한국정보전자통신기술학회논문지(jkiiect)'18-04, Vol.11 No.2
논문 18-11-02-181 http://dx.doi.org/10.17661/jkiiect.2018.11.2.181

텍스트 마이닝을 이용한 소비자 소비패턴 분석 기법 설계

정은희*, 이병관**

An Analysis Scheme Design of Customer Spending Pattern using Text Mining

Eun-Hee Jeong*, Byung-Kwan Lee**

요 약 본 논문에서는 텍스트 마이닝을 이용한 소비자의 소비패턴 분석 기법을 제안하였다. 제안하는 소비패턴 분석기법에서는 첫째, 피어슨의 상관계수를 이용하여 사용자의 평가점수에 대한 유사도를 분석하고, 둘째, 텍스트 마이닝 기법 중의 하나의 TD-IDF의 코사인 유사도를 이용하여 사용자의 리뷰들간의 유사도를 분석하고, 셋째, Sentiwordnet를 이용하여 평가점수와 리뷰의 일치성을 분석하였다. 그리고 제안하는 소비패턴 분석 기법은 평가점수의 유사도와 리뷰의 유사도를 이용하여 근접 이웃들을 선정하고, 선정된 이웃에 소비패턴에 적합한 추천리스트를 제공하였다. 추천리스트의 정확도는 피어슨 상관계수가 0.79, TD-IDF가 0.73, 그리고 제안하는 소비패턴분석기법이 0.82로 나타났다. 즉, 제안하는 소비패턴분석기법은 소비자의 정량적인 평가점수와 정성적인 리뷰를 모두 이용하므로 소비 패턴을 좀 더 정확하게 분석할 수 있었다.

Abstract In this paper, we propose an analysis scheme of customer spending pattern using text mining. In proposed consumption pattern analysis scheme, first we analyze user's rating similarity using Pearson correlation, second we analyze user's review similarity using TF-IDF cosine similarity, third we analyze the consistency of the rating and review using Sendiwordnet. And we select the nearest neighbors using rating similarity and review similarity, and provide the recommended list that is proper with consumption pattern. The precision of recommended list are 0.79 for the Pearson correlation, 0.73 for the TF-IDF, and 0.82 for the proposed consumption pattern. That is, the proposed consumption pattern analysis scheme can more accurately analyze consumption pattern because it uses both quantitative rating and qualitative reviews of consumers.

Key Words : Collaborative Filtering, Consumption pattern, Cosine similarity, Peason correlation, Text mining, TF-IDF, User review analysis

1. 서론

스마트 기기 보급이 늘어나고 활성화되면서 웹을 통한 쇼핑으로 고객들의 구매 패턴이 변화되고 있다[1]. 이런 가운데 최근 Web 2.0의 도래와 함께 텍스트를 이용해 인터넷 상에 자신의 의견을 표출하는 것이 점차 보편화되어 가고 있을 뿐만 아니라[2,3], 쇼핑몰과 같은 상거래 플랫폼에서도 상품에 대한 고객들의 의견을 공유할 수 있는 사용자 리뷰가 크게 활성화되고 있는 추세이다. 그리고 이와 같은 리뷰에는 해당 상품에 대해 고객이 갖고 있는 선호에 대한 보다 상세하고, 신뢰할 수 있는 정보를 담고 있어 추천 시스템에서 활용하기에 매우 유용하다[3,4,5,6]. 특히, 추천 시스템은 리뷰로부터 고객에 대한 정보를 추출하여 고객이 원하는 서비스를 제공함으로써 기업의 이미지를 제고시키며, 또한 이윤의 극대화 실현을 위해 이미 기업에서는 도입하고 있다.

This study has been worked with the support of a research grant of Kangwon National University in 2016.
*Department of Regional Economics, Kangwon National University
**Corresponding Author: Department of Software, Catholic Kwandong University (bklee@cku.ac.kr)
Received March 19, 2018 Revised April 02, 2018 Accepted April 05, 2018

이 추천시스템은 정량적이고 명시적인 상품에 대한 평점이나 구매 여부와 같은 정보를 수리적으로 처리하기 쉽다는 장점은 있지만, 과연 이러한 정보들이 정확하게 고객의 선호체계를 대표할 수 있는 가에 대해서는 의문이 제기되고 있다[3,4,5].

본 논문에서는 정성적인 사용자의 리뷰를 정량적으로 분석하기 위해 텍스트 마이닝 기술을 사용하는 소비자의 소비 패턴 분석기법을 설계하여 추천시스템이 갖는 의문점을 해결하고, 설계된 소비패턴 분석 기법을 이용하여 소비자에게 좀 더 정확한 소비패턴을 추천함으로써 소비자의 만족도를 향상시킬 뿐만 아니라 기업의 경쟁력을 강화시키고자 한다.

2. 관련연구

2.1 TF-IDF 텍스트 마이닝 기법

텍스트 마이닝은 자연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾고자 할 때 사용한다. 이때, TF-IDF(Term Frequency - Inverse Document Frequency)를 이용하는데, TF-IDF는 문서내에서 단어의 중요도를 빈도(Frequency)를 사용해서 계산하는 방법이다.

TF-IDF는 식 1과 같이 단어 빈도를 의미하는 TF 값과 문서 빈도의 역수를 의미하는 IDF 값을 곱하여 구한다.

$$TF-IDF_{i,j} = tf_{i,j} \times idf_i \quad \text{..... 식(1)}$$

여기서, $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$와 $idf_i = \log \frac{|D|}{|\{d_j | t_j \in d_j\}|}$를 의미하며, $n_{i,j}$는 단어 t_i가 문서 d_j에서 출현한 회수, $\sum_k n_{k,j}$는 문서 d_j에서 모든 단어가 출현한 회수, $|D|$는 문서집합에 포함되어 있는 문서의 수, 그리고 $|\{d_j | t_j \in d_j\}|$는 단어 t_j가 등장하는 문서의 수를 의미한다[6].

2.2 추천시스템

추천 시스템(recommended system)이란 특정 사용자를 위한 Top-N 추천 상품 목록을 생성하거나 추천 대상 상품들에 대한 해당 사용자의 평가 점수를 예측하는 방법을 통해, 그들이 전자상거래 사이트에서 구매를 희망하는 상품을 쉽게 찾을 수 있도록 도와주는 데이터 분석기반의 정보 여과(information filtering) 시스템을 말한다[3,7,8].

대부분의 추천 시스템은 내용기반 알고리즘보다 좀 더 우수한 추천 정확도를 보이는 협업 필터링 알고리즘을 선호하는 편이다. 협업 필터링 알고리즘은 사용자간의 유사성을 평가하여 추천 결과를 생성하는데 그 절차는 다음과 같다.

첫째, 피어슨 상관계수를 이용하여 사용자 x와 y들 간의 유사도 계산한다[9].

$$S_{x,y} = \frac{\sum_i (r_{x,i} - \overline{r_x}) \cdot (r_{y,i} - \overline{r_y})}{\sqrt{\sum_i (r_{x,i} - \overline{r_x})^2 \cdot \sum_i (r_{y,i} - \overline{r_y})^2}} \quad \text{.식(2)}$$

여기서 $S_{x,y}$는 사용자 x와 y의 유사도(similarity)를 의미하고, i는 사용자 x와 y가 평가한 상품, $r_{x,i}$는 사용자 x가 상품 i에 대해 평가한 값, $\overline{r_x}$는 사용자 x의 평가 평균값을 의미한다. 그리고 $r_{y,i}$는 사용자 y가 상품 i에 대해 평가한 값, $\overline{r_y}$는 사용자 y의 평가 평균값을 의미한다.

둘째, 사용자 x와 유사도가 높은 사용자들을 이웃으로 선택하여 집합 N을 생성한다.

셋째, 선택된 이웃들의 평가점수를 이용하여 추천 대상 사용자 x의 평가점수를 예측한다.

$$P_{x,i} = \overline{r_x} + \sum_{z \in N} (r_{z,i} - \overline{r_z}) \cdot \frac{S_{x,z}}{\sum_{z \in N} |S_{x,z}|} \quad \text{.... 식(3)}$$

여기서 $\overline{r_x}$는 추천 대상 사용자 x의 평가점수 평균값이고, $r_{z,i}$는 이웃 사용자 z가 상품 i에 대해 평가한 평가점수, $\overline{r_z}$는 이웃 사용자의 평가 평균값, 그리고

$S_{x,z}$는 추천 대상 사용자 x와 이웃 사용자 z간의 유사도를 나타낸다.

3. 소비패턴분석기법

본 논문에서 제안하는 소비패턴 분석 기법은 사용자 기반 협업적 필터링과 텍스트 마이닝을 결합시킨 기법이다. 제안하는 소비패턴 분석 기법은 사용자들이 등록한 평가점수로 유사도를 산출하고, 텍스트 마이닝 기법인 TF-IDF의 코사인 유사도를 산출하여 사용자들을 필터링 하고, 사용자의 리뷰와 평가점수를 비교하여 소비자의 소비패턴을 생성하여 추천리스트를 생성하여 이웃들에게 추천함으로써 추천리스트의 정확도를 향상시키고자 한다.

제안하는 소비패턴 분석 기법의 전체적인 구성과 흐름도는 그림 1과 같다.

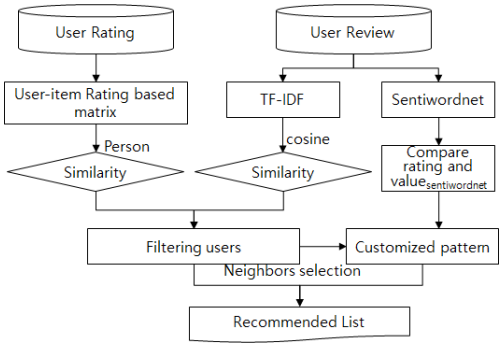

그림 1. 소비패턴 분석 기법 구성요소와 흐름도
Fig. 1. The component of customer spending pattern analysis technique and flowchart

3.1 사용자간 평가점수 유사도 분석

제안하는 소비패턴분석기법에서는 사용자들간의 구매상품에 대한 평점을 분석하고, 사용자들간의 평점 유사도를 산출한다. 이때, 평점 유사도는 피어슨 상관계수를 이용하여 계산한다.

사용자간 평가점수 유사도 분석 절차는 다음과 같다.

[단계 1] 사용자들의 구매 상품에 대한 평점으로 사용자-상품 평점 목록을 생성한다.

[단계 2] 상품에 대한 평점 등록 횟수가 기준치 이하인 사용자들은 사용자-상품 평점 목록에서 삭제한다. 본 논문에서는 평점 등록 횟수가 100건 이상인 사용자 정보들을 사용하였다.

[단계 3] 사용자-상품 평점 목록의 사용자간의 평가점수 유사도는 식(2)의 피어슨 상관계수를 이용하여 계산한다.

[단계 4] 사용자-상품 평점 목록에서 평가점수 유사도가 높은 사용자들을 선별한다.

3.2 사용자간 리뷰 유사도 분석

제안하는 소비패턴분석기법에서는 사용자의 리뷰에서 키워드를 추출하여 사용자간의 리뷰 유사도를 계산한다. 사용자들의 리뷰에서 키워드를 추출하는 절차는 다음과 같다.

[단계 1] 3.1절 사용자간 평가점수 유사도 분석에서 사용한 사용자들의 리뷰에 TF-IDF을 이용하여 리뷰에 대한 키워드를 추출한다.

[단계 2] 추출된 키워드들은 리뷰의 아이덴티티를 높게 반영한다고 할 수 있으므로, 동일한 상품에 대한 사용자들의 리뷰에서 추출된 키워드들을 이용하여 리뷰에 대한 유사도를 측정한다. 유사도는 식(4)의 코사인 유사도를 이용하여 산출한다.

$$S_{TF-IDF} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \cdots \text{식(4)}$$

[단계 3] 사용자-상품 리뷰 유사도가 높은 사용자들을 선별한다.

3.3 사용자간 유사도 분석

제안하는 소비패턴분석기법에서는 사용자간 평가점수 유사도 분석결과와 사용자간 리뷰 유사도 분석 결과를 합산하여 최종적으로 사용자간 유사도를 산출한다. 그리고 사용자간 유사도가 임계치보다 높은 N개의 사용자들을 선택하여 이웃으로 선정하고, 이웃들의 선호 아이템을 사용자가에 추천하도록 설계한다.

사용자간 유사도를 이용하여 사용자에게 선호 아이

템을 추천하는 절차는 다음과 같다.

[단계 1] 3.1절의 사용자-상품 평점 유사도와 3.2절의 사용자-상품 리뷰 유사도를 합산하는 식(5)를 이용하여 사용자간 유사도를 계산한다.

$$S_{user} = (S_{rating} + S_{TF-IDF})/2 \quad \cdots\cdots\cdots \text{식(5)}$$

[단계 2] 사용자간 유사도가 높은 N개의 사용자들을 이웃으로 선정한다.

[단계 3] 사용자의 이웃들이 소비한 상품중에서 사용자가 소비하지 않은 상품들을 추천한다.

4. 시뮬레이션 분석

4.1 데이터 수집

제안하는 소비패턴분석기법의 성능을 검증하기 위해 본 연구에서는 아마존 영화 리뷰 데이터셋을 이용하였다. 아마존 영화 리뷰 데이터셋은 SNAP(Stanford Network Analysis Project)에서 제공하는 데이터셋으로 1997년 8월부터 2012년 10월까지 등록된 영화 리뷰 데이터이다[10].

표 1. 아마존 영화 리뷰 데이터셋 통계 정보[10]
Table 1. Statistics information of Amazon movie review dataset

Dataset statistics	information
Number of reviews	7,911,684
Number of users	889,176
Number of products	253,059
Users with > 50 reviews	16,341
Median no. of words per review	101
Timespan	Aug. 1997 - Oct. 2012

아마존 영화 리뷰 데이터셋 통계 정보는 표 1에서 설명하고 있듯이 약 89만명의 사용자가 약 25만개의 영화에 대해 약 8백만건의 리뷰를 등록한 데이터셋이다.

사용자의 소비 패턴을 분석하기 위해서는 사용자의 성향을 분석할 필요가 있다. 그리하여 영화 리뷰 전체를 이용하지 않고, 100건 이상 리뷰를 등록한 사용자들을 추출하여 사용자의 성향을 분석하였다.

4.2 소비패턴분석기법 분석

제안하는 소비패턴분석기법은 Python을 이용하여 구현하였고, 소비패턴분석기법에 의해 추천된 상품이 사용자의 성향에 맞는 상품이었는지를 평가하여 제안하는 소비패턴분석기법의 정확도를 분석하였다.

4.2.1 사용자간 평가점수 유사도 분석 결과

3.1절의 절차에 따라 사용자간 평가점수 유사도를 분석하기 위해 동일한 제품들을 구매한 사용자들을 추출하여 표 2와 같은 사용자간 평가점수표를 생성하였다. 사용자간의 평가점수 유사도는 피어슨의 상관계수를 이용하였고, 사용자간의 좀 더 정확한 유사도를 측정하기 위해 동일한 품목에 대한 평가값을 추출하여 사용자간 평가점수 유사도를 계산하였다.

표 2. 사용자간 평가점수(예시)
Table 2. rating among users(sample)

User ID / productID	User A	User B	User C	User D	User E
B000J10FLY	0	0	5	5	0
B001TGV882	0	2	4	5	0
B005ZMUP8K	4	4	5	3	3
B000MMMTAK	1	2	5	2	0
B002YJMMBA	3	3	0	3	5
B000JLTR90	3	3	4	0	3
B001OKUREO	0	4	0	4	3
B000YAF4MA	4	4	0	5	4
B002V0GZ9M	0	4	5	0	4
B003QTUQGU	0	4	0	5	5

표 3의 사용자간 평가점수 유사도 분석 결과에서 알 수 있듯이 사용자 A와 B의 유사도가 가장 큰 것으로 나타났으며, 사용자 C와 D, 사용자 A와 D 순으로 나타났다.

표 3. 사용자간 평가점수 유사도 분석(예: 5명)
Table 3. Similarity analysis of rating among users(sample: 5 persons)

User ID / Similarity	User A	User B	User C	User D	User E
User A	1.00	0.92	0.07	0.61	0.15
User B	0.92	1.00	0.49	0.32	0.45
User C	0.07	0.49	1.00	0.65	0.52
User D	0.61	0.32	0.65	1.00	0.21
User E	0.15	0.44	0.52	0.21	1.00

4.2.2 사용자간 리뷰 유사도 분석 결과

3.2절의 절차에 따라 TF-IDF의 코사인 유사성을 분석하였다. 사용자간의 리뷰 유사도 측정을 위해, 동일한 제품에 대한 사용자의 리뷰를 TF-IDF로 분석하여 키워드를 추출하고, 키워드들에 대한 코사인 유사성을 측정하였다.

그림 2는 사용자간 리뷰 유사도를 분석한 코드이다. 키워드를 추출할 때, min_df를 리뷰의 길이에 따라 값을 0.1에서 0.4 사이의 값을 설정하였다.

```
import pymysql
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
np.set_printoptions(precision=2)
from sklearn.feature_extraction.text import TfidfTransformer
import nltk.stem
conn = pymysql.connect(host='localhost', user='root', passwd='', db='movie', charset='utf8', use_unicode=True)
cur = conn.cursor(pymysql.cursors.DictCursor)
cur.execute("select * from review_similarity")
results = cur.fetchall()
f1=open("./reviewAB.txt",'w')
for row in results:
    a = row['ReviewA']
    b = row['ReviewB']
    if (a != "." and b != ".") :
        f1.write(a + '\n')
        f1.write(b + '\n')
f1.close()
cur.close()
conn.close()
```

```
f2=open("./reviewAB.txt",'r', encoding='utf-8')
cnt=0
review1=[]
counts1=[]
while True:
    line = f1.readline()
    if not line: break
    review1.append(line)
    cnt = cnt + 1
f2.close()
vect1 = CountVectorizer(min_df=0.4, stop_words="english").fit(review1)
counts1 = vect1.fit_transform(review1)
counts1 = counts1.todense()
i=0
total_cosine=0.0
while True:
    cosine = cosine_similarity(counts1[i], counts1[i+10])
    i = i + 1
    total_cosine = total_cosine + cosine
    if i==10: break
avg_cosine = total_cosine / cnt
print(total_cosine)
print(avg_cosine)
```

그림 2. 사용자간 리뷰 유사도 분석 코드(사용자 A와 B)
Fig. 2. Similarity analysis code of review among users (user A and B)

표 4는 TF-IDF로 분석한 사용자들 리뷰의 키워드이다.

표 4. 사용자들의 리뷰에 대한 키워드 분석 결과(예 5명)
Table 4. The result of keyword analysis about user's review (sample : 5 persons)

User ID	Keyowrds
User A	acting, beliefs, critique, dark, enjoy, enjoyable, film, films, good, hold, howard, life, like, little, long, movie, number, personal, political, quite, release, say, slow, stars, story, storyline, viewing, watch, …
User B	actors, aspects, audience, better, book, cast, character, director, effects, end, family, father, film, fine, good, hands, important, interesting, life, like, long, love, manner, men, mixed, movie, musical, old, people, real, robert, score, script, small, special, story, time, version, viewer, work, world, written, young, …

User C	action, bizarre, blood, book, brought, certainly, characters, directing, doctor, doesn, end, especially, fans, film, foe, good, great, lake, life, little, lives, love, man, men, missed, movie, new, plays, plot, possible, pregnant, reality, scene, scenes, sea, story, struggling, suddenly, time, town, viewers, water, way, writing, years, young, …
User D	actually, bit, characters, different, doesn, don, film, funny, good, know, like, mean, movie, people, quite, really, reason, remember, right, say, scenes, story, storyline, strange, sure, thing, time, watching, way, …
User E	american, apparently, bit, book, clearly, couple, don, especially, extremely, fact, film, good, got, great, help, job, just, know, life, like, little, long, man, maybe, message, movie, old, people, play, police, pretty, probably, room, say, seen, sense, stars, story, tell, thing, time, water, way, young, …

표 5는 각 제품에 대한 사용자들의 리뷰에 대한 유사도를 분석한 결과 중에서 상품 8개에 대한 사용자 A와 B, 사용자 B와 D의 리뷰 유사도 결과를 설명한 것이다.

표 5. 제품에 대한 사용자 리뷰 유사도 분석 결과(예 8개)
Table 5. Similarity analysis of user's review about product (sample: 8 products)

Product ID	Cosine similarity (User A,B)	Product ID	Cosine similarity (User B,D)
B000KKQNRO	0.54	B001TGV882	0.19
B005ZMUP8K	0.83	B005ZMUP8K	0.28
B001QH32CE	0.60	B005ZMUP8K	0.49
B000MMMTAK	0.31	B002YJMMBA	0.17
B0083SI986	0.92	B001OKUGREO	0.82
B000JLTR90	0.87	B0068FZ05Q	0.72
B000ZLFALI	0.55	B00YAF4MA	0.90
B000YAF4MA	0.94	B003QTLIQGU	0.39

사용자 A와 B의 경우, 제품 B000YAF4MA에 대한 코사인 유사도가 0,94로 가장 높으며, 제품 B000MMMTAK에 대한 코사인 유사도가 0.31로 가장 작다.

사용자 B와 D의 경우, 제품 B000YAF4MA에 대한 코사인 유사도가 0.90로 가장 높으며, 제품 B002YJMMBA에 대한 코사인 유사도가 0.17로 가장 작다.

표 6은 사용자간의 리뷰 유사도를 코사인 유사도를 이용하여 분석한 결과값이다. 표 6에서 설명하고 있듯이, 동일한 제품에 대한 리뷰들에 대한 키워드 유사성은 사용자 A와 B가 0.65로 가장 높으며, 사용자 B와 D가 0.57 , 그리고 사용자 B와 E가 0.54 순으로 나타났다.

표 6. 사용자간 리뷰 유사도 분석 결과(예 5명)
Table 6. Similarity analysis of review among users(sample: 5 persons)

User ID / Similarity	User A	User B	User C	User D	User E
User A	1.00	0.65	0.50	0.51	0.28
User B	0.65	1.00	0.59	0.54	0.57
User C	0.50	0.59	1.00	0.4	0.52
User D	0.51	0.54	0.4	1.00	0.40
User E	0.28	0.57	0.52	0.40	1.00

표 3과 표 6의 결과를 비교해볼 때, 사용자간 평가점수 유사도는 사용자 A와 C가 가장 낮은 것으로 나타났었는데, 사용자간 리뷰 유사도는 사용자 A와 E가 가장 낮은 것으로 나타났으며, 전체적으로 평가점수 유사도가 낮으면, 리뷰 유사도도 낮게 나타남을 알 수 있다. 그리고 사용자간 리뷰 유사도는 하나의 rating 점수가 아닌 여러개의 키워드들을 이용하여 유사도를 평가하기 때문에 사용자간 평가점수 유사도와 다르게 최고값과 최저값의 차이가 크지 않음을 알 수 있다.

4.2.3 사용자간 유사도 평가 결과

사용자간의 유사도는 사용자의 평가점수와 리뷰를 이용하여 산출한 각각의 유사도를 식(5)에 대입하여 계산하였고, 그 결과는 표 7과 같다.

표 7에서 볼 수 있듯이 사용자 A의 이웃으로 B, D를 선정하였고, 사용자 B의 이웃으로 A, C, E를 선정하였고, 사용자 C의 이웃으로 B, D, E를 선정하였고, 사용자 D의 이웃으로 A, C를 선정하였고, 사용자 E의 이웃으로 B, C를 선정하였다. 그리고 이 이웃들이 구매한 제품들

을 사용자 A, B, C, D, E에게 추천하도록 하였다.

표 7. 사용자간 유사도 분석 결과(예 5명)
Table 7. Similarity analysis of among users(sample: 5 persons)

User ID / Similarity	User A	User B	User C	User D	User E
User A	1.00	0.79	0.29	0.56	0.22
User B	0.79	1.00	0.55	0.43	0.51
User C	0.28	0.54	1.00	0.53	0.52
User D	0.56	0.43	0.53	1.00	0.31
User E	0.21	0.51	0.52	0.31	1.00

사용자들의 리뷰에 대한 긍정과 부정은 Sentiwordnet를 이용하여 분석하였다. 표 8은 사용자 중에 User B의 결과이다. 표 8에서 설명하고 있듯이 User B의 ranking 값과 리뷰에 대한 긍정비율값과 부정비율값이 일치하지 않는 결과들이 있다. 즉, User B의 경우 productID가 B001TGV882와 BoooMMMTAK 처럼 긍정적인 내용으로 리뷰를 등록하였지만, 평가점수를 낮게 등록하거나 또는 productID가 B0083Sl986의 경우 부정적인 내용으로 리뷰를 등록하였지만, 평가점수를 높게 등록한 결과를 볼 수 있다.

표 8. 리뷰의 긍정/부정 분석결과와 rating(예 user B)
Table 8. Review's positive/negative analysis and rating (sample: User B)

ProductID	Ranking	Positive Rate	Negative Rate
B003AI2VGA	3	10.46	11.96
B000KKQNRO	4	7.72	4.62
B000NOlVT0	3	6.95	6.9
B001TGV882	2	9.59	6.02
B005ZMUP8K	4	12.27	9.42
B001QH32CE	4	10.12	11.59
B000MMMTAK	2	13.01	8.12
B0083Sl986	5	9.72	11.51
B002YJMMBA	3	17.54	11.65
B000JLTR90	3	4.61	3.00
B001OKUREO	4	8.74	10.82
B000ZLFALI	5	18.72	14.93
B0068FZ05Q	4	8.24	2.95
B000YAF4MA	4	16.68	10.54
B002V0GZ9M	4	13.67	9.58
B003QTUQGU	4	14.02	11.33

표 9는 사용자들의 평가점수와 리뷰의 긍정값 및 부정값의 일치성을 평가한 결과이다. 가장 높은 일치성을 보인 사용자는 User A로 0.9이고, 나머지 사용자들은 거의 비슷한 일치성으로 나타났다. 즉 대부분의 사용자들이 평가점수와 다르게 리뷰를 등록하므로 평가점수만으로 제품에 대해 평가하는 것은 사용자의 의도를 정확하게 평가하기 어렵다는 것을 알 수 있다.

표 9. 리뷰의 긍정/부정과 일치성 분석(예 5명)
Table 9. Consistency analysis result of review's positive/negative and rating (sample: 5 persons)

	User A	User B	User C	User D	User E
Consistency rate	0.9	0.69	0.7	0.64	0.64

표 10은 추천리스트의 정확도를 비교한 결과이다. 사용자들간의 유사도로 선정된 이웃들의 구매품목을 참조하여 생성한 추천리스트를 제공한 결과, 피어슨의 상관계수의 정확도는 0.79이고, TF- IDF의 정확도는 0.73, 그리고 제안하는 소비패턴분석기법의 정확도는 0.82로 나타났다.

표 10. 추천리스트의 정확도 결과
Table 10. The precision result of recommended list

	Person Correlation	TF - IDF	Proposed Scheme
Precision Rate	0.79	0.73	0.82

즉, 제안하는 소비패턴분석기법은 평가점수 분석결과와 리뷰 분석 결과를 모두 반영하여 추천리스트를 작성하였기 때문에 정확도가 가장 높은 것으로 나타났다.

5. 결론

본 논문에서는 정량적인 평가점수와 정성적인 사용자의 리뷰를 분석하는 소비자의 소비패턴 분석 기법을 제안하였다. 제안하는 소비패턴 분석기법에서는 피어슨의 상관계수를 이용하여 정량적인 평가점수를 분석하였고, 텍스트 마이닝 기법 중의 하나의 TD-IDF의 코사인 유사도를 이용하여 사용자의 리뷰를 정량적으로 분석하 였다. 또한, 리뷰의 긍정 및 부정의 일치성

분석에는 Sentiwordnet를 이용하였다.

그 결과, 제안하는 소비패턴 분석 기법으로 이웃을 선정하고, 선정된 이웃에 추천한 추천리스트의 정확도가 피어슨의 상관계수와 TD-IDF의 코사인 유사도를 이용하여 생성한 추천리스트 보다 높았다.

즉, 제안하는 소비패턴 분석 기법을 이용하여 소비자에게 좀 더 정확한 소비패턴을 추천함으로써 소비자의 만족도를 향상시킬 뿐만 아니라 기업의 경쟁력을 강화시킬 수 있을 것이다.

REFERENCES

[1] Shin, C. H., J.W. Lee, H.N. Yang, and I.Y. Choi, "The research on Recommender for New Customers Using Collaborative Filtering and Social Network Analysis," Journal of Intelligence and Information Systems, vol.18, no.4, pp.19-42, 2012.

[2] Chen, P.Y., S. Dhanasobhon, and M.D. Smith, "An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.com," Proceedings of International Conference on Information Systems(ICIS), 94, 2007.

[3] B. K. Jeon, H. C. Ahn, "A Collaborative Filtering System Combined with User's Review Mining : Application to the Recommendation of Smartphone Apps," Journal of Intelligence and Information Systems, vol.21, no.2, pp.1-18, 2015.

[4] B. K. Jeon, "A Study on the Combination of Collaborative Filtering and User's Review Mining," Kookmin University Graduate School of Business IT, Master thesis, 2016.02.

[5] Zhang, Z., D. Zhang, and J. Lai, "urCF: User Review Enhanced Collaborative Filtering," Proceedings of the 20th Americas Conference on Information Systems, 2014.

[6] S. J. Lee, H. J. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF," The Journal of Society for e-Business Studies, vol.14, no.4, pp.59-73, 2009.

[7] Choeh, J.Y., S.L. Lee and Y.B. Cho, "Applying Rating Score's Reliability of Customers to Enhance Prediction Accuracy in Recommender System," Journal of Digital Contents Society, vol.13, no.7, pp.379-385, 2013.

[8] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.,"Item-based collaborative filtering recommendation algorithms," Proceedings of the 10th International Conference on World Wide Web, pp.285~295, 2001.

[9] Jeong, E.H, and Lee, B.K., "A Design of Customized Market Analysis Scheme Using SVM and Collaboration Filtering Scheme," The Journal of Korea Institute of Information, Electronics, and Communication Technology, vol.9, no.6, pp.609-616, 2016.

[10] Amazon movies review dataset, https://snap.stanford.edu/data/web-Movies.html

저자약력

정 은 희(Eun-Hee Jeong) [종신회원]

- 1998년 2월 : 관동대학교 일반대학원 전자계산공학과 (공학석사)
- 2003년 2월 : 관동대학교 일반대학원 전자계산공학과 (공학박사)
- 2003년 9월 ~ 현재 : 강원대학교 글로벌인재학부 지역경제전공 교수

<관심분야> 전자상거래 보안, 빅 데이터, 헬스케어, IoT 보안, 기계학습

이 병 관(Byung-Kwan Lee) [종신회원]

- 1986년 2월 : 중앙대학교 전자계산공학과 (공학석사)
- 1990년 2월 : 중앙대학교 전자계산공학과 (공학박사)
- 1988년 3월 ~ 현재 : 가톨릭관동대학교 소프트웨어학과 교수

<관심분야> 네트워크 보안, 빅 데이터, 데이터 마이닝, IoT 보안, 인공지능