고객 데이터가 풍부해지고 다양해짐에 따라 보다 정확하게 고객의 선호도나 행태를 파악하는 것이 가능하게 되었다. 즉, 퍼스널 빅데이터(personal big data)를 추천시스템에 활용하여 기존의 알고리즘 개발 기반에서 데이터의 품질 향상 기반의 성능 향상 연구의 필요성이 대두되고 있다.

본 연구에서는 실제 퍼스널 빅데이터로부터 마케팅의 시장세분화 이론에 근거한 5종의 다중 프로파일을 개발하고, 고객의 선호도나 행태를 다양한 관점에서 표현할 수 있도록 하였다. 또한, 이를 기반으로 협업필터링의 이웃(neighborhood) 탐색과정에 적용할 수 있는 3가지의 앙상블 기법을 제안하고 추천 성능을 개선하고자 한다.

논문의 구성은 다음과 같다. 먼저 관련 선행 연구들을 제 2장에서 요약하고, 제 3장에서는 프로파일 개발과 유사도 계산, 그리고 추천시스템 성능 개선을 위한 앙상블 방법론을 제안한다. 그리고 제 4장에서는 실제 퍼스널 빅데이터를 기반으로 한 실험과 성능 평가에 대해 기술하며, 마지막으로 제 5장에서는 본 연구의 의의 및 한계점에 대해 서술하는 것으로 구성된다.

2. 선행 연구

2.1 협업필터링 기반 추천시스템

협업필터링(Collaborative Filtering)은 사용자와 상품에 대한 정보를 기반으로 프로파일을 생성하고, 이를 바탕으로 목표고객과 유사한 구매 선호도를 가진 다른 사용자가 구매하거나 선호하는 항목을 추천하는 방법이다(Kim et al., 2012; Cabral et al., 2014). 이러한 협업필터링 기반 추천시스템은 Goldberg et al.(1992)에 의해 처음 소개된 이후 지금까지 성능이 가장 우수한 방법 중 하나로 알려져 있으며, 영화, 음악, 유통 등 다양한 분야에서 적용되고 있다. 기존의 협업필터링 기반 추천시스템 연구의 동향을 살펴보면, 주로 상품에 대한 사용자의 평점(rating)이나 구매 여부 데이터로부터 단일 프로파일을 생성하여 새로운 알고리즘을 개발하거나 기존 알고리즘을 응용하는 위주로 연구되어왔다(Billsus and Pazzani, 1998; Lee and Park, 2007; Bell et al., 2009; Bar et al., 2013; Lee and Kim, 2013).

이러한 알고리즘 개선 측면에서 앙상블 기법을 적용한 연구 중 대표적인 것이 Netflix의 영화 추천시스템 성능 개선 사례이다. Bell et al.(2009)와 Piotte and Chabbert(2009)의 연구에 따르면, Netflix 고객들의 영화 평점 정보를 기반으로 단일 프로파일을 생성한 후 인공신경망과 의사결정나무 등의 데이터 마이닝 기법과 협업필터링 기법을 앙상블하여 추천 성능을 약 10% 개선하였다. 또한, Bar et al.(2013)의 연구에서도 MovieLens의 영화 평점 정보를 사용하여 하나의 프로파일을 생성하고 사용자 간의 k-NN 모델과 상품 간의 k-NN 모델의 예측 결과를 결합하는 fusion 기법 등을 제안하여 추천 성능이 개선됨을 보였다. 영화 뿐만 아니라 유통 분야의 연구도 있다. Lee and Kim(2013)은 온라인 상점의 구매 데이터를 이용하여 사용자-상품 매트릭스 기반 프로파일을 생성하였고, 로지스틱 회귀분석, 의사결정나무, 인공신경망 모형을 구축하여 앙상블 하였다.

반면, 본 연구와 같이 다중 프로파일을 이용하여 추천시스템 성능을 개선하는 앙상블 연구도 다수 존재한다. 대표적으로 Pazzani(1999)의 연구를 꼽을 수 있다. 이 연구에 따르면, 레스토랑에

대한 고객의 선호 정보와 해당 고객에 대한 인구통계학적 정보, 그리고 레스토랑의 음식(상품) 정보 등 서로 유형이 다른 프로파일들에 대해 협업필터링 기법과 내용기반 필터링 기법 등을 각각 사용하여 앙상블하였고 기존 추천시스템 성능보다 우수하다는 것을 입증하였다. 이와 유사하게 건강식품에 대한 구매 정보와 인구통계학적 정보, 해당 상품을 검색한 질의어 정보를 각각 협업필터링 기법과 내용기반 필터링 기법으로 앙상블하여 추천 성능을 개선한 사례도 있다(Kim et al., 2012). 이는 정형데이터가 아닌 검색 질의어라는 비정형 데이터를 프로파일로 생성하였다는 점에 의의가 있다. 이 밖에도 Cabral et al.(2014)의 연구처럼 영화와 관련된 다양한 메타데이터를 이용하여 제안한 앙상블 알고리즘별 성능을 비교한 사례도 있다.

그러나 앞서 언급한 다중 프로파일 기반 추천시스템 연구들은 평점이나 구매정보에 인구통계학적 정보나 상품 정보를 추가하는데 그쳐, 고객의 선호도나 행태를 다양한 측면에서 적절히 표현하지 못하고 있을 뿐만 아니라, 서로 성격이 다른 이종의 알고리즘들을 결합함으로써 추후 확장성 측면에서 제약이 존재한다,

2.2 퍼스널 빅데이터

빅데이터란 기존 데이터베이스의 관리 역량을 넘어서는 데이터 집합으로서, 데이터의 양(Volume), 다양성(Variety), 이용 속도(Velocity) 측면에서 특징을 가지고 있다(Bok and Yoo, 2014; Ward and Barker, 2013). 특히 웹 로그 데이터, 센서 데이터, 위치 데이터, 소셜 데이터 등 새로운 데이터가 생성되면서 이러한 데이터를 활용할 수 있는 역량과 방법 등이 중요하게 인식되고 있다.

그 중에서 퍼스널 빅데이터(Personal big data)는 Kim et al.(2012)에 따르면, '사용자의 활동에 의해 생성된 빅데이터 속성을 지닌 데이터'라고 정의되고 있다. 여기서 빅데이터의 속성은 앞서 언급한 3V를 뜻하는데, 먼저 퍼스널 빅데이터는 일생에 거쳐 기록됨과 동시에 비디오나 오디오 자료도 포함되기 때문에 빅데이터의 Volume 속성을 지닌다. 또한 다양한 기기들로부터 정형, 비정형, 반정형의 데이터를 제공받으며, 이러한 데이터들은 실시간으로 쌓이므로 각각 Variety와 Velocity 속성을 만족한다고 할 수 있다.

Gurrin et al.(2014)의 연구에서는 라이프 로그(Life log)에 대해 서술하였는데, 라이프 로그가 쌓이는 것을 라이프 로깅(Life logging)이라고 정의하고 있다. 이러한 라이프 로깅은 일반적으로 우리 주변 어디에서나 일어나고 있으며, 이렇게 쌓이는 빅데이터는 큰 도전이자 기회라고 서술하였다. 특히 일상생활에서 쌓이는 라이프 로그가 개개인의 측면에서는 퍼스널 빅데이터임을 정의하고 있으며 이 연구 역시 퍼스널 빅데이터가 빅데이터의 3V 속성을 지닌다고 주장하였다.

2.3 사용자 프로파일링과 시장세분화 이론

2.3.1 사용자 프로파일링

사용자 프로파일링(User Profiling)은 일반적으로 지식기반(Knowledge-based)이거나 행동기반(Behaviour-based)이다. 지식기반은 설문이나 인터뷰를 통해 사용자의 정보를 얻어 모델을 구축하는 방식이며, 행동기반은 사용자의 행동에서 의미 있는 패턴을 발견하기 위하여 기계학습 방법을 이용하는 것이다(Middleton et al., 2004). 전

통적인 협업필터링 기반 추천시스템에서는 웹 사용자들이 스스로 관심있는 상품에 대해 직접 평점(rating)을 입력하는 명시적인 방법을 이용하여 프로파일링하는데(Park et al., 2006), 이는 지식기반 프로파일이라고 할 수 있다.

반면, 행동기반 프로파일에는 주로 상품구매 등 사용자의 행동 정보가 사용된다. Middleton et al.(2004)에 따르면, 대부분의 추천시스템에 사용되는 사용자 프로파일링은 사용자가 어떤 것에 관심이 있는지를 나타내는 행동기반이라고 하였다. Weng et al.(2004)의 연구에서는 만약 사용자가 해당 상품을 구매하였다면 1, 구매하지 않았다면 0의 값을 사용하여 구매정보를 이진형(binary) 프로파일로 표현하였다. 그러나 이러한 사용자 프로파일링 기법은 사용자에 대한 정보가 상품 구매에 제한되어 있을 때 적합하며, 사용자 행동에 관한 변수와 데이터가 다양할 경우 사용자의 특성을 규명할 수 있는 세분화 기준이 추가적으로 필요하다고 할 수 있다(Kim, 2012; Park et al., 2006).

2.3.2 시장세분화 이론

시장 세분화(Market segmentation) 이론은 서로 다른 특징, 행동양식 등으로 시장을 나눈다는 개념이며, 이를 위해서는 유사한 개인특성을 갖는 소비자 그룹이 발굴되어야 한다(Kim and Oh, 2009). 시장 세분화를 위한 방법은 사전 세분화(Priori segmentation)과 사후 세분화(Posteriori segmentation)로 나눌 수 있다. 사전 세분화는 분석자가 미리 세분화 기준을 결정한 다음 데이터를 수집하고 세그먼트 특성을 분석하는 경우이며, 사후 세분화는 데이터 수집이 먼저 이루어지는 경우이다. 시장 세분화는 사전 세분화에서 사후 세분화 방법으로 이동해왔다고 볼 수 있는데(Mazanec, 2000), 사후 세분화 방법이 하나의 변수가 아닌 다양한 소비자 정보를 바탕으로 여러 관점에서 세그먼트를 도출할 수 있기 때문이다. 본 연구에서는 사후 세분화 방법을 이용하여 사용자 행동에 관한 유형을 분류하고자 한다.

기존의 선행연구에 따르면, 시장 세분화 이론은 소비자 시장에서 시장을 세분화하고 그 세분시장의 프로파일을 개발하기 위한 4가지 기준변수를 제시한다(Claycamp and Massy, 1968). 세분화 기준은 인구통계학적 변수, 지리적 변수, 심리적 변수, 행동적 변수로서 그 중 인구통계학적 변수를 기준으로 한 세분화 방식이 가장 많이 사용되고 있다. 그러나 Kim(2002)은 고객정보뿐만 아니라 웹로그 데이터, 외부환경 정보 등을 복합적으로 사용하여 확장된 의미의 로그분석을 제안하였다. 이를 통해 사용자 특성별로 개인화된 서비스를 제공할 수 있는 기반이 된다고 하였다. 본 연구에서도 시장 세분화 이론 중 지리적 기준을 포함시킨 인구통계학적 기준과 행동적 기준에 초점을 맞추어 사용자의 특성을 파악할 수 있는 프로파일을 규명하고자 한다. 개성이나 욕구 등의 심리적 기준 변수는 설문이나 인터뷰를 통해 얻을 수 있는 사용자 정보이기 때문에, 본 연구에서는 측정 불가능하다고 판단하였다. 따라서, 사용자들의 기본 정보를 기반으로 한 인구통계학적 프로파일과 웹 상에서의 행동 정보를 토대로 한 고객 행동 프로파일을 규명하고자 한다. Kim and Oh(2009)의 연구에서도 고객리뷰에 대한 텍스트 마이닝 분석을 하기 전에 시장세분화 이론을 도입하여 비슷한 유형의 고객 그룹을 발굴하여 정확도를 높인 사례가 있다.

소비자 시장에서의 세분화 기준은 <Table 1>과 같다(Claycamp and Massy, 1968). 그 중에서

웹 상에서 적절하다고 판단되는 인구통계학적 기준, 지리적 기준, 행동적 기준에 따른 변수들을 채택하여 프로파일로 생성하였다. 특히 본 연구에서 초점을 맞추어 도입할 행동적 기준의 하위 개념에 대해 세부적으로 설명하자면, 먼저 구매 또는 사용상황에 따른 변수, 사용자가 추구하는 편익이나 혜택에 따른 변수, 사용량 또는 사용률에 따른 변수 등으로 나눌 수 있다(Claycamp and Massy, 1968; Kim, 2012). 본 연구에서는 퍼스널 빅데이터로부터 웹 사용자의 속성을 규명하기 위해 인구통계학적 기준과 지리적 기준을 기반으로 한 인구통계학적 프로파일과, 행동적 기준을 기반으로 한 rating, 선호 사이트, 인터넷 사용행태, 검색 키워드 토픽 프로파일을 생성하였으며, 이에 대한 내용은 3.1절에서 자세히 다룬다.

〈Table 1〉 Criterion on user attribute identification

	consumer market	Web
demographic criterion	gender, age, income, occupation, religion, etc.	gender, age, occupation, etc.
Geographic criterion	location, size of city, population density, climate, etc.	Internet access location
Psychographic criterion	lifestyle, personality, desire, etc.	impossible to measure
Behavior criterion	pursuit benefits, product preference, etc.	pattern of internet usage, searching keyword, site preference, etc.

3. 제안 추천 방법론

제안하는 추천방법론은 다음과 같다. 먼저, 퍼스널 빅데이터로부터 사용자의 선호도나 행태를 다양한 관점에서 파악할 수 있는 5종의 다중 프로파일을 생성한다. 다음으로, 생성한 프로파일을 기반으로 사용자 간 유사도를 계산하여 협업 필터링의 이웃(neighborhood) 탐색과정에 적용한다. 이 과정에서 각 프로파일을 결합하여 유사도를 계산하는 3가지 앙상블 기법이 사용된다. 마지막으로, 선택된 이웃이 가장 선호하는 상품을 해당 고객에게 추천한다.

3.1 프로파일 생성

본 연구에서 웹 사용자의 속성을 규명하기 위한 프로파일은 마케팅에서 일반적으로 사용하는 소비시장에서의 시장세분화 이론에 근거하였다. 그 중에서 웹 사용자에 적합한 변수를 채택하였고, 해당 변수를 기준으로 rating, 선호사이트, 인구통계학, 인터넷 사용행태, 검색 키워드 프로파일과 같이 5종의 다중 프로파일을 생성하였다. 시장 세분화 이론의 기준에 따른 프로파일 생성 목록은 <Table 2>와 같다. 인구통계학적 기준에 따라 인구통계학적 프로파일을 생성하였으며, 행동적 기준의 하위 개념에 기초하여 rating 프로파일(상품에 대한 관심 정보), 선호 사이트 프로파일과(인터넷 사용패턴에 대한 정보), 인터넷 사용행태 프로파일(인터넷 사용량 또는 사용률

〈Table 2〉 5 profiles based on market segmentation theory

criterion	profile
demographic criterion	demographic
behavior criterion	rating
	site preference
	internet usage
	topic

에 대한 정보), 그리고 검색 키워드 토픽 프로파일(사용자가 추구하는 편익 및 관심사에 대한 정보)을 개발하였다.

	processed food	furniture/ DIY	men's clothing	laptop/ desktop	health food	book/CD
user 5	1	0	0	1	0	0
user 9	1	1	1	0	0	0
user 66	0	0	0	0	0	1
user 171	0	1	0	0	0	0
user 252	0	0	0	0	0	0
user 405	0	0	0	1	1	0

〈Figure 1〉 Part of rating profile

먼저, <Figure 1>은 단일 프로파일을 이용한 추천시스템 연구에서 주로 이용된 구매 여부나 평점(rating) 프로파일인 사용자-상품 매트릭스이다. 본 연구에서는 특정 상품에 대한 페이지를 클릭하는 행동이 그 상품에 대한 구매 관심이라고 판단하였다. 따라서 이진형(binary)의 매트릭스로서, 상품의 중분류 카테고리 145개에 대해 해당 상품 페이지를 한 번이라도 클릭한 적이 있으면 1, 없으면 0으로 표시하였다. 즉, 1번 사용자(user 1)는 '가공식품(processed food)', '노트북/컴퓨터(laptop/desktop)' 카테고리에 해당하는 상품 페이지를 클릭한 적이 1번 이상 있으며, 해당 카테고리에 속한 상품에 관심이 있다고 보았다. 이는 식 (1)에서와 같이, 값이 1이면 사용자 i가 상품 j에 대해 관심이 있음을 의미하고 0이면 관심이 없거나 해당 상품을 모른다는 것을 의미한다. 이러한 값을 바탕으로 rating 프로파일을 구성하였다.

$$B_{ij} = \begin{cases} 1 : 사용자 i가 \ 상품 j에 \ 대해 \ 관심 \\ 0 : 사용자 i가 \ 상품 j에 \ 대해 \ 무관심 \end{cases} \quad (1)$$

	Portal	Shopping	News	Entertain ment	Finance	E-Mail
user 5	0.093	0.059	0.430	0.218	0	0
user 9	0.001	0.323	0	0.039	0.001	0.211
user 66	0.280	0.028	0	0.260	0.014	0
user 171	0.001	0	0.001	0.037	0.519	0.001
user 252	0.216	0.253	0.198	0	0	0.001
user 405	0.027	0	0.315	0	0.266	0

〈Figure 2〉 Part of site preference profile

선호 사이트 프로파일은 <Figure 2>과 같다. 이는 어떤 카테고리의 사이트에 주로 접속하는지를 나타내며, 사용자의 총 인터넷 체류시간 중 사이트 카테고리 22개에 대해 각각 체류한 시간의 비율을 토대로 생성하였다. 본 연구에서는 특정 페이지에 오래 체류했다는 것이 그에 대한 선호도가 반영된 행동이라고 판단하였다. 프로파일 값은 0과 1사이의 값을 가지며, 값이 클수록 해당 사이트의 카테고리에 오래 체류했다는 것을 의미한다. 사용자별 체류시간의 비율을 계산하였기 때문에, 1명당 카테고리 비율들의 전체 합은 1이다.

	Gender	Age	Occupation	Location	Marrige	School
user 5	Female	23	Student	Seoul	N	high-school
user 9	Male	49	Public official	Ulsan	Y	university
user 66	Male	54	Self-employed	Kyungsang	Y	high-school
user 171	Female	21	Student	Kyungsang	N	high-school
user 252	Male	35	Public official	Seoul	N	university
user 405	Female	18	Student	Jeonra	N	middle-school

〈Figure 3〉 Part of demographic profile

그 다음은 인구통계학적 프로파일로서, <Figure 3>에서 확인할 수 있다. 이 프로파일은 접속 지

역과 같은 지리적 변수를 포함하여 사용자의 성별, 나이, 직업, 결혼여부, 최종학력 정보를 나타낸다. 본 연구에서 사용한 실험데이터는 성별(Gender)은 남자와 여자, 나이(Age)는 10세부터 91세까지 분포되어 있고 결혼여부(Marrige) 변수는 미혼과 기혼으로 나뉜다. 또한 직업군(Occupation)은 20개, 지역(Location)은 13개, 최종학력(School)은 6개로 분류되어 있다.

	DAY	TIME	VSITES	COV	VDAYS	D_TIME	D_COV	SCH_ KEYWORD
user 5	balanced	PM	250	4.650	201	2579	1.026	193
user 9	weekdays	work	83	3.298	119	859	1.429	29
user 66	balanced	PM	793	5.940	119	4112	0.675	956
user 171	weekend	PM	345	5.873	206	1872	0.836	2374
user 252	balanced	AM	316	4.821	111	4778	0.896	2911

〈Figure 4〉 Part of internet usage profile

네 번째 프로파일은 인터넷 사용행태 프로파일이다. 이는 주로 어느 요일/시간에 접속하는지 등의 인터넷 사용패턴을 나타낸다 (Figure 4 참조). ‘접속 요일(DAY)’과 ‘접속 시간(TIME)’ 변수의 경우 각각 클러스터링하여 평일에 접속한 비율이 높은 주중형, 주말에 접속한 비율이 높은 주말형, 주중과 주말에 골고루 접속한 균등형으로 그룹을 나누고, 시간은 오전, 오후, 일과시간 그룹으로 나누었다. 이 밖에도 'VSITES'는 본 실험 데이터의 수집 기간인 1년 동안 해당 사용자가 접속한 사이트의 총 개수를 의미하며, ‘COV’ 변수는 사이트 카테고리 22개 간에 얼마나 많이 이동하였는 지에 대한 변동계수(Coefficient of variation)를 의미한다. 또한, ‘VDAYS’는 365일 중 인터넷에 접속한 총 일 수, ‘D_TIMES’는 초 단위의 접속 체류시간의 총 합, ‘D_COV’는 총 접속시간에 대한 변동계수, 'SCH_KEYWORD‘는 사용자가 1년 동안 포털 사이트에 검색한 키워드의 총 개수를 의미한다. 이처럼 인터넷 사용행태 프로파일은 웹 상에서의 행동을 바탕으로 사용자의 속성을 규명할 수 있도록 8개의 변수들로 구성되었다.

	topic 1	topic 5	topic 6	topic 13	topic 21	topic 35
user 5	0.093	0.059	0.430	0.218	0	0
user 9	0.001	0.323	0	0.039	0.001	0.211
user 66	0.280	0.028	0	0.260	0.014	0
user 171	0.001	0	0.001	0.037	0.519	0.001
user 252	0.216	0.253	0.198	0	0	0.001
user 405	0.027	0	0.315	0	0.266	0

〈Figure 5〉 Part of topic profile

마지막으로 검색 키워드 토픽 프로파일(Figure 5 참조)은 포털 사이트 검색 기록에서 추출한 키워드를 바탕으로 생성한 프로파일이다. 사용자의 관심을 파악하기 위해 포털사이트의 검색 기록으로부터 토픽을 추출하는 과정은 Hyun et al.(2015) 연구에서 제안한 방법론 중 하나인 ‘포털사이트 검색 키워드 기반의 고객 클러스터링(Module 3)’ 방법과 같이 수행하였다. 해당 방법론은 사용자가 검색한 키워드로 고객의 관심을 직접적으로 파악할 수 있는 방법이다(Hyun et al., 2015). 수행 프로세스로는 먼저 검색 키워드를 사용자별로 하나의 문서로 나타낸다. 사용자별 검색 키워드 문서를 바탕으로 토픽 분석을 수행하여 각각 다른 주제의 토픽 50개를 도출한다. 그 후 각 사용자별 검색 키워드 문서를 통해 50개의 토픽에 얼마나 관여되었는지에 대한 값을 나타낸 것이 <Figure 5>이다. 즉, 토픽 1,2,3 만으로 예시를 들면, 1번 사용자(user 1)는 토픽 3(topic 3)에 해당하는 주제에 대해 가장 많이 검

색하였고 큰 관심을 가지고 있음을 뜻한다. 또한 정도의 차이는 존재하지만 1번 사용자와 2번 사용자는 토픽 1과 2에 대해 공통된 관심사를 가지고 있다고 볼 수 있다.

3.2 유사도 계산

협업필터링 기반 추천시스템 연구에서 사용자의 유사성을 계산하기 위한 유사도 측정 방법은 다양하며, 변수 및 매트릭스의 유형(type)별로 적합한 측정 방법을 사용해야 한다(Kim et al., 2012). 본 연구에서는 이진형(binary), 연속형(continuous), 혼합형(mixed) 프로파일에 적합한 유사도를 사용하여 사용자 간 유사성을 계산하였다.

$$J(A,B)=\frac{|A\bigcap B|}{|A\bigcup B|}=\frac{|A\bigcap B|}{|A|+|B|-|A\bigcap B|} \quad (2)$$

식 (2)는 매트릭스의 값이 1 또는 0과 같은 이진형(binary)으로 이루어져 있을 경우, 유사도를 계산 하는 방법인 Jaccard 인덱스이다(Niwattanakul, 2013). 이는 사건의 발생을 기준으로 유사도를 계산하는 방법이라고 할 수 있다. 즉 A 또는 B가 출현한 모든 사건들 중에서 A와 B가 동시에 출현한 사건들의 비율을 뜻하는데, 본 연구의 rating profile 에서는 A와 B가 각각 서로 다른 사용자를 의미하며, 출현한 사건은 상품에 대한 클릭 여부가 된다. 예를 들어 1번 사용자(user 1)는 '가공식품', '노트북/컴퓨터'를 클릭한 적이 있고, 2번 사용자(user 2)는 '가공식품', '건강식품', '여성의류' 카테고리를 클릭한 적이 있다고 할 때, jaccard 인덱스 방법을 사용하면 1번과 2번 사용자가 클릭한 모든 카테고리 중에서 동시에 클릭한 카테고리들의 비율을 계산한 것이다. 즉, 모든 카테고리 '가공식품', '노트북/컴퓨터', '건강식품', '여성의류' 카테고리 중에서 두 사용자 모두 클릭한 적이 잇는 '가공식품' 카테고리의 비율, 1/5 = 0.2가 1번 사용자와 2번 사용자 간의 유사도가 된다.

$$\omega=\cos(\vec{i},\vec{j})=\frac{\vec{i}\ \bullet\ \vec{j}}{\|\vec{i}\|\ *\ \|\vec{j}\|} \quad (3)$$

cosine 유사도는 내적공간의 두 벡터 간 각도의 코사인 값을 이용하여 벡터간의 유사한 정도를 측정하는 방법으로(Linden et al., 2003), 식 (3)을 통해 계산된다. 여기서 i와 j는 서로 다른 사용자이며, 사이트 체류시간의 비율, 토픽 관여도 등이 벡터로 사용될 수 있다. 유사도 결과 값은 0부터 1사이의 값으로 나타나는데, 1에 가까울수록 비슷한 성향을 가지고 0이면 다른 성향을 지니고 있다는 것을 의미한다. 선호 사이트 프로파일을 예로 들면, 같은 카테고리에 한해, 비슷한 체류시간을 보낸 사용자들의 코사인 유사도 값이 높게 나올 것이다. 이는 두 사용자의 선호 사이트가 유사하며 비슷한 성향을 가진다는 것을 의미한다.

$$s_{ij}=\frac{\sum_k w_{ijk}s_{ijk}}{\sum_k w_{ijk}} \quad (4)$$

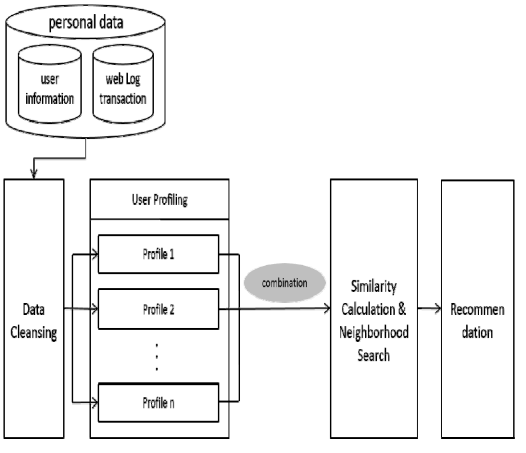
gower 유사도는 혼합형(mixed) 데이터의 유사도를 측정하는 가장 널리 알려진 방식이며(Gower, 1971), 계산은 식 (4)와 같다. 여기서 i와 j는 서로 다른 사용자이다. 또한, s_{ijk}는 k번째 변수에 대한 기여도를 의미하고 w_{ijk}는 k번째 변수의 유효성 여부에 따라 1또는 0으로 출력된다.

식 (3)의 분모의 $\sum_k w_{ijk}$는 변수들의 유사성의 합으로 나눠진다. 본 연구의 인구통계학적 프로파일을 예로 들면, 연속형 변수인 '나이'와 범주형 속성인 '직업' 등 서로 다른 유형의 변수들 각각에 대한 기여도 및 유효성을 바탕으로 사용자 간의 유사 정도를 계산하는 것이다.

본 실험에서는 위와 같이 각 프로파일에 맞는 유사도 계산 방법을 이용하여, 이진형(binary) 프로파일은 jaccard 유사도, 연속형(continuous) 프로파일은 cosine 유사도, 혼합형(mixed) 프로파일은 gower 유사도를 이용하여 사용자 간 유사도를 계산하였다.

3.3 앙상블 방법론

본 연구에서 제안하는 앙상블 방법론은 크게 2가지로 나뉜다. 첫째는 프로파일 전체를 통합하여 하나의 유사도를 계산하는 기법이고, 둘째는 개별 프로파일별 유사도를 계산하여 결합하는 기법이다. 이 유사도 결합 기법은 또 다시 2가지로 방법으로 나뉘는데, 하나는 단순 평균으로 유사도를 결합하는 방법이고 또 다른 하나는 가중평균으로 유사도를 결합하는 방법이다. 이러한 3가지 방법론은 각각 프로파일 통합 유사도, 개별 프로파일 유사도 평균, 개별 프로파일 유사도 가중평균 기법이다.



〈Figure 6〉 Process of ensemble methodology 1

[Combined profile]

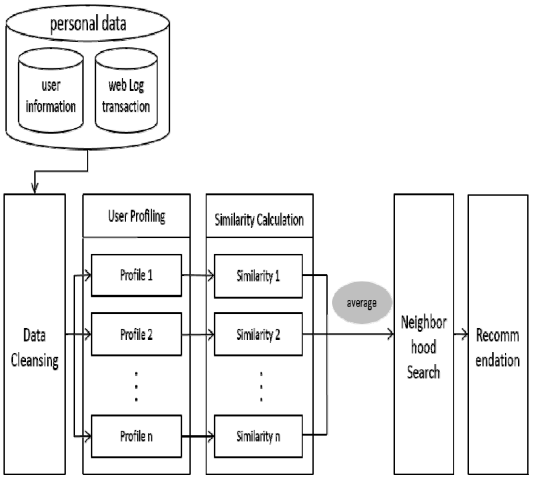
	processed food	furniture/ DIY	health food	portal	shopping	news	gender	age	occupation
user 9	1	0	0	0.093	0.059	0.430	Female	23	student
user17	1	1	1	0.001	0.323	0	Male	49	public official
user 81	0	0	0	0.280	0.028	0	Male	54	self-employed
user 134	0	1	0	0.001	0	0.001	Female	21	student
user 287	0	0	0	0.216	0.253	0.198	Male	35	public official
user402	0	0	0	0.027		0.315	Female	18	Student

	user 2	user 3	user 4	user 5
user 1	0.0333	0	0.4221	0.0931
user 2	0	0	0.1922	0.0019
user 3	0.1777	0.0333	0.0010	0.2803
user 4	0.3114	0.1556	0.0333	0.0011

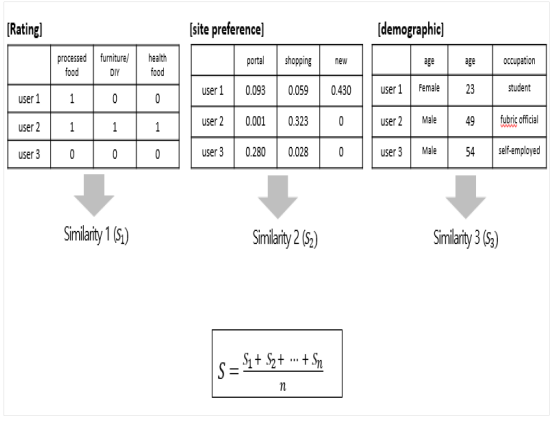
〈Figure 7〉 Example of ensemble methodology 1

본 절에서는 3가지의 제안 앙상블 방법론에 대해 설명하고 프로세스를 정의하고자 한다. 먼저, 프로파일 통합 유사도 기법은 퍼스널 빅데이터를 정제하여 개발한 다중 프로파일을 하나의 전체 프로파일로 결합하여, 결합한 프로파일로부터 하나의 유사도를 계산하는 방법이다. 프로세스는 <Figure 6>과 같으며, 데이터 예시는 <Figure 7>과 같다.

두 번째로, 개별 프로파일 유사도 평균 기법은 앞서 언급했듯이, 프로파일 결합이 아닌 유사도 결합이라는 점에서 프로파일 통합 유사도 기법과 큰 차이가 있다. 프로세스는 <Figure 8>과 같다. 퍼스널 빅데이터로부터 생성한 다중 프로파일을 기반으로 각각의 프로파일별 유사도를 계산하여 이를 단순 평균하며, 도출된 평균 유사도를 이용하여 이웃을 탐색하고 상품을 추천하는



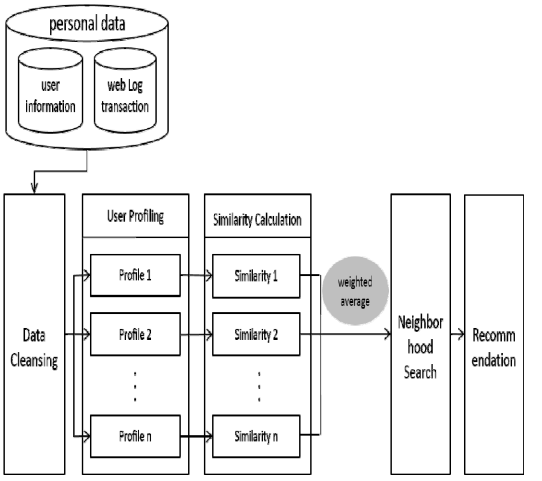
〈Figure 8〉 Process of ensemble methodology 2



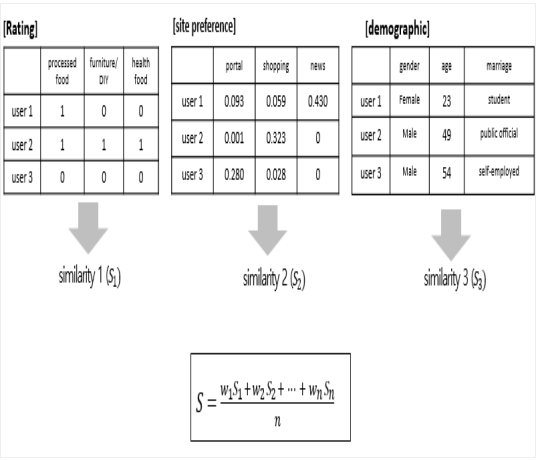
〈Figure 9〉 Example of ensemble methodology 2

방법이다. 이에 대한 데이터 예시는 <Figure 9>와 같다.

마지막으로 개별 프로파일 유사도 가중 평균 기법은 개별 프로파일 유사도 평균 기법과 유사하지만, 유사도를 결합할 때 단순 평균이 아닌 프로파일별 가중치를 주어 앙상블한다는 점에서 차이점이 있다. 가중치는 각 프로파일별로 부여하는데, 각 단일 프로파일의 추천 성능을 나타내



〈Figure 10〉 Process of ensemble methodology 3



〈Figure 11〉 Example of ensemble methodology 3

는 F1 값을 사용한다. 즉, 단일 프로파일의 F값과 해당 프로파일의 유사도를 각각 곱한 후, 이를 평균하여 가중 평균 유사도를 구한다. 개별 프로파일에 대한 가중 평균 유사도를 기반으로 이웃을 탐색하고 상품을 추천하는 방법이다. 프로세스는 <Figure 10>과 같으며 데이터 예시는 <Figure 11>과 같다.

4. 실험 및 평가

4.1 실험 데이터

본 실험에 사용된 데이터는 국내 한 인터넷 사이트 순위 분석 전문 업체의 패널 1000명에 대해 수집한 1년(2012년 7월 1일 ~ 2013년 6월 30일) 동안의 웹로그 트랜잭션 데이터 및 사용자 정보 자료이다. 웹로그 트랜잭션 데이터의 경우, 해당 사용자가 웹에 접속한 시간과 접속한 사이트, 그리고 해당 사이트에서의 체류시간, 검색 키워드 등이 건수별로 저장되어 있는 기록이며, 사용자 정보 자료는 1000명의 사용자에 대한 기본 인구통계학적 정보가 담겨있는 자료이다. 본 연구에서는 패널들이 클릭한 온라인 쇼핑몰의 상품 정보를 추가적으로 얻기 위해, 웹로그 정보 중 국내 인터넷 쇼핑몰 'A사' 사이트의 URL을 추출하였다. 그 후 크롤링을 통해 각 UFL에 해당하는 상품의 이름, 가격, 상품 카테고리 등의 정보를 수집하였다.

4.2 실험 방법

실험을 위해 우선 인터넷 쇼핑몰 사이트 'A사'에 대한 방문 데이터를 트랜잭션 기준으로 5-fold로 분할하였으며, 훈련용 데이터셋(train set)과 평가용 데이터셋(test set)은 각각 60%와 40%로 설정하여 분석하였다. 따라서, 신뢰도 높은 결과를 산출하기 위해 반복 분석을 실시하였고 그 결과를 평균한 값을 추천 성능 지표로 활용하였다. 본 실험에서는 제 3장에서 제안한 방법론을 훈련용 데이터셋(train set)에 적용하여 이웃을 선정하고, 예측한 추천 목록 중 N개의 최상위 상품을 추천하는 Top-N List 방식을 이용하였다. 이러한 추천 결과를 평가용 데이터셋(test set)의 실제 데이터와 비교하여 추천 성능을 측정하였다. 검색 키워드 토픽 프로파일을 생성하기 위한 토픽 분석은 SAS E-miner를 활용하였고, 데이터 전처리 및 추천시스템 구현에는 오픈 소스 분석 툴인 R을 사용하였다.

4.3 평가 방법

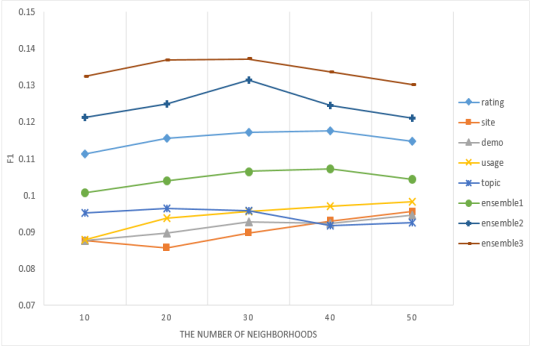
본 연구에서 추천 시스템 성능은 다수의 선행연구(Herlocker et al., 2004; Billsus and Pazzani, 1998)에서 사용한 F1 지표를 이용하여 평가하였다. 식 (4)에서 Precision은 추천한 전체 아이템 개수 중 실제 구매한 아이템의 개수를 나눈 비율이고, Recall은 실제 구매한 전체 아이템 개수 중에서 추천된 아이템 개수를 나눈 비율이다. 따라서, F1은 Precision과 Recall 값을 동일한 가중치를 주고 모두 반영하여 하나의 지표로 정확성을 산출하기 위해 고안되었다. 지표는 0부터 1까지의 값을 가지며 값이 높을수록 예측 정확도가 높음을 의미한다.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (4)$$

4.4 실험 결과

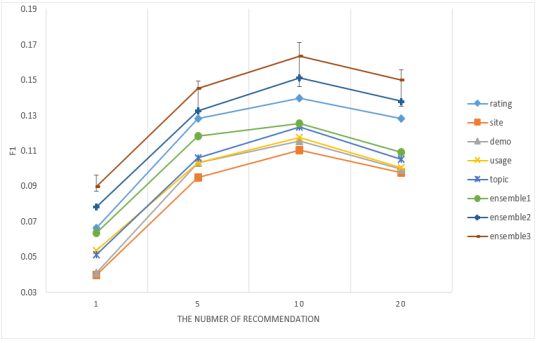
앞 절에서 설명한 F1 지표를 활용하여, 본 연구에서 제안한 프로파일 통합 유사도, 개별 프로파일 유사도 평균 개별 프로파일 유사도 가중 평균 앙상블 방법론을 적용한 추천시스템 성능 평가를 실시하였다. 추천 시스템의 성능은 이웃의 수와 추천 상품의 수에 따라 민감하게 달라진다. 이 두 가지 변수의 효과를 살펴보기 위해 이웃의 수는 10, 20, 30, 40, 50명으로 변화시키고, 추천

상품의 수는 1, 5, 10, 20개로 증가시키면서 성능을 측정하였다. 먼저 이웃의 수 변화에 따른 성능 비교를 알아보기 위해 각 이웃의 수에 대해 1, 5, 10, 20개 상품을 추천했을 때의 F1값을 나타냈으며, 그 결과는 <Figure 12>과 같다.



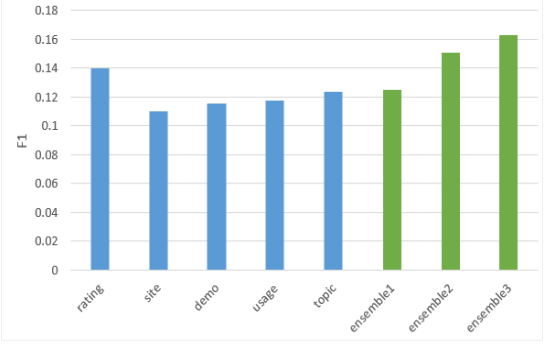
〈Figure 12〉 Performance comparison on the number of neighborhoods

<Figure 12>의 범례에서 위의 5개(rating, site, demo, usage, topic)은 본 연구에서 생성한 rating, 선호 사이트, 인구통계학, 인터넷 사용행태, 검색 키워드 토픽 프로파일을 의미하며, 각 단일 프로파일에 대한 F1값을 나타낸다. 그 다음 3가지(ensemble 1,2,3)는 본 연구에서 제안한 프로파일 통합 유사도, 개별 프로파일 유사도 평균, 개별 프로파일 유사도 가중 평균 방법을 각각 적용한 F1값을 뜻한다. 그림을 보면 가장 높은 성능을 보이고 있는 것은 앙상블 3인 유사도 가중 평균 기법이며, 두 번째로 높은 성능을 보이는 것은 앙상블 2인 유사도 평균 기법이다. 그림의 전체적인 꺾은선 그래프 추이를 보면 F1값이 어느 정도 증가하였다가 감소하는 추세를 보이고 있으며, 대부분 이웃의 수가 20일 때 성능이 가장 높은 편이라고 판단된다.



〈Figure 13〉 Performance comparison on the number of recommendation

<Figure 13>은 추천 상품 수 변화에 따른 성능 비교를 알아보기 위해, 성능이 가장 높았던 20명의 이웃으로 상품을 추천했을 때의 F1값을 나타낸 그래프이다. 그림을 보면 앞의 <Figure 12>과 같이, F1값이 증가하다가 감소하는 추세로 바뀌는데, 10개의 상품을 추천했을 때 모든 프로파일과 방법론이 가장 높은 성능을 보인 것을 알 수 있다.



〈Figure 14〉 Performance comparison of each profile and ensemble recommendation

<Figure 14>은 앞의 두 번의 실험결과를 토대

로, 20명의 이웃 수와 10개의 추천 상품 수를 적용한 성능을 나타낸 그래프이다. 왼쪽의 5개 막대 그래프(rating, site, demo, usage, topic0는 5종의 단일 프로파일에 대한 성능이며, 나머지 3개 막대 그래프(ensemble 1,2,3)는 앙상블 방법론 3가지를 적용한 것에 대한 성능을 의미한다. 실험결과, 단일 프로파일 중에서는 rating profile이 0.139로 가장 우수한 성능을 보였다. 본 연구에서 제안한 앙상블 방법에 대한 성능을 살펴보면, 먼저 프로파일 통합 유사도 기법의 경우 5종의 단일 프로파일 성능의 평균치 수준으로, rating profile보다는 낮은 성능을 보였다(0.125). 반면, 개별 프로파일 유사도 평균과 가중 평균 방법은 각각 0.150와 0.163로 단일 프로파일보다 높은 성능을 나타냈으며, 특히 개별 프로파일 유사도 가중 평균 기법이 가장 높은 성능을 보였다. rating profile의 성능과 비교했을 때, 개별 프로파일 유사도 가중 평균 기법은 16.85%, 개별 프로파일 유사도 평균 기법은 8.1%의 성능 개선이 이루어지는 것으로 나타났다.

5. 결론

본 연구는 두 가지 측면에서 의의가 있다. 첫째, 마케팅의 시장세분화 이론을 이용하여 실제 퍼스널 빅데이터로부터 고객의 선호도나 행태를 다양한 관점에서 표현할 수 있는 5종의 프로파일을 개발하였다. 이는 기존의 다중 프로파일 기반 연구와 비교하여, 단순히 인구통계학적 정보나 상품 정보와 같이 고객의 선호도나 특성에만 제한되는 것이 아니라 실제 고객의 행태를 마케팅적 시각에서 보다 구체적으로 규명하였다는 점에서 의의가 있다. 또한 빅데이터 환경에서 추천시스템을 개발하고자 할 때 어떠한 정보를 이용하여 고객의 특성을 규명하는 프로파일을 만드는 것이 효과적인 지 제안하였다는 점에서 가치가 있다고 할 수 있다. 둘째, 프로파일 및 유사도를 앙상블하여 협업필터링의 이웃 선정 과정에 적용할 수 있는 방법론을 제안하였다. 이를 통해 다중 프로파일을 어떻게 결합하여 사용하는 것이 효과적인지를 제시하였다는 점에서 의의가 있다. 또한, 본 연구는 새로운 데이터를 기반으로 기존 알고리즘을 적용한 앙상블 연구이므로, 향후 알고리즘 개선 연구에 본 방법론을 결합하여 추가적인 추천 성능 향상을 기대할 수 있다.

반면, 본 연구에서는 퍼스널 빅데이터를 웹로그 데이터로 한정지어 실험하였고 인터넷 행태를 기반으로 사용자의 속성을 규명했다는 점에서 한계가 있다. 따라서 향후 연구에서는 소셜 데이터, 위치 데이터 등 다양한 유형의 퍼스널 빅데이터를 정의하고 동일한 방법론으로 실험하여 추천 성능을 검증해 볼 필요가 있다.

본 연구에서는 다양한 추천시스템 알고리즘 중에서 사용자 기반 협업필터링 알고리즘에만 제시한 방법론을 적용하여 평가하였다. 따라서 다양한 추천 알고리즘에 적용할 경우에도 추천 성능이 향상되는지 분석하여, 어떠한 알고리즘 또는 어떻게 앙상블된 알고리즘에 적용하는 것이 최적인지를 찾기 위한 추가연구가 필요하다.

참고문헌(References)

Bar, A., G. Rokach, G. Shani, B. Shapira, and A. Schclar, “Improving simple collaborative filtering models using ensemble methods,”

Multiple Classifier Systems, Springer, (2013), 1～12.

Billsus, D. and M. J. Pazzani, "Learning Collaborative Information Filters," *ICML*, Vol.98, (1998), 46～54.

Bok, K. S. and J. S. Yoo, "Activation Policy and Case Study of Big Data," *The Journal of Korean Institute of Communication Sciences*, Vol.31, No.11(2014), 3～13.

Cabral, B., R. D. Beltro, and M. G., Manzato, "Combining Multiple Metadata Types in Movies Recommendation Using Ensemble Algorithms," *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, (2014), 231～238.

Claycamp, H. J. and W. F. Massy, "A Theory of Market Segmentation," *Journal of Marketing Research*, Vol.5, No.4(1968), 388～394.

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry, "Using Collaborative filtering to weave an information Tapestry," *Communications of the ACM*, Vol.35, No.12(1992), 61～70.

Gower, J. C., "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, Vol.27, No.4(1971), 857～871.

Gurrin, C., A. F. Smeaton, and A. R. Doherty, "LifeLogging: Personal Big Data," *Foundations and Trends in Information Retrieval*, Vol.8, No.1(2014), 1～107.

Herlocker, J. L., J. A. Konstan, and J. Riedl, "An algorithmic framework for performing collaborative filtering," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), 230～237.

Hyun, Y., N. Kim, and Y. Cho, "Interest-based Customer Segmentation Methodology Using Topic Modeling," *Journal of Information Technology Applications & Management*, Vol.22, No.1(2015), 77～93.

Herlocker, J. L., J. A. Konstan, L. G., Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems*, Vol.22, No.1(2004), 5～53.

Kim, N.-H., "A Study on the Improvement of Web-log Analysis in Internet shopping-Mall," *Proceedings of Korea Intelligent Information System Society*, (2002), 134～139.

Kim, J.-H., B.-H. Ahn, and D. Jeong, "A Recommender System using Mixed Filtering for Health Products," *The Journal of Internet Electronic Commerce Research*, Vol.12, No.2(2012), 109～124.

Kim, K. H. and S. R., Oh, "Methodology for Applying Text Mining Techniques to Analyzing Online Customer Reviews for Market Segmentation," *Journal of the Korea Contents Association*, Vol.9, No.8(2009), 272～284.

Kim, Y., J. Moon, H. J. Lee, and C. S., Bae, "Knowledge Digest Engine for Personal Bigdata Analysis," *Human Centric Technology and Service in Smart Space*, Springer Netherlands, 2012.

Lee, J. S. and S. D. Park, "Performance Improvement of a Movie Recommendation System using Genre-wise Collaborative Filtering," *Journal of Intelligence and Information Systems*, Vol.13, No.4(2007), 65～78.

Lee, Y. and K.-j. Kim, "Product Recommender Systems using Multi-Model Ensemble Techniques,"

Journal of Intelligence and Information Systems, Vol.19, No.2(2013), 39～54.

Linden, G., B. Simth, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing,* Vol.7, No.1(2003), 76～80.

Mazanec, J. A. "Market Segmentation," J. Jafari(Ed), *Encyclopedia of Tourism*, London:Routledge, 2000.

Middleton, S. E., Shadbolt, N. R., and De Roure, D. C., "Ontological User Profiling in Recommneder Systems," *ACM Transactions on Information Systems*, Vol.22, No.1(2004), 54～88.

Niwattanakul,S., J. Singthongchai, E. Naenudorn, and S, Wanapu, "Using Jaccard Coefficient for Keywords Similarity," *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol.1(2013).

Park, Y.-J., E.-J. Jung, and K.-N. Chang, "Customer Behavior Based Customer Profiling Technique for Personalized Products Recommendation," *Korean Management Science Review*, Vol.23, No.3(2006), 183～194.

Pazzani, M., "A Framework for Collaborative, Content-Based, and Demographic Filtering," *Artificial Intelligence Review*, Vol.13, No.5-6(1999), 393～408.

Piotte, M. and M. Chabbert, "The Pramatic theory solution to the Netflix grand prize," *Netflix prize documentation*, 2009.

Ward, J. S. and A. Barker, "Undefined By Data: A Survey of Big Data Definitions," *The Computing Research Repository*, 2013.

Weng, S. S. and M. J. Liu, "Feature-based recommendations for one-to-one marketing," *Expert Systems with Applications*, Vol.26, No.4(2004), 493～508.

Abstract

A Multimodal Profile Ensemble Approach to Development of Recommender Systems Using Big Data

Minjeong Kim* · Yoonho Cho**

The recommender system is a system which recommends products to the customers who are likely to be interested in. Based on automated information filtering technology, various recommender systems have been developed. Collaborative filtering (CF), one of the most successful recommendation algorithms, has been applied in a number of different domains such as recommending Web pages, books, movies, music and products. But, it has been known that CF has a critical shortcoming. CF finds neighbors whose preferences are like those of the target customer and recommends products those customers have most liked. Thus, CF works properly only when there's a sufficient number of ratings on common product from customers. When there's a shortage of customer ratings, CF makes the formation of a neighborhood inaccurate, thereby resulting in poor recommendations. To improve the performance of CF based recommender systems, most of the related studies have been focused on the development of novel algorithms under the assumption of using a single profile, which is created from user's rating information for items, purchase transactions, or Web access logs. With the advent of big data, companies got to collect more data and to use a variety of information with big size. So, many companies recognize it very importantly to utilize big data because it makes companies to improve their competitiveness and to create new value. In particular, on the rise is the issue of utilizing personal big data in the recommender system. It is why personal big data facilitate more accurate identification of the preferences or behaviors of users.

The proposed recommendation methodology is as follows: First, multimodal user profiles are created from personal big data in order to grasp the preferences and behavior of users from various viewpoints. We derive five user profiles based on the personal information such as rating, site preference, demographic, Internet usage, and topic in text. Next, the similarity between users is calculated based on the profiles and

* Department of Data Science, Kookmin University
** Corresponding Author: Yoonho Cho
School of Business Administration, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea
Tel: +82-2-910-4950, E-mail: www4u@kookmin.ac.kr

then neighbors of users are found from the results. One of three ensemble approaches is applied to calculate the similarity. Each ensemble approach uses the similarity of combined profile, the average similarity of each profile, and the weighted average similarity of each profile, respectively. Finally, the products that people among the neighborhood prefer most to are recommended to the target users.

For the experiments, we used the demographic data and a very large volume of Web log transaction for 5,000 panel users of a company that is specialized to analyzing ranks of Web sites. R and SAS E-miner was used to implement the proposed recommender system and to conduct the topic analysis using the keyword search, respectively. To evaluate the recommendation performance, we used 60% of data for training and 40% of data for test. The 5-fold cross validation was also conducted to enhance the reliability of our experiments. A widely used combination metric called F1 metric that gives equal weight to both recall and precision was employed for our evaluation. As the results of evaluation, the proposed methodology achieved the significant improvement over the single profile based CF algorithm. In particular, the ensemble approach using weighted average similarity shows the highest performance. That is, the rate of improvement in F1 is 16.9 percent for the ensemble approach using weighted average similarity and 8.1 percent for the ensemble approach using average similarity of each profile. From these results, we conclude that the multimodal profile ensemble approach is a viable solution to the problems encountered when there's a shortage of customer ratings.

This study has significance in suggesting what kind of information could we use to create profile in the environment of big data and how could we combine and utilize them effectively. However, our methodology should be further studied to consider for its real-world application. We need to compare the differences in recommendation accuracy by applying the proposed method to different recommendation algorithms and then to identify which combination of them would show the best performance.

Key Words : Big Data, Recommender System, Collaborative Filtering, Multimodal Profile, Ensemble Methodology

Received : November 25, 2015 Revised : December 14, 2015 Accepted : December 14, 2015
Corresponding Author : Yoonho Cho

저 자 소 개

김 민 정
현재 국민대학교 데이터사이언스학과 석사과정에 재학 중이며, 한국산업기술대학교 e-비즈니스학과에서 학사 학위를 취득하였다. 주요 관심분야는 고객관계관리(CRM), 마케팅 애널리틱스, 데이터 마이닝, 추천시스템 등이다.

조 윤 호
현재 국민대학교 경영학부 빅데이터경영통계전공 교수로 재직 중이다. 서울대학교 계산통계학과를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 비즈니스애널리틱스, 빅데이터 마이닝, 추천시스템, 소셜네트워크분석, 고객관계관리 등이다.