ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 September: 21(3): 101~116 http://dx.doi.org/10.13088/jiis.2015.21.3.101

사용자 관심 이슈 분석을 통한 추천시스템 성능 향상 방안

최성이
국민대학교 비즈니스IT전문대학원
(csy0000@kookmin.ac.kr)

현윤진
국민대학교 비즈니스IT전문대학원
(yoonjin0630@koojmin.ac.kr)

김남규
국민대학교 경영대학 경영정보학부
(ngkim@kookmin.ac.kr)

많은 기관들이 데이터에 기반을 둔 의사결정을 수행해 왔으며, 특히 수치자료를 비롯한 정형 데이터가 이러한 목적으로 널리 활용되어 왔다. 하지만 최근에는 스마트기기와 소셜미디어의 발달로 인해 다양한 형태를 가진 방대한 양의 정보가 생성, 공유, 저장되면서, 전통적인 정형 데이터 기반 의사결정으로부터 비정형 빅데이터 기반 의사결정으로 관심의 전환이 이루어지고 있다. 데이터 기반 의사결정의 대표적 분야인 추천시스템 분야에서도 성능 향상을 위해 비정형 데이터를 활용해야 한다는 필요성이 최근 꾸준히 제기되고 있다. 특히 사용자의 성향이나 선호도는 고객의 니즈와 직결되기 때문에, 비정형 데이터 분석을 통해 사용자의 성향을 파악하고 이를 통해 상품 추천 및 구매 예측의 정확도를 향상시키기 위한 노력이 매우 시급하게 이루어질 필요가 있다. 따라서 본 연구에서는 사용자의 성향을 측정하여 재구매 예측 정확도, 특히 카테고리별 재구매 예측 정확도를 높임으로써, 궁극적으로 추천시스템의 성능을 향상시킬 수 있는 방안을 제시한다. 구체적으로는 사용자의 일상적인 인터넷 사용 기록을 분석하여 고객이 조회하는 뉴스 기사의 이슈를 식별하고 다양한 이슈에 대한 고객의 관심을 계량화한 후, 이를 활용하여 고객의 카테고리별 재구매 여부를 예측하는 모델을 제안하고자 한다. 실제 웹 트랜잭션으로부터 도출된 인터넷 뉴스 조회 기록 및 쇼핑몰 구매 기록을 대상으로 실험을 수행한 결과, 고객의 과거 구매이력만을 활용한 카테고리 재구매 예측 모형에 비해 본 연구에서 제안한 모형, 즉 고객의 과거 구매이력과 관심 이슈를 모두 활용한 예측 모형의 정확도가 다소 우수한 것으로 나타났다.

주제어 : 데이터 마이닝, 빅데이터 분석, 추천시스템, 텍스트 마이닝, 토픽 분석

논문접수일 : 2015년 8월 31일 논문수정일 : 2015년 9월 9일 게재확정일 : 2015년 9월 9일
투고유형 : 국문일반 교신저자 : 김남규

1. 서론

많은 기관들이 데이터에 기반을 둔 의사결정을 수행해 왔으며, 특히 수치자료를 비롯한 정형 데이터가 이러한 목적으로 널리 활용되어 왔다. 실제로 데이터에 기반을 둔 의사결정은 비구조적인 문제를 해결하는데 도움을 주며, 이를 통해 약 5 ~ 6%의 생산성을 향상시킬 수 있음이 Brynjolfsson et al.(2011)에 의해 밝혀진 바 있다. 특히 최근에는 스마트기기와 소셜미디어의 발달로 인해 다양한 형태를 가진 방대한 양의 정보가 유통되면서 빅데이터(Big Data)를 활용하여 새로운 방식으로 사회 문제를 해결하기 위한 연구가 매우 활발하게 이루어지고 있다(Kang et al., 2012). 이로 인해 다양한 형태의 텍스트 콘텐츠, 멀티미디어 콘텐츠 등을 효과적으로 처리하기 위한 다양한 기술이 고안되어 왔으며, 그 결과 전통적인 정형 데이터 기반의 의사결정으로부터

비정형 빅데이터 기반 의사결정으로의 관심 전환이 이루어지게 되었다. 예를 들면 소셜미디어를 기반으로 사용자의 여론 및 감성을 분석하는 시스템인 소셜위즈덤에 관한 연구(Heo et al., 2013), 여러 유형의 재난으로 인한 위험을 사전에 방지하는 재난관리 시스템에 관한 연구(Min and Jeong, 2013), 주식 시장에 관한 뉴스 분석에 근거하여 주가지수의 방향을 예측한 연구(Kim et al., 2012; Yu et al., 2013), 텍스트 분석에 근거하여 사회적 이슈를 다양한 관점에서 재구조화한 연구(Hyun et al., 2015; Kim et al., 2014), 토픽 분석을 통해 웹 카테고리별 방문자의 주요 관심 이슈를 식별한 연구(Choi and Kim, 2014) 등이 있다. 이러한 연구들은 빅데이터에 기반을 둔 의사결정 지원 시스템이 민간분야뿐 아니라 공공, 경제, 위기관리 등 다양한 분야의 문제 해결에 효과적으로 사용될 수 있음을 시사한다.

이러한 다양한 응용 중 추천시스템은 데이터 기반 의사결정의 대표적 분야 중 하나로 인식되고 있으며, 특히 인터넷 쇼핑몰의 상품 추천에 대한 연구가 활발하게 이루어져 왔다. 대표적인 예로 방문객과 유사한 선호 체계를 가진 다른 사용자들의 고객 정보 및 구매이력을 활용하여 새로운 아이템을 추천(Funakoshi and Ohguro, 2000)한 아마존닷컴(Amazon.com)의 예를 들 수 있으며, 최근에는 추천시스템의 성능 향상을 위해 비정형 빅데이터 분석을 활용하는 시도(Choi and Hwang 2012)도 많은 주목을 받고 있다. 추천시스템의 성능 향상을 위한 기존의 노력은 크게 (1)알고리즘 개선과 (2)양질의 데이터 확보라는 두 가지 측면에서 모색되어 왔다. 알고리즘을 개선을 통해 추천시스템의 성능을 고도화 시키려는 노력은 매우 다양한 형태로 활발하게 이루어져 왔지만, 이에 비해 새로운 데이터 확보를 통한 추천시스템 성능 향상에 대한 연구는 상대적으로 부족했던 것으로 파악된다. 이러한 관점에서 전술한 바와 같이 최근 다양한 경로를 통해 축적되는 비정형 빅데이터를 가공하고 활용하여 추천시스템의 성능을 향상시키기 위한 시도는 매우 시의적절하며 반드시 필요한 시도인 것으로 보인다. 특히 사용자의 성향이나 선호도는 고객의 니즈와 직결된다는 점을 감안할 때 빅데이터 분석을 통해 사용자의 성향을 파악하고, 이를 통해 재구매를 예측하고 상품을 추천하기 위한 연구가 반드시 이루어져야 한다.

이러한 필요에 기인하여 본 연구에서는 사용자의 관심을 측정하여 추천시스템의 성능을 향상시킬 수 있는 방안을 제시하고자 한다. 하지만 현재 인터넷 쇼핑몰에서 얻을 수 있는 사용자 성향 정보는 사용자가 사이트 등록시 임의로 입력한 취미와 관심분야 등에 국한된다는 한계가 있다. 이러한 정보는 객관성이 보장되지 않을 뿐 아니라 충분히 세분화되어 있지 않고 갱신이 수시로 이루어지지 않기 때문에, 이러한 정보를 통해 사용자의 최근 관심을 파악하기엔 어려움이 있다. 따라서 본 연구에서는 사용자의 일상적인 인터넷 사용패턴을 분석하여 다양한 분야에 대한 고객의 성향을 계량화하고, 이를 활용하여 고객의 재구매를 예측함으로써 궁극적으로 추천시스템의 성능을 향상시킬 수 있는 방안을 제시하고자 한다.

이후 본 논문의 구성은 다음과 같다. 다음 장인 2장에서는 제안 기법과 관련된 기존 연구를 요약한다. 3장에서는 고객의 성향을 계량화하는 방안을 소개하고, 4장에서는 제안 방법론을 실제 데이터에 적용한 실험을 통해 제안 방법론의 성능을 평가한다. 마지막인 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

기존 추천시스템 관련 연구의 상당히 많은 수가 사용자의 연관성을 기반으로 상품을 추천하는 협업필터링(Collaborative Filtering)(Billsus and Pazzani, 1998)에 근거하여 수행되어 왔다. 이들 연구는 일반적으로 방문객과 사용자들의 기호정보(Taste Information)를 측정한 뒤, 이를 기반으로 고객군의 과거 행동에 근거하여 관심사나 취향을 예측한다(Jeong et al., 2015). 이 외에도 상품의 속성 정보를 활용한 내용기반(Content-based)추천(Balabanovic and Shoham, 1997), 사용자의 인구통계학 정보, 구매이력, 웹 방문 기록 등의 분석을 통한 규칙기반(Rule-based)추천(Chun and Hong, 2001) 등이 추천시스템 연구의 큰 축을 이루고 있다. 이러한 큰 흐름 내에서 웹 마이닝, 상품 계층도, 군집 분석, 유전자 알고리즘, 신경망, 사회연결망 등을 활용하여 추천시스템의 성능을 향상시키기 위한 연구가 활발하게 이루어지고 있다(Ahn et al., 2014; Kim et al., 2005; Kim and Kim, 2005).

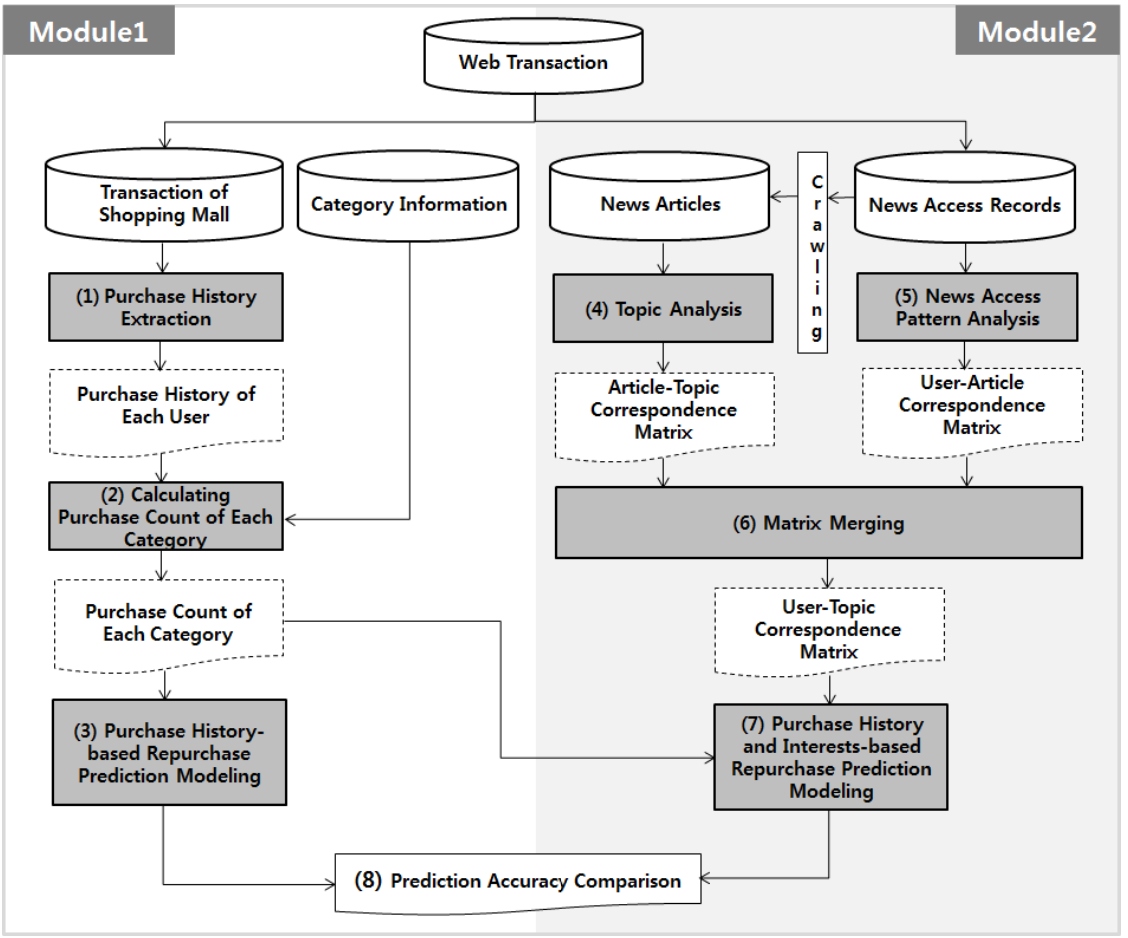
추천시스템의 성능 향상을 위해 다양한 외부 데이터의 활용 방안을 모색하는 시도가 이루어져 왔으며, 인간의 감정을 활용한 연구(Ahn, 2014), 사용자의 위치 정보를 사용한 연구(Kim and Oh, 2012), 그리고 날씨 정보(Roh et al., 2008)를 사용한 연구 등은 이러한 시도의 대표적 예라 할 수 있다. 또한 최근 다양한 채널을 통해 유통되는 텍스트 데이터의 양이 급증하고 있을 뿐 아니라 이러한 텍스트에 대한 분석 기술의 수준이 향상됨에 따라, 비정형 텍스트 데이터 분석을 통해 보다 효과적인 추천시스템을 구축하는 방안에 대한 관심이 높아지고 있다. 트윗(Tweet) 분석을 활용한 팔로워 추천시스템(Armentano et al., 2013), 라디오 신청곡 사연 문서를 활용한 음악 추천시스템(Hyung et al., 2014), 특정 도메인 상품의 사용자 리뷰를 활용한 상품 추천시스템(Aciar et al., 2006) 등이 텍스트 분석을 활용한 추천시스템의 대표적인 예이다.

텍스트 분석을 활용한 추천시스템 연구의 대부분은 텍스트 데이터를 포함하고 있는 상품 혹은 서비스에 대한 정보, 혹은 상품 및 서비스에 대한 사용자 리뷰를 기반으로 수행되어 왔다. 즉 이들 연구에서 분석한 비정형 텍스트 데이터는 해당 사이트 내부의 데이터로 국한된다. 하지만 온라인 쇼핑몰의 경우 각 상품에 대한 상품평 이외의 내부 텍스트 데이터를 찾기 어려울 뿐 아니라, 상품평은 각 상품에 초점을 두고 작성되므로 사용자의 일반적인 성향을 파악하기 위해 사용되기에는 무리가 따른다는 한계를 갖는다. 따라서 전통적으로 온라인 쇼핑몰은 사용자의 성향을 파악하기 위해 사용자가 직접 입력한 취미, 관심분야, 선호도 등의 정형 데이터를 이용해왔으나, 이들 데이터는 객관성, 구체성, 갱신 주기 등의 측면에서 사용자의 성향을 정확하게 나타내기에는 부족함이 있다.

따라서 본 연구에서는 특정 온라인 쇼핑몰 사용자의 성향을 파악하기 위해 해당 사이트 외부에서 사용자가 조회한 인터넷 뉴스에 대한 텍스트 분석을 수행하고, 이를 통해 사용자의 성향을 정형화하여 추천시스템의 성능을 향상시킬 수 있는 방안을 제시하고자 한다.

3. 사용자의 관심 이슈 분석을 통한 카테고리별 재구매 예측

3.1 연구 개요



〈Figure 1〉 Research Overview

본 절에서는 고객이 조회한 뉴스 기사를 분석하여 고객별 관심 이슈를 식별하고, 이를 활용하여 해당 고객의 상품 카테고리별 향후 재구매 여부 예측 모형을 설계한다. 본 연구의 전체 개요는 <Figure 1>과 같으며, 각 프로세스의 세부 과정은 이후 절에서 상세히 소개한다.

<Figure 1>은 크게 두 개의 Module로 구성되어 있다. Module 1의 경우 프로세스 (1)~(3)으로 구성되며, 구매이력에만 기반하여 카테고리별 재구매 여부를 예측한다. Module 1은 본 연구에서 제안하는 방법론과의 성능 비교를 위해 본 연구 개요에 포함되었다. 반면 Module 2의 프로세스 (4)~(7)은 본 연구의 핵심 프로세스를 구성하며, 고객이 조회한 뉴스에 대한 토픽 분석을 통해 각 고객별 관심 이슈를 도출하는 과정이다. 프로세스 (4)에서는 크롤링을 통해 획득한 뉴스 기사에 대해 토픽 분석을 수행하여 기사와 토픽 간 대응 매트릭스를 도출하고, 프로세스 (5)에서

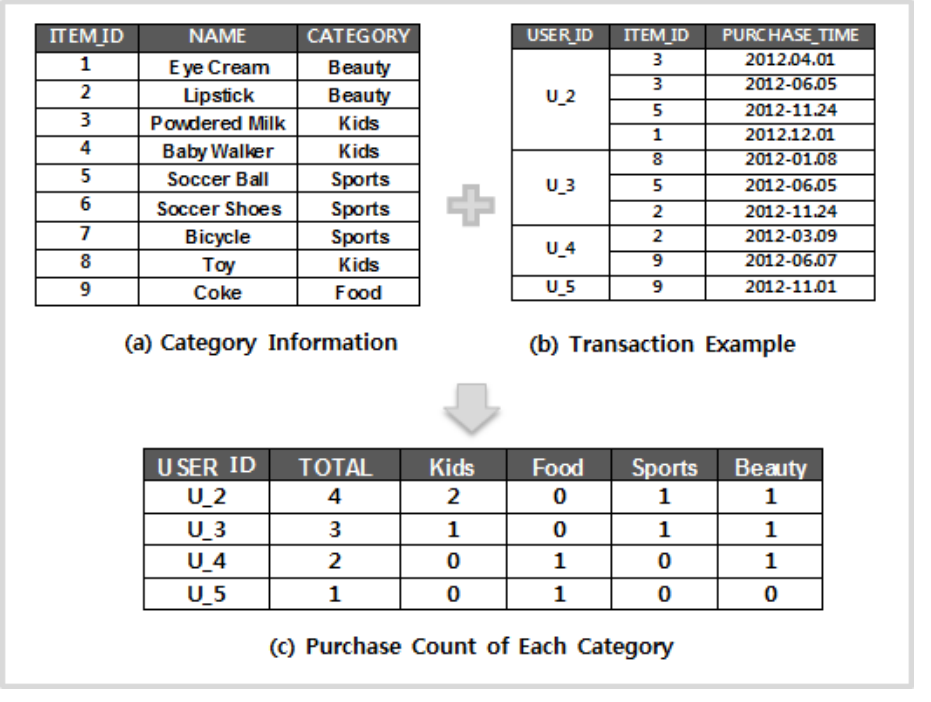
는 사용자의 뉴스 접속 기록를 분석하여 고객과 기사 간 대응 매트릭스를 도출한다. 이후 프로세스 (6)은 프로세스 (4)~(5)의 결과를 병합하여 고객과 토픽 간 대응 매트릭스를 도출함으로써 고객의 관심 이슈를 계량화하여 나타낸다. 마지막으로 프로세스 (7)은 토픽 대응도를 입력 변수로, 카테고리별 재구매 여부를 목표 변수로 정의하여 카테고리별 재구매 예측 모형을 구축한다.

3.2 순수 구매이력 기반 카테고리별 재구매 예측 모형

본 절에서는 Module 1의 프로세스 (1)~(3)에 해당되는 구매이력 기반 카테고리별 재구매 예측 모형 수립 과정을 설명한다. 우선 프로세스 (1)은 USER_ID, ITEM_ID, 그리고 PURCHASE_TIME으로 구성되는 고객별 구매이력을 생성한다. 다음으로 프로세스 (2)에서는 각 고객의 카테고리별 구매 회수를 집계한다. 이를 위해 프로세스 (1)에서 도출한 고객별 구매이력과 함께 <Figure 2(a)>와 같은 각 상품의 카테고리 정보가 사용되며, 그 결과로 도출되는 고객의 카테고리별 구매 회수는 <Figure 2(c)>와 같이 나타난다. 이를 입력 변수로 활용하여 카테고리별 재구매 예측 모형을 수립하는 과정은 4장의 실험을 통해 소개한다.

3.3 사용자의 관심 이슈를 반영한 카테고리별 재구매 예측 모형

본 절에서는 본 연구의 핵심인 <Figure 1>의



ITEM_ID	NAME	CATEGORY
1	Eye Cream	Beauty
2	Lipstick	Beauty
3	Powdered Milk	Kids
4	Baby Walker	Kids
5	Soccer Ball	Sports
6	Soccer Shoes	Sports
7	Bicycle	Sports
8	Toy	Kids
9	Coke	Food

(a) Category Information

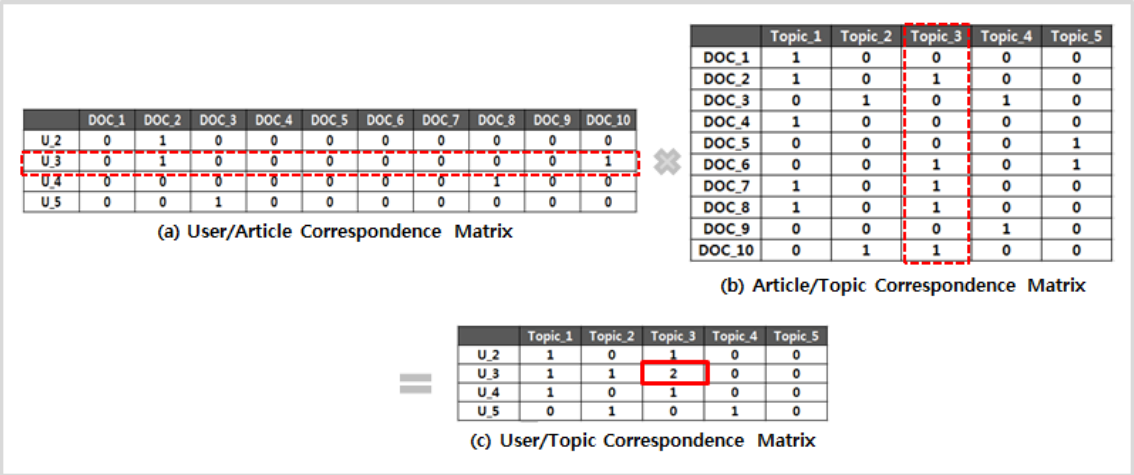
USER_ID	ITEM_ID	PURCHASE_TIME
U_2	3	2012.04.01
3	2012-06.05
5	2012-11.24
1	2012.12.01
U_3	8	2012-01.08
5	2012-06.05
2	2012-11.24
U_4	2	2012-03.09
9	2012-06.07
U_5	9	2012-11.01

(b) Transaction Example

USER ID	TOTAL	Kids	Food	Sports	Beauty
U_2	4	2	0	1	1
U_3	3	1	0	1	1
U_4	2	0	1	0	1
U_5	1	0	1	0	0

(c) Purchase Count of Each Category

〈Figure 2〉 Calculating Purchase Count of Each Category



〈Figure 3〉 Constructing User/Article Correspondence Matrix

Module 2에 대해서 설명한다. 프로세스 (4)는 수집한 뉴스 기사 전체에 대한 토픽 분석을 통해 주요 이슈을 추출하고 각 토픽과 기사의 대응 관계를 <Figure 3(b)>와 같은 행렬 형태로 나타낸다. <Figure 3(b)>에서 하나의 이슈에 다수의 문서가 포함될 수 있으며, 하나의 문서가 다수의 이슈에 대응될 수 있음을 확인할 수 있다. 토픽 분석을 통해 이슈를 추출하는 과정은 이미 많은 연구에서 소개되었을 뿐 아니라 상용화된 분석 도구를 사용하여 손쉽게 수행 가능하므로 본 논문에서는 이 과정을 자세히 다루지 않는다. 다음으로 프로세스 (5)는 사용자의 뉴스 접속 기록을 요약하여 사용자와 기사 간 행렬을 구축한다. 본 연구에서는 각 사용자별로 각 기사에 대한 조회 여부만을 기록하여 <Figure 3(a)>와 같은 이진 형태의 행렬로 나타낸다.

앞에서 도출한 <Figure 3(b)>의 기사/토픽 대응행렬과 <Figure 3(a)>의 사용자/기사 대응행렬을 병합하여 사용자/토픽 대응행렬을 도출하는 과정은 프로세스 (6)에서 수행된다. 이 과정은 실제로는 행렬곱 연산에 의해 수행되며, 수행 결과가 <Figure 3(c)>에 소개되어 있다. 예를 들어 <Figure 3>에서 사용자 U_3의 이슈 Topic_3에 대한 관심은 DOC_1부터 DOC_10까지의 10개의 뉴스 기사에 대한 조회 기록을 통해 측정될 수 있으며, 본 예에서는 Topic_3에 포함되는 기사 DOC_2, DOC_6, DOC_7, DOC_8, 그리고 DOC_10 중 U_2가 조회한 기사는 DOC_2와 DOC_10 두 개인 것으로 나타났다. 즉 <Figure 3(a)>의 U_3에 해당되는 행과 <Figure 3(b)>의 Topic_3에 해당되는 열의 벡터곱을 통해 <Figure 3(c)>에서 "2"라는 값을 도출할 수 있으며, 이 과정이 <Figure 3(a)>의 전체 행과 <Figure 3(b)>의 전체 열에 반복 적용되어 행렬곱 연산을 통해 <Figure 3(c)> 전체를 도출할 수 있다.

한편 <Figure 3(c)>에서 사용자의 이슈별 관심도는 상대적 비율이 아닌 절대 수치로 표현되어 있다. 예를 들어 U_2와 U_3이 Topic_1에 대해 갖는 관심은 모두 '1'로 나타나므로, 두 사용자의 해당 이슈에 대한 관심도는 동일한 것으로 해석

된다. 하지만 U_2의 경우 '1'이라는 값이 모든 이슈에 대해 갖는 관심의 50%에 해당되는 반면, U_3의 경우 '1'이라는 값은 모든 이슈에 대해 갖는 관심의 25%에 불과하다. 따라서 U_2는 U_3에 비해 상대적으로 Topic_1에 대해 갖는 관심이 크다고 할 수 있으며, 이러한 관점을 반영하기 위해 사용자의 이슈별 관심도는 표준화를 통해 상대적 비율로 변환되어야 한다. 표준화된 관심도는 <Figure 3(c)>의 각 셀의 값을 각 행의 총합, 즉 사용자별 관심의 총합으로 나눔으로써 도출할 수 있으며, 이 과정이 <Figure 4>에 나타나있다.

	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	TOTAL
U_2	1	0	1	0	0	2
U_3	1	1	2	0	0	4
U_4	1	0	1	0	0	2
U_5	0	1	0	1	0	2

(a) Raw User/Topic Correspondence Matrix

	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	TOTAL
U_2	0.5	0	0.5	0	0	1
U_3	0.25	0.25	0.5	0	0	1
U_4	0.5	0	0.5	0	0	1
U_5	0	0.5	0	0.5	0	1

(b) Normalized User/Topic Correspondence Matrix

〈Figure 4〉 Normalizing User/Topic Correspondence Matrix

이렇게 도출된 <Figure 4(b)>의 표준화된 사용자의 이슈별 관심도는 프로세스 (7)의 사용자 관심 이슈 반영 카테고리 재구매 예측 모형의 입력으로 활용된다. 본 모형과 순수 구매이력 기반 예측 모형의 예측 정확도 비교가 프로세스 (8)에서 이루어지며, 이 과정은 다음 절에서 실험 결과와 함께 소개한다.

4. 실험

4.1 실험 개요

본 절에서는 제안 방법론을 실제 데이터에 적용한 실험 결과를 소개한다. 우선 실험을 위해 인터넷 순위 분석 전문 업체로부터 패널 5,000명의 웹 트랜잭션 기록을 제공받았다. 이 기록에는 각 패널에 대한 비식별화된 인구통계정보와 함께 각 패널들이 방문한 사이트의 URL, 방문시각, 체류시각 등의 정보가 포함되어 있다. 다음으로 패널들이 2012년 7월부터 2013년 6월까지 방문한 인터넷 뉴스의 URL 약 15,000건을 추출하고, 크롤링을 통해 이들 기사에 대한 원문을 수집하였다. 전체 패널 중 해당 기간 위의 뉴스 기사를 한 건이라도 조회한 패널은 총 2,615명으로 나타났다. 쇼핑몰 구매기록은 이들 패널들이 국내 한 인터넷 쇼핑몰 'G'사에서 상품을 구매한 기록을 사용하였으며, 2,615명의 패널 중 해당 기간에 'G'사에서 상품을 한 번이라도 구매한 사람은 총 359명으로 파악되었다. 따라서 본 실험에서는 해당 기간 중 뉴스 조회 기록과 상품 구매 기록을 모두 가진 패널 359명의 웹 사용 기록만을 분석에 사용하였다.

4.2 구매이력 기반 입력 변수 및 목적 변수 생성

본 부절에서는 순수 구매이력 기반 카테고리별 재구매 예측 모형과 본 연구에서 제안하는 사용자 관심 이슈를 반영한 카테고리별 재구매 예측 모형의 성능을 비교하기 위한 실험 과정 및 결과를 소개한다. 우선 'G사'의 분류 기준에 따라 각 상품을 14개의 카테고리로 분류하였다. 14개 카테고리의 명칭과 각 카테고리의 상품 중 본

실험에 사용된 품목의 개수가 <Table 1>에 요약되어 있다 .

〈Table 1〉 Category Information

Category	Num. Items	Category	Num. Items
Appliance	626	Food	1,758
Beauty	1,227	Furniture	1,272
Car	617	Hobby	322
Computer	1,036	Kids	1,981
Culture	597	Living	1,937
Digital	1,252	Sports	930
Fashion	3,611	Trip	51

한편 구매이력에 기반하여 카테고리별 재구매 여부를 예측하기 위해 14개 카테고리에 대한 구매 정보로부터 총 31개의 입력 변수를 생성하였다. 1년간의 구매 기록으로 구성된 전체 데이터 중 전반부 9개월 데이터로부터 입력 변수를 생성하고 후반부 3개월 데이터로부터 목적 변수를 생성하였다. 분석 기간 전체의 총 구매횟수가 향후 구매에 영향을 줄 것으로 판단하여 입력 변수가 생성된 분석 기간 전체, 즉 최근 9개월간의 총 구매횟수 및 동일 기간 14개 카테고리별 구매횟수를 입력 변수로 사용하였다. 한편 전체 기간뿐 아니라 최근의 구매횟수가 향후 구매에 더욱 강한 영향을 줄 것으로 판단하여 최근 3개월간의 총 구매횟수 및 동일 기간 14개 카테고리별 구매횟수를 입력 변수로 추가 도출하였다. 마지막으로 전체 구매횟수 중 최근 구매횟수가 차지하는 비율이 향후 구매에 미치는 영향을 분석하기 위해 최근 9개월 총 구매횟수에 대한 최근 3개월 총 구매횟수의 비율을 입력 변수로 사용하였다. 한편 목적 변수로는 향후 3개월 내 각 카테고리의 상품 구매 여부를 사용하였다. 순수 구매이력 기반 카테고리별 재구매 예측 모형 수립을 위해 사용된 입력 변수와 목적 변수에 대한 설명이 <Table 2>에 요약되어 있다.

〈Table 2〉 Variables for Category Repurchase Prediction Model

Role	Name	Definition
Input (31 Variables.)	Total_3M	Total number of purchases in the last 3 months
	Total_9M	Total number of purchases in the last 9 months
	Ratio_3M/9M	Freq_3M / Freq_9M
	Appliance_3M	Number of purchases of items from category "Home Appliances" compared to total number of purchases in the last 3 months
	...	...
	Trip_3M	Number of purchases of items from category "Travel" compared to total number of purchases in the last 3 months
	Applicance_9M	Number of purchases of items from category "Home Appliances" compared to total number of purchases in the last 9 months
	...	...
	Trip_9M	Number of purchases of items from category "Travel" compared to total number of purchases in the last 9 months
Target (14 Variables.)	T_Appliance	Yes or No to whether items in category "Home Appliances" will be purchased during the next 3 months
	...	...
	T_Trip	Yes or No to whether items in category "Travel" will be purchased during the next 3 months

4.3 사용자 관심 이슈 기반 입력 변수 생성

본 부절에서는 뉴스 기사에 대한 토픽 분석을 먼저 수행하고, 사용자가 어떤 이슈의 기사 몇 건을 조회하였는지를 분석하여 각 사용자의 관심 이슈를 계량화한 실험의 과정 및 결과를 소개한다. 토픽 분석에는 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하였으며, 전체 과정은 파싱(Parsing), 필터링(Filtering), 그리고 토픽 분석 순으로 이루어졌다. 분석의 결과로 도출될 토픽의 수에 대해서는 명확한

기준이 제시되어 있지 않으나, 통상적으로 전체 문서 수의 1/1000 이상 1/100 이하 토픽 수가 사용되어 왔다. 따라서 본 연구에서는 토픽의 수를 15개 이상 150개 이하로 변경시키며 파일럿 실험을 수행하였으며, 그 결과 토픽의 수가 30개일 때 유사 토픽이 중복 출현하는 빈도가 낮으면서 의미있는 토픽의 누락이 적게 나타남을 파악하여 본 실험의 토픽 수를 30개로 설정하였다. 전체 사용자 359명의 30개 이슈에 대한 표준화된 대응도의 일부, 즉 사용자 10명의 5개 이슈에 대한 표준화된 대응도가 <Figure 5>에 제시되어 있다. 예를 들면 USER_ID가 '9'인 사용자의 경우 <Figure 5>에 표시된 5개의 이슈 중에는 T1의 이슈에 대한 관심도가 가장 높고 반대로 T4에 대한 관심도가 가장 낮음을 알 수 있다.

	USER_ID	T1	T2	T3	T4	T5	...
1	9	0.085	0.017	0.068	0.000	0.034	
2	24	0.000	0.000	0.000	0.000	0.000	
3	26	0.009	0.002	0.018	0.004	0.029	
4	27	0.000	0.071	0.071	0.000	0.071	
5	31	0.000	0.000	0.000	0.000	0.000	
6	42	0.031	0.031	0.031	0.000	0.031	...
7	64	0.019	0.056	0.019	0.000	0.019	
8	70	0.124	0.005	0.175	0.000	0.078	
9	91	0.067	0.011	0.068	0.005	0.062	
10	101	0.000	0.000	0.000	0.000	0.000	
...							

〈Figure 5〉 Normalized User/Topic Correspondence Matrix (Part)

이렇게 도출된 표준화된 사용자별 이슈 대응도를 입력 변수로 사용함으로써, 사용자의 관심 이슈를 반영한 카테고리별 재구매 예측 모형을 수립할 수 있다.

4.4 카테고리별 재구매 예측 모형의 정확성 분석

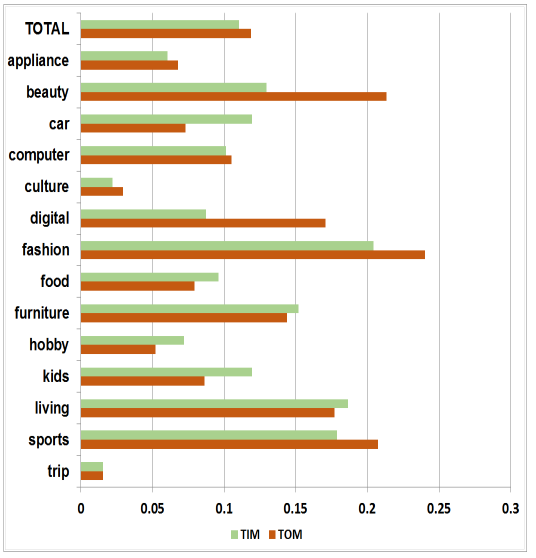
본 부절에서는 순수 구매이력 기반의 예측 모형(TOM: Transaction-Only Model), 순수 사용자 관심 이슈 기반의 예측 모형(IOM: Interests-Only Model), 그리고 구매이력 및 사용자 관심 이슈를 동시에 고려한 신규 예측 모형(TIM: Transaction and Interests Model)의 정확성 분석 결과를 제시한다. TOM의 경우 <Table 2>에 소개된 31개의 입력 변수만을 사용하여 14개의 목적 변수를 예측한다. 반면 TIM의 경우 우선 <Figure 5>에 제시된 30개 이슈에 대한 대응도를 입력 변수로 사용하여 1차 분석을 수행한 뒤 특정 임계값 이상의 변수 중요도를 갖는 이슈만을 선별하고, 이들을 <Table 2>의 31개 입력 변수와 통합하여 최종 모형의 입력 변수로 사용한다.

예측 실험은 1년 동안의 구매이력을 최초 9개월과 이후 3개월로 분할한 뒤, 9개월의 데이터로부터 규칙을 도출하고, 도출된 규칙이 이후 3개월의 카테고리별 재구매 여부를 예측하는 정확성을 분석하였다. 하지만 각 카테고리별로 이후 3개월의 재구매 여부를 확인한 결과 대부분의 카테고리에서 재구매 고객의 수가 재구매를 하지 않은 고객의 수에 비해 현저하게 적게 나타남을 알 수 있었다. 따라서 비균형 집합으로부터 편향된 규칙이 도출되는 것을 방지하기 위해 과대표본추출(Oversampling)을 사용하였다. 카테고리별 원 데이터의 재구매 비율과 과대표본추출 이후의 재구매 비율이 <Table 3>에 요약되어 있다. <Table 3>의 각 셀은 전체 고객 중 재구매 고객과 재구매를 하지 않은 고객의 비율을 나타내며, 괄호 내의 숫자는 고객 수를 의미한다 .

〈Table 3〉 Repurchase Ratio Before/After Oversampling

Category	Before Sampling		After Oversampling	
	Repurchase=Yes	Repurchase=No	Repurchase=Yes	Repurchase=No
Appliance	8% (30)	92% (329)	50% (330)	50% (329)
Beauty	13% (47)	87% (312)	48% (282)	53% (312)
Car	9% (32)	91% (327)	50% (320)	51% (327)
Computer	12% (43)	88% (316)	49% (301)	51% (316)
Culture	3% (9)	97% (350)	48% (324)	52% (350)
Digital	13% (48)	87% (311)	48% (288)	52% (311)
Fashion	19% (67)	81% (292)	48% (268)	52% (292)
Food	13% (47)	87% (312)	48% (282)	53% (312)
Furniture	10% (35)	90% (324)	49% (315)	51% (324)
Hobby	7% (24)	93% (335)	46% (288)	54% (335)
Kids	14% (49)	86% (310)	49% (294)	51% (310)
Living	22% (80)	78% (279)	46% (240)	54% (279)
Sports	12% (42)	88% (317)	48% (294)	52% (317)
Trip	1% (4)	99% (355)	45% (288)	55% (355)

이상의 과정을 거쳐 IOM, TOM, 그리고 TIM의 세 가지 모형의 예측 정확도를 오분류 비율(Misclassification Ratio) 관점에서 비교한 결과가 <Figure 6>에 요약되어 있다. <Figure 6(a)>는 세 가지 모형을 모두 인공신경망(Artificial Neural Networks)에 기반하여 구축한 경우의 오분류 비율을, 그리고 <Figure 6(b)>는 세 가지 모형을 모두 의사결정나무(Decision Tree)에 기반하여 구축한 경우의 오분류 비율을 나타낸다. <Figure 6(a)>와 <Figure 6(b)> 모두에서 IOM은 다른 두 모형에 비해 전체 카테고리에 걸쳐 오분류 비율이 높게 나타남을 알 수 있었다. 한편 성능이 우수하게 나타난 TOM과 TIM의 카테고리별 오분류 비율을 그래프로 도식화하여 비교한 결과가 <Figure 7>과 <Figure 8>에 나타나있다.



〈Figure 7〉 Comparison of Misclassification Ratio between TIM and TOM (Artificial Neural Network-based Analysis)

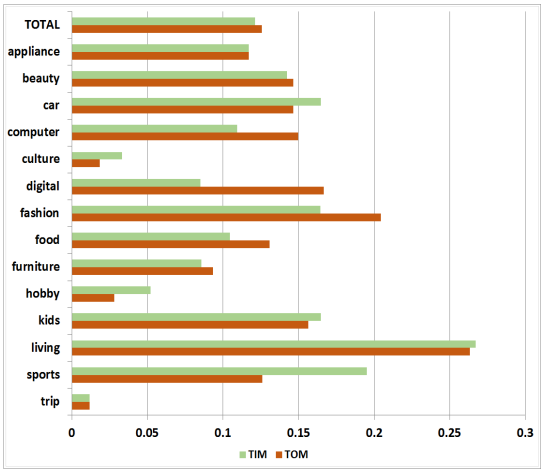
	appliance	beauty	car	computer	culture	digital	fashion	food	furniture	hobby	kids	living	sports	trip
IOM	0.1698	0.3933	0.2195	0.2389	0.0369	0.3042	0.3644	0.2552	0.2646	0.14	0.2922	0.3917	0.2073	0.0233
TOM	0.0679	0.2134	0.0731	0.1053	0.0295	0.1708	0.24	0.0794	0.144	0.052	0.0864	0.177	0.2073	0.0155
TIM	0.0604	0.1297	0.1193	0.1012	0.0221	0.0872	0.2044	0.0962	0.1518	0.072	0.1193	0.1866	0.1788	0.0155

(a) Artificial Neural Network-based Analysis

	appliance	beauty	car	computer	culture	digital	fashion	food	furniture	hobby	kids	living	sports	trip
IOM	0.166	0.2594	0.1538	0.2551	0.048	0.2375	0.3644	0.2552	0.1984	0.1	0.3128	0.3541	0.1951	0.0233
TOM	0.117	0.1464	0.1465	0.1498	0.0185	0.1667	0.2044	0.1308	0.0934	0.028	0.1564	0.2632	0.126	0.0116
TIM	0.117	0.1423	0.1646	0.1093	0.0332	0.0849	0.1644	0.1046	0.0856	0.052	0.1646	0.2671	0.1951	0.0116

(b) Decision Tree-based Analysis

〈Figure 6〉 Misclassification Ratio of Three Prediction Models



〈Figure 8〉 Comparison of Misclassification Ratio between TIM and TOM (Decision Tree-based Analysis)

<Figure 6> ~ <Figure 8>에서 인공신경망 기반 분석의 경우 전체 카테고리에 대한 오분류 비율의 평균은 TOM이 0.1187, TIM이 0.1103으로 제안 방법론인 TIM이 근소하게 우수한 것으로 나타났으며, 카테고리별로는 appliance, beauty, computer, culture, digital, fashion, 그리고 sports에서 TIM의 오분류 비율이 낮게 나타났다. 한편 의사결정나무 기반 분석의 경우 전체 카테고리에 대한 오분류 비율의 평균 역시 TOM이 0.1256, TIM이 0.1212로 TIM이 근소하게 우수한 것으로 나타났으며, 카테고리별로는 beauty, computer, digital, fashion, food, 그리고 furniture에서 TIM의 오분류 비율이 낮게 나타났다.

5. 결론

본 연구에서는 사용자의 관심을 측정하여 추천시스템의 성능을 향상시킬 수 있는 방안을 제시하는 것을 목적으로 수행되었다. 현재 대부분의 인터넷 쇼핑몰에서 얻을 수 있는 사용자 성향 정보는 사용자가 사이트 등록시 임의로 입력한 취미와 관심분야 등에 국한된다는 한계가 있다. 이러한 정보는 객관성이 보장되지 않을 뿐 아니라 충분히 세분화되어 있지 않고 갱신이 수시로 이루어지지 않기 때문에, 이러한 정보를 통해 사용자의 최근 관심을 파악하기엔 어려움이 있다. 따라서 본 연구에서는 사용자의 일상적인 인터넷 사용패턴을 분석하여 다양한 분야에 대한 고객의 성향을 계량화하고, 이를 활용하여 고객의 카테고리별 재구매 가능성을 예측함으로써 궁극적으로 추천시스템의 성능을 향상시킬 수 있는 방안을 제시하였다. 제안 방법론을 활용하여 쇼핑몰을 비롯한 인터넷 사이트들은 자사 고객의 최근 성향을 보다 객관적인 형태로 관리할 수 있으며, 이를 통해 궁극적으로 고객 만족도 향상 및 수익 증대를 달성할 수 있을 것으로 기대한다.

제안 방법론의 실무 적용 가능성을 파악하기 위해, 국내 한 인터넷 쇼핑몰의 구매기록, 인터넷 뉴스 조회기록 및 기사 내용을 사용한 실험을 수행하였다. 실험 결과 의사결정나무와 인공신경망 분석 모두에서 사용자 관심 이슈와 구매이력을 모두 활용한 재구매 예측 모형이 구매이력에만 기반한 예측 모형보다 낮은 오분류 비율을 보임을 알 수 있었다. 하지만 기존 모델과의 오분류 비율의 차이가 크게 나타나지 않는다는 점은 연구의 한계로 남는다. 또한 기존 모델과의 성능 비교가 오분류 비율의 평균 차이에 근거하여 이루어졌으나 보다 엄밀한 관찰을 위해서는 이러한 차이가 통계적으로 의미가 있는지에 대한 검증이 이루어질 필요가 있으며, 오분류 비율이 아닌 F-Score 등 다른 척도를 활용한 검증도

수행되어야 한다. 또한 인공신경망 및 의사결정나무 이외의 다른 알고리즘 기반의 다양한 예측 모델에 대한 검증을 통해 제안 모형의 공신력을 확보할 수 있을 것으로 기대한다. 한편 실험 결과 제안 방법론에 의한 성능 개선 정도가 카테고리별로 상이하게 나타나는 현상을 파악하였다. 이러한 현상은 쇼핑몰 카테고리의 분류 기준과 인터넷 뉴스 기사의 분류 기준의 일치도, 카테고리별 구매 고객의 인터넷 뉴스 접속 빈도, 실험에 사용된 카테고리별 표본 수의 차이 등에 기인한 것으로 보이나, 구체적인 원인은 본 연구에서 규명하지 못하였다. 향후 연구에서 카테고리별 성능 개선 차이의 원인을 분석하는 것은 매우 흥미로운 시도가 될 것으로 판단하며, 이러한 접근을 통해 제안 방법론의 전체 성능도 향상시킬 수 있을 것으로 기대한다.

참고문헌(References)

Aciar, S., D. Zhang, S. Simoff, and J. Debenham, "Recommender System Based on Consumer Product Reviews," *WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence table of contents*, (2006), 719~723.

Ahn, H., "Improvement of a Context-aware Recommender System through User's Emotional State Prediction," *Journal of Information Technology Applications & Management*, Vol. 21, No.4(2014), 203~223.

Ahn, S. M., I. H. Kim, B. Choi, Y. Cho, E. Kim, and M. K. Kim, "Understanding the Performance of Collaborative Filtering Recommendation through Social Network Analysis," *Journal of Society for e-Business Studies*, Vol.17, No.2 (2014), 129~147.

Armentano, M. G., D. Godoy, and A. A. Amandi, "Followee Recommendation Based on Text Analysis of Micro-blogging Activity," *Information Systems*, Vol.38, No.8(2013), 1116~1127.

Balabanovic, M. and Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," *Communication of the ACM*, Vol.40, No.3 (1997), 66~72.

Billsus, D. and M. J. Pazzani, "Learning Collaborative Information Filters," *Proceedings of 15th International Conference on Machine Learning*, (1998), 46~45.

Brynjolfsson, E., L. M. Hitt, and H. H. Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?," *Social Science Research Network*, (2011), 33~34.

Choi, H. G. and I. J. Hwang, "Emotion-based Music Recommendation System based on Twitter Document Analysis," *Journal of KIIS: Computing Pratice and Letters*, Vol.18, No.11(2012), 762~767.

Choi, S. and N. Kim, "Identifying the Interests of Web Category Visitors Using Topic Analysis," *Journal of Information Technology Applications & Management*, Vol.21, No.4(2014), 415~429.

Chun, I. G. and I. S. Hong, "The Implementation of Knowledge-based Recommender System for Electronic Commerce Using Java Expert System Library," *Proceedings of IEEE International Symposium on Industrial Electronics*, (2001), 1766~1770.

Funakoshi, K. and T. Ohguro, "A Content-Based

Collaborative Recommender System with Detailed Use of Evaluations," *Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, (2000), 253~256.

Heo, J., P. Ryu, Y. Choi, H. Kim, and C. Ock, "An Issue Event Search System based on Big Data for Decision Supporting: Social Wisdom," *Jounal of KIISE : Software and Applications*, Vol.40, No.7(2013), 381~394.

Hyun, Y., N. Kim, and Y. Cho, "A Multi-Dimensional Issue Clustering from the Perspective Consumers Interests and R&D," *Journal of Information Technology Services*, Vol.14, No.1(2015), 237~249.

Hyung, Z., K. Lee, and K. Lee, "Music Recommendation Using Text Analysis on Song Requests to Radio Stations," *Expert Systems with Applications*, Vol.41, No.5 (2014), 2608~2618.

Jeong, I.-Y., X. Yang, and H.-k. Jung, "A Study on Movies Recommendation System of Hybrid Filtering-Based," *Journal of Korea Institute of Informationand Communication Engineering*, Vol.19, No.1(2015), 113~118.

Kang, M. M., S. R. Kim, and S. M. Park, "Analysis and utilization of Big Data," *Korea Information Science Society review*, Vol30, No.6(2012), 25~32.

Kim, J., N. Kim, and Y. Cho, "User-Perspective Issue Clustering Using Multi-Layered Two-Mode Network Analysis," *Journal of Intelligent Information Systems*, Vol.20, No.2(2014), 93~107.

Kim, J. K., D. H. Ahn, and Y. H. Cho, "A Personalized Recommender System, WebCF-PT: A Collaborative Filtering using Web Mining and Product Taxonomy," *Asia Pacific Journal of Information Systems*, Vol.15, No.1 (2005), 63~79.

Kim, K. J. and B. G. Kim, "Product Recommender System for Online Shopping Malls using Data Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.11, No.1(2005), 191~205.

Kim, S. and A. H. Oh, "Offline Book Recommendation System using User Location Information," *HCI*, (2012), 53~55.

Kim, Y., N. Kim, and S. R. Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining," *Journal of Intelligence and Information Systems*, Vol.18, No.2(2012), 143~156.

Min, G. Y. and D. H. Jeong, "Research on Assessment of Impact of Big Data Attributes to Disaster Response Decision-Making Process," *Journal of Society for e-Business Studies*, Vol.18, No.3(2013). 17~43.

Roh, J., K. Yoon, J. Kim, and J.-h. Lee, "A Music Recommendation System Using Collaborative Filtering and Context Awareness," *proceeding of KIIS Fall conference*, Vol.18, No.2(2008), 76~79.

Yu, E., Y. Kim, N. Kim, and S. R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligent Information Systems*, Vol.19, No.1(2013), 95~110.

Abstract

Improving Performance of Recommendation Systems Using Topic Modeling

Seongi Choi* · Yoonjin Hyun* · Namgyu Kim**

Recently, due to the development of smart devices and social media, vast amounts of information with the various forms were accumulated. Particularly, considerable research efforts are being directed towards analyzing unstructured big data to resolve various social problems. Accordingly, focus of data-driven decision-making is being moved from structured data analysis to unstructured one. Also, in the field of recommendation system, which is the typical area of data-driven decision-making, the need of using unstructured data has been steadily increased to improve system performance. Approaches to improve the performance of recommendation systems can be found in two aspects- improving algorithms and acquiring useful data with high quality. Traditionally, most efforts to improve the performance of recommendation system were made by the former approach, while the latter approach has not attracted much attention relatively. In this sense, efforts to utilize unstructured data from variable sources are very timely and necessary. Particularly, as the interests of users are directly connected with their needs, identifying the interests of the user through unstructured big data analysis can be a crew for improving performance of recommendation systems. In this sense, this study proposes the methodology of improving recommendation system by measuring interests of the user. Specially, this study proposes the method to quantify interests of the user by analyzing user's internet usage patterns, and to predict user's repurchase based upon the discovered preferences.

There are two important modules in this study. The first module predicts repurchase probability of each category through analyzing users' purchase history. We include the first module to our research scope for comparing the accuracy of traditional purchase-based prediction model to our new model presented in the second module. This procedure extracts purchase history of users. The core part of our methodology is in the second module. This module extracts users' interests by analyzing news articles the users have

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

read. The second module constructs a correspondence matrix between topics and news articles by performing topic modeling on real world news articles. And then, the module analyzes users' news access patterns and then constructs a correspondence matrix between articles and users. After that, by merging the results of the previous processes in the second module, we can obtain a correspondence matrix between users and topics. This matrix describes users' interests in a structured manner. Finally, by using the matrix, the second module builds a model for predicting repurchase probability of each category.

In this paper, we also provide experimental results of our performance evaluation. The outline of data used our experiments is as follows. We acquired web transaction data of 5,000 panels from a company that is specialized to analyzing ranks of internet sites. At first we extracted 15,000 URLs of news articles published from July 2012 to June 2013 from the original data and we crawled main contents of the news articles. After that we selected 2,615 users who have read at least one of the extracted news articles. Among the 2,615 users, we discovered that the number of target users who purchase at least one items from our target shopping mall 'G' is 359. In the experiments, we analyzed purchase history and news access records of the 359 internet users. From the performance evaluation, we found that our prediction model using both users' interests and purchase history outperforms a prediction model using only users' purchase history from a view point of misclassification ratio. In detail, our model outperformed the traditional one in appliance, beauty, computer, culture, digital, fashion, and sports categories when artificial neural network based models were used. Similarly, our model outperformed the traditional one in beauty, computer, digital, fashion, food, and furniture categories when decision tree based models were used although the improvement is very small.

Key Words : Big Data Analysis, Data Mining, Recommendation Systems, Text Mining, Topic Modeling

Received : August 31, 2015 Revised : September 9, 2015 Accepted : Septenber 9, 2015
Type of Submission : Normal Track Corresponding Author : Namgyu Kim

저 자 소 개

최 성 이
현재 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하고 있다. 원광대학교 정보전자상거래학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝, 토픽 모델링 및 데이터베이스 등이다.

현 윤 진
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다. 국민대학교 비즈니스IT학부에서 학사 학위를 취득하고, 동 대학원에서 석사 학위를 취득하였다. 주요 관심분야는 텍스트 마이닝 및 데이터 마이닝이다.

김 남 규
현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.