Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Improving Predictive Accuracy of User-based Collaborative Filtering Using Word2Vec

Boo-Sik Kang*

Division of Service Managements, MokWon University

ABSTRACT

Word2Vec is a most popular method in text mining area, recently. It converts words to vectors using association among words in sentences. Similar words are nearly located in the vector space. Improving predictive accuracy of recommender algorithms is a major work in the area of recommender systems. User-based collaborative filtering recommends products using the information about product preference of Neighbors. This study proposed a method to compute user similarity using vectors of users by Word2Vec instead of using traditional method. In order to use Word2Vec, we separate sentences first, and then find corpus that is meaningful word set of the sentences. For using Word2Vec in user-based movie recommender, we find users that have seen same movies first, we substitute an user to a word and user list of a movie to corpus of one sentence. There can be several methods to compose the sentences in recommender systems. This study considers two methods, first method constructs a sentence per movie and second method can construct several sentences per movie. After sentence construction, it enters corpus of sentences into Word2Vec and computes vectors of users, and then computes user similarity by coefficient corelation method using the vectors of users. Using the similarity, it recommends products by user-based collaborative filtering. To validate, the proposed methods were applied to filmtrust dataset. The experimental results of repeating 10-fold cross validation three times showed that mean MAE of user-based collaborative filtering(wvCF3.0) applying Word2Vec improved the predictive accuracy greatly than that of conventional collaborative filtering method(uCF). Also, it showed that the sentence expansion method(wvCFthree) constructing several sentences per movie is better than the one sentence method(wvCF3.0) constructing one sentence per movie for improving the predictive accuracy. To test statistical significance between uCF and wvCF3.0, and between wvCF3.0 and wvCFthree, we experimented paired t-test and confirmed the statistical significance.

KEYWORDS: Word2Vec, Predictive accuracy, Collaborative filtering, Recommender systems, User similarity

ARTICLE INFO: Received 30 January 2018, Revised 5 February 2018, Accepted 8 February 2018.

*Corresponding author is with the Division of Service Management, MokWon University, 88 Doanbuk-ro Seo-gu Daejeon, 35349, KOREA.

E-mail address: bookang@mokwon.ac.kr

1. 서 론

많은 온라인 쇼핑몰에서 상품추천을 하고 있는데, 상품추천에 있어 중요한 요소 중의 하나는 상품추천의 정확도이다. 상품추천시스템에서 가장 많이 활용되는 기법은 협업필터링 추천방식으로, 협업필터링은 크게 아이템기반 협업필터링과 사용자기반 협업필터링으로 구분할 수 있다[1,2]. 아이템기반 추천방식은 사용자가 구매했거나 좋아하는 아이템과 유사한 아이템을 추천하는 방식이다[1,3]. 사용자기반 추천방식은 사용자간의 유사성을 측정하여 유사한 특성을 갖는 이웃 사용자들의 구매 및 선호정보를 이용하여 추천하는 방식이다[2,3]. 사용자기반 협업필터링 추천시스템에서 추천 상품의 예측 정확도를 높이기 위한 주요 과정 중의 하나는 좀 더 정확하게 사용자 유사성을 찾아 반영하는 것이다. 지금까지는 주로 사용자의 상품 구매 여부, 구매 상품에 대한 사용자의 평점 등을 이용하여 유사성을 측정하였다[3].

최근 자연어 처리 분야중 하나인 텍스트 분석 분야에서 문서분류[4], 감성사전 구축[5], 한국어 신문기사 분류[6], 문서기반 검색방법[7], 감성분석에 의미론적 요소 반영[8]등 Word2Vec이 활발하게 사용되고 있다. Word2Vec은 단어를 벡터로 변환하는 기법으로, 문장 내부의 단어 간 연관성을 파악해 벡터로 변환한다. 유사한 단어는 벡터 공간에서 가까운 거리에 위치하게 된다[9]. Word2Vec은 텍스트 분석외의 영역에서도 응용할 수 있는 데, 상품추천 시스템에서 아이템을 단어로 간주하여 추천하는 아이템기반 추천시스템[10,11]도 제안되고 있다.

사용자 기반 협업필터링 방식에서 사용자 유사성 척도로 사용자 구매 상품과 상품 평점의 상관계수를 일반적으로 많이 사용한다[3]. Word2Vec을 이용하면 사용자를 다차원의 벡터로 표현하고 이를 이용하여 사용자 유사성을 측정할 수 있다. Word2Vec은 문장내의 단어들의 연관성을 벡터로 표현하기 때문에, Word2Vec을 사용하기 위해서는 먼저 사용자와 상품간의 정보를 이용하여 문장을 구성해야 한다. 문장 구성 형태에 따라 사용자의 벡터 값이 달라진다.

본 연구에서는 다양한 문장 구성을 통해 Word2Vec에서 추출한 사용자 벡터를 사용하여 사용자 유사도를 구하고, 이를 활용해 사용자기반 협업필터링의 추천 상품의 예측 정확성을 높이는 방안을 제시한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 사용자기반 협업필터링 방식과 Word2Vec에 대해서 살펴본다. 제 3장에서는 본 연구에서 제안한 방안에 대해 설명한다. 제 4장에서는 제안한 방법을 filmtrust 데이터[12]에 적용하여 실험을 하고 분석한다. 제 5장에서는 결론을 기술한다.

2. 협업필터링과 사회연결망 분석

2.1 사용자기반 협업필터링

사용자기반 협업필터링 방식은 추천 대상자와 성향이 유사한 이웃사용자들을 찾아내고, 이웃사용자들의 구매정보를 활용하여 추천 대상자가 구매하지 않은 상품들 중에서 가장 구매가능성이 높게 예측되는 상품들을 추천한다[3,13]. 일반적인 추천 단계는 다음과 같다[3].

1단계에서는 사용자간 유사도를 계산한다. 유사도를 측정하는 방법으로는 피어슨 상관계수를 가장 많이 사용되며 식(1)과 같이 계산된다.

$$s(x,y)=\frac{\sum_{j=1}^{n}\left(v_{xj}-\overline{v_x}\right)\left(v_{yj}-\overline{v_y}\right)}{\sqrt{\sum_{j=1}^{n}\left(v_{xj}-\overline{v_x}\right)^2}\sqrt{\sum_{j=1}^{n}\left(v_{yj}-\overline{v_y}\right)^2}} \quad (1)$$

식(1)에서 $s(x,y)$는 사용자 x와 y의 유사도를 나타내고, v_{xj}와 v_{yj}는 시용자 x와 y가 상품 j에 부여한 선호도이며, $\overline{v_x}$와 $\overline{v_y}$는 사용자 x와 y의 평균선호도를 의미한다.

2단계에서는 추천 대상자와 유사도가 높은 N명의 이웃사용자를 선정한다.

3단계에서는 추천 대상자를 위한 상품별 선호도 평가점수를 식(2)와 같이 계산한다.

$$p(x,i)=\overline{v_x}+\frac{\sum_{k\in N} s(x,k)\times(v_{ki}-\overline{v_k})}{\sum_{k\in N}|s(x,k)|} \quad (2)$$

식(2)에서 $p(x,i)$는 추천 대상자 x를 위한 추천 대상 상품 i의 예측 평가점수이다. N은 2단계에서 선정한 이웃사용자 집합을 나타낸다.

4단계에서는 추천 대상자 a에게 선호도 예측 평가점수가 높은 상품 순으로 추천 한다.

2.2 Word2Vec

Word2Vec은 2013년 구글에서 발표된 연구로, 자연언어, 음성, 이미지와 같이 사람과 관련된 영역에서 많이 사용되고 있으며, 모델 내부에서 심층 신경망을 이용해 문장내의 단어를 다차원의 벡터로 변환하여 주는 모델이다[9]. 상품추천시스템에서 아이템을 단어로 간주하여 추천하는 아이템기반 추천시스템[10,11]이 제안되고 있는 데, 이 연구에서는 사용자를 단어로 간주하여 추천시스템에 활용하는 사용자기반 추천에 대해 살펴본다.

Word2Vec 기능은 파이썬에서 Gensim[14] 라이브러리를 주로 사용한다. Word2Vec을 사용하는 과정은 개략적으로 다음과 같다. 먼저, 문서를 읽어들이고, 문서를 여러 문장으로 구분한다. 한 문장을 의미있는 단어들의 조합으로 구성된 말뭉치로 정리한다. 각 문장은 다양한 단어들로 구성되어 있어 문서를 문장-단어 매트릭스로 표현하면 매우 큰 사이즈의 희박행렬로 표현된다. 정리된 전체 말뭉치를 Word2Vec에 입력으로 넣어 출력으로 각 단어의 다차원 벡터를 구한다. 다차원의 크기 M은 사용자가 결정해야 하는 Word2Vec의 하이퍼 파라미터 중 하나이다. 즉 출력으로 얻고 싶은 차원의 크기를 M으로 설정한다. 또한 M차원의 벡터는 문장 내 단어 간 연관성을 파악하여 구하게 되는 데, 단어 간 연관성을 파악하는 과정에서 몇 단계 인접단어까지 고려하여 구할 것인지를 사용자가 결정하여 주어야 한다. 이를 윈도우 크기라고 하며, 만약 윈도우 크기가 1이면 단어의 앞뒤에 나오는 단어만을 고려하여 연관성을 구하고, 윈도우 크기가 10이라면 앞 10단어, 뒤 10단어까지를 고려하여 연관성을 구한다.

3. 제안 추천 알고리즘

본 연구는 사용자기반 협업필터링의 예측 정확도를 높이기 위해서 Word2Vec을 활용하는 방안에 대한 연구이다. 사용자의 유사도를 측정하기 위해 기존의 사용자 평점을 이용하지 않고 Word2Vec으로 구한 사용자 벡터를 활용하는 방안에 대해 살펴본다. 이를 위한 절차는 다음 <그림 1>과 같다.

말뭉치 구성단계에서는 사용자 구매 상품 및 평점 정보를 이용하여 먼저 문장을 구성한다. 문장을 구성하는 단어로 사용자ID를 사용한다. 예를 들어, A라는 상품을 구매한 사용자ID 리스트를 찾으면, 사용자ID 리스트는 문장에 대응되고 단어의 조합은 문장의 말뭉치에 해당한다. 모든 상품에 대한 구매 사용자ID 리스트를 찾으면 전체 말뭉치를 구성하게 된다. 사용자 평점 정보가 있는 경우에는 문장의 구성이 다양하게 만들어 질 수 있다. A라는 상품에 해당하는 문장을 구성할 때, 하나의 문장으로 구성

할 수도 있고 여러 개의 문장으로 구성될 수도 있다. 가장 최적의 문장 구성 형태는 실험을 통해 찾는다.

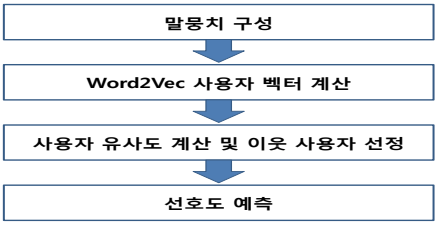

그림 1. 추천 알고리즘 절차

Figure 1. Procedure of the recommendation algorithm

Word2Vec 사용자 벡터 계산단계에서는 전 단계에서 구성한 전체 말뭉치를 입력으로 넣어 사용자 벡터를 계산한다. 이 때 하이퍼 파라미터로 사용자 벡터 크기와 윈도우 크기를 설정한다. 최적의 파라미터 값은 실험을 통해 찾는다.

사용자 유사도는 전 단계에서 구한 사용자 벡터를 가지고, 식(1)을 이용하여 사용자 x와 y간의 유사도 $s(x,y)$를 계산한다. 이웃사용자 선정은 추천 대상자와 유사도가 높은 순서로 N명을 선정한다.

선호도 예측단계에서는 식(2)를 이용하여 추천 대상자 x를 위한 추천 대상 상품 i의 평가점수를 예측한다.

추천 대상자를 위한 상품별 선호도 예측작업이 끝나면 선호도가 높은 상품 순으로 추천 대상자에게 상품을 추천한다.

4. 실험분석

4.1 실험데이터

본 연구에서 제안된 방안을 평가하기 위해 LibRec[15]에서 제공하는 filmtrust 데이터를 사용하였다. filmtrust 데이터는 1508명의 사용자가 2071개 영화에 대해 0.5점에서 4점사이의 영화 평점을 부여한 데이터로 <userid, movieid, movieRating> 3개의 속성과 35,497개의 사례로 구성되어 있고 데이터 밀도는 1.14%이다[3,12]

4.2 실험 및 결과분석

본 연구에서 제안한 방안의 예측 정확도 평가를 위한 평가척도로 협업필터링 연구에서 일반적으로 많이 이용하는 MAE(Mean Absolute Error)를 사용하였다[1,3]. 전체 데이터는 학습데이터와 검증데이터로 구분되고, 학습데이터를 이용하여 상품추천 모델을 구축하고 검증데이터내의 영화 평점을 예측한다. 검증데이터의 실제 영화 평점과 예측 평점 차의 절대값 평균이 MAE이다. 평가를 위한 실험은 사용자 벡터를 얻기 위해 파이썬과 Gensim 내의 Word2Vec을 사용하였고, 기존 연구와의 예측 정확성 비교는 R3.3.2버전의 RStudio 환경하에서 실시하였다.

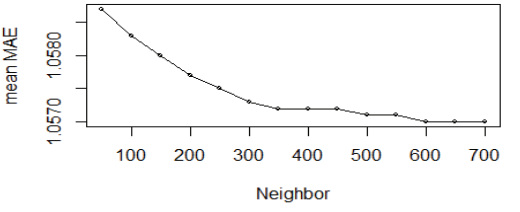

그림 2. 이웃사용자와 평균 MAE

Figure 2. Neighbor and mean MAE

실험은 전통적인 사용자기반 협업필터링(uCF)에 의한 예측 정확도와 Word2Vec의 사용자 벡터기반 협업필터링(wvCF)의 예측 정확도를 비교하였다. 먼저 CF 방식에서 최적 이웃사용자 수 N을 결정하기 위해 선행연구[3]와 같이 사전 실험에서 이웃사용자 수의 변화에 따른 MAE를 살펴보았다. 10-겹

상호검증을 실시하여 실험한 결과 이웃사용자 수 N에 따른 평균 MAE는 <그림 2>와 같았다. 이웃사용자의 수가 600부터 평균 MAE가 최저를 나타냈고, 향후 실험에서는 이웃사용자 수 N을 600으로 설정하였다.

Word2Vec의 하이퍼 파라미터인 벡터 크기 M과 윈도우 크기 W에 대한 사전 실험에서는 M과 W의 크기 변화에 따른 예측정확도의 영향은 크지 않았다. 그 중에서도 좀 더 나은 결과를 보인 M=200, W=10으로 설정하여 실험하였다.

표 1. 영화-사용자 평점 예
Table 1. An example of movie-user ratings

movie	user							
	1	2	3	4	5	6	7	8
A	4	3	3.5		4	3	1	
B		3.5	4			4		3.5

문장을 구성하는 방안에 대해 몇 가지를 사전 실험하였다. 예를 들어 A 영화에 대해 평점 3점 이상을 부여한 사용자ID를 하나의 문장으로 만드는 방안(1)과 A 영화에 대해 같은 평점을 부여한 사용자ID만으로 문장으로 만드는 방안(2)이 있다. (1)의 경우에는 하나의 문장 안에 3점 이상을 부여한 모든 사용자ID가 단어 조합을 구성하게 된다. <표 1>에서 영화 A에 대해 {1, 2, 3, 5, 6}, 영화 B에 대해 {2, 3, 6, 8}의 2개 문장이 구성된다. (2)의 경우에는 4점을 부여한 사용자ID만으로 문장을 구성하면 {1, 5}, {3, 6}의 2개 문장이 만들어 지고, 3.5점을 부여한 사용자ID만으로 문장을 구성하면 {2, 8}의 1개 문장이 구성되고, 3점을 부여한 사용자ID만으로 문장을 구성하면 {2, 6}의 1개 문장이 만들어 진다. 3점 이상 문장 확장 방법은 영화 A에서 {1, 5}, {2, 6}, 영화 B에 대해 {3, 6}, {2, 8}의 4개 문장을 만들게 된다. 문장 구성 목적이 사용자간 유사성을 측정하는 것이기 때문에 한 문장에 사용자가 1명인 경우에는 문장을 구성하지 않았다. 사전 실험 결과에서는 (2)방안의 예측 정확도가 높았다. 이후 실험은 (2)의 방법을 중심으로 실험하였다.

전통적인 사용자기반 협업필터링(uCF)과 Word2Vec의 사용자벡터기반 협업필터링(wvCF)의 예측 정확도를 비교하기 위해 R언어의 cvTools를 이용하여 10-겹 상호검증을 3회 실시하였다. 즉 각각의 제안된 방안에 대해 전체 데이터의 10%는 검증데이터, 90%는 학습데이터로 구성하여 검증하는 상호검증을 총 30회 실시한 결과는 <표 2>와 같다.

표 2. 상호검증시험의 평균 MAE
Table 2. Mean MAE of cross validation test

method	mean MAE
uCF	1.0125
wvCF4.0	0.8372
wvCF3.5	0.8372
wvCF3.0	0.8369*
wvCF2.5	0.8382
wvCF2.0	0.8385
wvCF1.5	0.8392
wvCF1.0	0.8399
wvCF0.5	0.8603

<표 2>에서 uCF는 전통적인 사용자기반 협업필터링 방식을 의미한다. wvCF4.0은 한 영화에 대해 평점을 4.0을 부여한 사용자ID만으로 문장을 구성하여 사용자 벡터를 구하고 상품추천을 한 방법이다. 실험결과 영화에 대해 3점을 부여한 사용자ID만으로 문장을 구성하는 경우(wvCF3.0)가 가장 예측정확도가 높은 결과를 보였지만 전체적으로 보면 높은 점수대를 부여한 사용자ID로 문장을 구성하는 경우가 하위 점수를 부여한 사용자ID로 문장을 구성하는 경우에 비해 예측정확도가 높은 경향을 나타내고 있다.

다음은 일점점수 이상을 부여한 사용자들로 문장을 확장 구성하여 평균 MAE를 비교하였다.

표 3. 문장 확장 방안의 평균 MAE
Table 3. Mean MAE of sentence expansion method

fold	MAE
wvCFone	0.8372
wvCFtwo	0.8363
wvCFthree	0.8362
wvCFfour	0.8362
wvCFfive	0.8362
wvCFsix	0.8362
wvCFseven	0.8362
wvCFfull	0.8

<표 3>에서 wvCFone은 영화에 대해 4점을 부여한 사용자ID만으로 문장을 구성하는 방법으로 <표 2>의 wvCF4.0과 같다. wvCFtwo는 영화에 4점을 부여한 사용자ID로 한 문장을 만들고, 영화에 3.5점을 부여한 사용자ID로 한 문장을 만들어 영화 당 2개 문장을 구성한다. wvCFthree는 영화에 4점 부여 사용자ID로 한 문장, 3.5점 부여한 사용자ID로 한 문장, 3점 부여 사용자ID로 한 문장을 구성하여 영화 당 3개 문장으로 확장한 방법이다. 실험결과를 보면 문장을 확장하는 방법이 예측정확도를 향상시킴을 알 수 있다. 3개 문장 이상을 확장한 경우에는 정확도 향상이 거의 없음을 볼 수 있다.

표 4. (a) uCF와 wvCF3.0의 쌍체 t-검정 결과
Table 4. (a) Results of pair-wise t-test of uCF and wvCF3.0

	mean	var.	t-value	p-value
uCF	1,0125	0.00083	36.7	0.00000
wvCF3.0	0.8369	0.00034		

(b) wvCF3.0과 wvCFthree의 쌍체 t-검정 결과
(b) Results of pair-wise t-test of wvCF3.0 and wvCFthree

	mean	var.	t-value	p-value
wvCF3.0	0.8369	0.00034	29	0.00000
wvCFthree	0.8362	0.00034		

예측 정확도 차이에 통계적 유의성 검정을 위해, 쌍체 t-검정을 uCF와 wvCF3.0, wvCF3.0과 wvCFthree에 대해 실시하여, <표 4>의 결과를 얻었다. 쌍체 t-검정의 결과 통계적으로 유의한 결과를 얻었다. 결과적으로 Word2Vec을 이용한 사용자 벡터기반 협업필터링 방식이 예측 정확도를 크게 높일 수 있음을 보여 주었다. Word2Vec을 이용하기 위한 문장 구성 방법은 중간평점 이상인 3점 이상을 부여한 사용자들을 대상으로 문장을 확장하는 방법이 가장 효과적인 것으로 나타났다.

5. 결 론

추천 알고리즘의 예측 정확도를 높이는 것은 상품추천시스템 영역에서 주요 과제 중 하나이다[3]. 상품추천시스템에서 많이 활용되고 있는 방법으로 협업필터링 방식이 있다[1,2]. 본 연구에서는 사용자기반 협업필터링 방식의 예측 정확도를 높이기 위해서, 사용자 유사도 측정에 Word2Vec의 사용자 벡터를 사용하는 방법에 대해 제안하였다. Word2Vec은 문장 내 단어 간의 연관성을 파악하여 단어를 벡터로 변환하여 준다. 어떤 상품을 구매한 사용자 집합이 있을 때, 사용자를 단어로 대응하고 상품별 구매 사용자 집합을 문장으로 대응하면 Word2Vec을 이용해 사용자 벡터를 구할 수 있다.

사용자 벡터의 유사성을 이용한 사용자기반 협업필터링 방식을 영화 평점 정보가 있는 filmtrust 데이터에 적용하여 실험한 결과 예측 정확도가 기존 사용자기반 협업필터링 방식에 비해 크게 높아짐을 보였다. 특히 중간 평점 이상을 부여한 영화 사용자만을 대상으로, 평점 당 문장을 구성하여 영화 당 문장의 수를 여러 개로 만드는 문장 확장 방법이 가장 효과적인 것으로 실험결과 나타났다.

한정된 데이터에 대한 실험결과를 일반화할 수는 없지만 Word2Vec을 사용자기반 추천시스템에 적용하는 경우 예측 정확도 향상에 효과적임을 연구결과가 보여 주고 있다.

References

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

[2] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, *EigenTaste: a constant time collaborative filtering algorithm,* Information Retrieval, Vol. 4, No. 2, pp. 133-151, 2001.

[3] B. S. Kang, *The study of recommendation algorithm's predictive accuracy improvement using structural holes on trust-based social networks,* Journal of Knowledge Information Technology and Systems, Vol. 12, No. 1, pp. 209-217, 2017.

[4] J. M. Kim, and J. H. Lee, *Text document classification based on recurrent neural network using word2vec*, Journal of Korean Institute of Intelligent Systems, Vol. 27, No. 6, pp. 560-565, 2017.

[5] C. Heo, and S. Y. Ohn, *A novel method for constructing sentiment dictionaries using word2vec and label propagation*, The Journal of Korean Institute of Next Generation Computing, Vol. 13, No. 2, pp. 93-101, 2017.

[6] D. W. Kim, and M. W. Koo, *Categorization of korean news articles based on convolutional neural networks using doc2vec and word2vec,* Journal of Korean Institute of Information Scientists and Engineers, Vol. 44, No. 7, pp. 742-747, 2017.

[7] W. J. Kim, D. H. Kim, and H. W, Jang, *Semantic extension search for documents using the word2vec*, Journal of the Korean Contents Associations, Vol. 16, No. 10, pp. 687-692, 2016.

[8] J. Y. Ahn, J. H. Bae, N. K. Nam, and M. Song, *A study of 'emotion trigger' by text mining techniques*, Journal of Intelligence and Information Systems, Vol. 21, No. 2, pp. 69-92, 2015.

[9] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality,* Advances in Neural Information Processing Systems 26, pp. 3111-3119, 2013.

[10] B. Oren, K. Noam, *Item2vec: neural item embedding for collaborative filtering,* 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing, pp. 1-6 Sep, 2016.

[11] V. Kuzmin, *Item2vec-based approach to a recommender system*, Bachelor's Thesis, University of Tartu, 2017.

[12] G. Guo, J. Zhang, and N. Yorke-Smith, *A novel bayesian similarity measure for recommender systems*, Proceedings of the 23th International Joint Conference on Artificial Intelligence, pp. 2619-2625, 2013.

[13] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, *An algorithm framework for performing collaborative filtering*, Proceedings of the 1999 Conference on Research and Development in Information Retrieval, pp. 230-237, 1999.

[14] Gensim, https://radimrehurek.com/gensim/. Sep. 2017.

[15] Librec, http://www.librec.net/datasets.html. Oct. 2016.

Journal of Knowledge Information Technology and Systems(JKITS), Vol. 13, No. 1, pp. 169-176, February 2018

Word2Vec을 이용한 사용자기반 협업필터링의 예측 정확도 개선

강부식

목원대학교 서비스경영학부

요 약

Word2Vec은 최근 텍스트 마이닝에서 가장 활발하게 사용되고 있는 방법이다. 텍스트 문장내의 단어 간의 연관성을 파악하여 문장의 단어를 벡터로 변환한다. 벡터 공간에서 유사한 단어일수록 가까이 위치하게 된다. 추천 알고리즘의 예측 정확도를 높이는 것은 상품추천시스템 영역에서 주요 과제 중 하나이다. 사용자 기반 협업필터링은 선호도 유사성이 높은 이웃 사용자의 상품 선호 정보를 이용하여 상품을 추천한다. 본 연구에서는 사용자기반 협업필터링 방식의 예측 정확도를 높이기 위해서 기존의 상품 정보를 이용하여 사용자 유사도를 계산하는 대신 Word2Vec의 사용자 벡터를 이용하여 사용자 유사도를 계산하는 방법에 대해 제안하였다. Word2Vec을 사용하기 위해서는 먼저 문장을 구분하고 문장내의 의미 있는 단어의 조합인 말뭉치를 찾아낸다. 사용자기반 영화추천방식에서 Word2Vec을 사용하기 위해서는, 먼저 동일 영화를 본 사용자 조합을 구하여 사용자를 단어로, 개별 영화의 사용자 조합을 문장의 말뭉치로 대체한다. 이때 문장을 구성하는 방안에 여러 가지가 있다. 개별 영화 당 하나의 문장을 구성하는 방안과 개별 영화당 여러 개의 문장을 구성하는 문장 확장 방안이 있다. 문장의 말뭉치를 입력으로 받아 Word2Vec으로 사용자 벡터를 구하면, 사용자 벡터의 상관관계로 사용자의 유사도를 구한다. 이를 이용하여 사용자기반 협업필터링으로 상품 추천을 한다. 제안한 방안의 검증을 위해 filmtrust 데이터에 적용하여 실험하였다. 10-겹 상호검증을 3회 실시한 결과 Word2Vec을 활용한 사용자기반 협업필터링 방식이 예측 정확도를 크게 개선시킴을 알 수 있었다. 또한 중간 이상의 평점을 부여한 사용자만을 대상으로 하여 개별 영화당 여러 개의 문장으로 확장하는 문장 확장 방식(wvCFthree)이 개별 문장 방식(wvCF3.0)에 비해 예측 정확도를 높임을 실험결과 알 수 있었다. 통계적 유의성을 검정하기 위해 uCF와 wvCF3.0, wvCF3.0과 wvCFthree에 대한 쌍체 t-검정을 실시한 결과 통계적으로 유의함을 확인하였다.

Boo Sik Kang received the bachelor's degree in the Department of Industrial Engineering from the KyungHee University in 1985. He received the M.S. degree and the Ph.D. degree in the Department of Industrial Engineering from KAIST in 1989 and 2000, respectively. From 1989 to 2001, he was a researcher at Korea Telecom. He was a professor in the Division of Service Management at MokWon University from 2001 to 2018. His current research interests include data mining, customer relationship management, service quality management. He is a life member of the KKITS.

E-mail address: bookang@mokwon.ac.kr