Journal of The Korea Society of Computer and Information
Vol. 21 No. 7, pp. 9-16, July 2016
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.7.009

Addressing the New User Problem of Recommender Systems Based on Word Embedding Learning and Skip-gram Modelling

Su-Mi Shin*, Kyung-Chang Kim**

Abstract

Collaborative filtering(CF) uses the purchase or item rating history of other users, but does not need additional properties or attributes of users and items. Hence CF is known th be the most successful recommendation technology. But conventional CF approach has some significant weakness, such as the new user problem. In this paper, we propose a approach using word embedding with skip-gram for learning distributed item representations. In particular, we show that this approach can be used to capture precise item for solving the "new user problem." The proposed approach has been tested on the Movielens databases. We compare the performance of the user based CF, item based CF and our approach by observing the change of recommendation results according to the different number of item rating information. The experimental results shows the improvement in our approach in measuring the precision applied to new user problem situations.

▸ Keyword : Recommender system, Word Embedding, Skip-gram, New User Problem

I. Introduction

추천 시스템(Recommendation System)이란 사용자들의 관심사 및 취향을 고려하여 고객의 성향에 부합하는 아이템을 예측하여 추천해주는 시스템으로 온라인 쇼핑몰이나 디지털 콘텐츠 서비스 업체를 중심으로 소비자의 의사결정을 도와주는 추천 서비스들이 지속적으로 발전시키고 있다. 추천 시스템에서 널리 이용되는 기법은 협업 필터링(Collaborative Filtering)이다[1]. 협업 필터링은 아이템의 속성이나 사용자의 속성에 제약이 없으며 다른 사용자들의 아이템선호 정보를 다양하게 활용할 수 있는 기회를 제공한다. 협업 필터링기법은 어떤 소비자가 선호하는 아이템이 있을 때, 이 소비자와 비슷한 취향을 가진 다른 사용자들이 선호하는 아이템을 이 소비자가 선호할 확률이 높은 점에 착안한 것이며, 이를 이용해서 비슷한 취향을 가진 사용자들에게 아이템을 추천한다. 따라서 협업 필터링은 비슷한 관심사를 갖는 사용자를 찾고, 이웃 사용자들이 아이템에 대해 평가한 정보를 토대로 아이템을 분석하는 과정이 필요하다. 협업 필터링 기법은 추천시스템에 가장먼저 도입된 성공기법이라고 평가되고 있지만 기본적으로 극복해야할 문제가 있다. 협업 필터링 기법이 추천을 위해 고려하는 사항은 사용자에 의한 아이템의 선호도이므로 구매나 평점 등 아이템의 선호도 정보가 많이 존재하는 경우 높은 추천 성능을 보이지만 아이템 선호도 정보가 없는 신규 사용자에게 아이템을 추천할 수 없다는 단점이 있다. 이와 같이 신규 사용자에 대한 추천이 불가능한 문제를 새로운 사용자 문제(New user problem)라고 하는데[2] 새로운 사용자 문제는 사용자가 무엇을 좋아할지에 대한 정보가 부족한 상태에서 추천할 만한 아이템을 예측해야하기 때문에 근본적인 해결이 어렵다. 이런 이유로 대부분의 연구들은 새로운 사용자 문제를 해결하기 위해 사용자의 인구통계학적 정보를 사용하는 방법[3]이나, 경험에 대한 개방성이나 외향성 등 과 같은 개인정보를 이용하는 방법[4], 초기에 사용자

•First Author: Su-Mi Shin, Corresponding Author: Kyung-Chang Kim
*Su-Mi Shin(sumi@kisti.re.kr), Dept. of Computer Engineering, Hongik University / Dept. of Information Service, KISTI
**Kyung-Chang Kim (kckim@hongik.ac.kr), Dept. of Computer Engineering, Hongik University
• Received: 2016. 06. 20, Revised: 2016. 07. 04, Accepted: 2016. 07. 09.

들에게 아이템 집합에 평가하게 하거나 여러 질문의 대답을 유도하는 적극 학습(active learning)을 이용하는 방법[5, 6] 등 부가적인 정보를 사용한다. 그러나 온라인 쇼핑몰이나 디지털 콘텐츠 서비스를 이용하려는 사용자로부터 부가정보를 요구하는 것은 사용자의 만족도를 저하시키는 원인이 될 수 있다.

본 연구에서는 사용자의 인구통계학 정보 등 부가적인 프로파일 정보를 추가로 요구하지 않으면서 선호하는 아이템간의 관계를 분석하여 아이템을 추천하는 방법으로 동일한 사용자로부터 소비되거나 평가된 아이템간 연관성을 학습하여 벡터로 변환하는 워드 임베딩 기법을 기반으로 하는 새로운 추천 시스템을 제안한다.

연구의 기본적인 아이디어는 동일한 사용자가 여러 아이템에 대해 평가한 아이템의 선호도를 시간 순으로 정렬하였을 때 일정범위에 있는 아이템의 선호도는 서로 연관되어 있으며 이 아이템 선호도 하나하나를 객체로 n차원에 임베딩하면 각 아이템의 선호도간의 유사도를 수치로 표현할 수 있다는 것이다.

예를 들어 한명의 사용자가 A라는 영화를 5점으로 평가하고 연달아 B라는 영화를 2점으로 평가했다면 'A영화 5점'이라는 객체와 "B영화 2점"이라는 객체는 서로 관련이 있으며 학습을 통해 임의의 차원에서 가까운 위치에 임베딩 되도록 함으로써 관련성 높은 객체라는 것을 수치화 할 수 있다.

객체를 n차원에 임베딩하는 방법으로는 하나의 주어진 단어로부터 주위 단어를 예측하는 Skip-gram 알고리즘을 이용하며, 각 객체별로 생성된 벡터 값을 기반으로 거리가 가까운 객체를 찾고 이를 사용자에게 추천하는 아이템 목록을 생성한다.

본 논문은 다음과 같이 구성된다. 2장에서는 협업필터링과 워드 임베딩 기법을 알아보고 3장에서는 본 연구에서 제안하는 워드 임베딩 기반의 추천시스템에 대해 상세히 기술한다. 4장에서는 제안한 방법을 평가하고 5장에서는 향후 연구방향을 제시한다.

II. Background and Related Works

1. Collaborative Filtering and The New User Problem

협업필터링 기법은 목표 사용자와 유사한 구매이력을 보이는 이웃 사용자들이 보여준 아이템에 대한 선호도를 바탕으로 목표 사용자에게 유용한 아이템을 추천하는 방법이다. 즉, 협업필터링 기법은 사람들의 관심사나 취향이 무작위로 분포된 것이 아니라 그룹이나 사람들의 취향 사이에는 일반적인 트렌드와 패턴이 있다는 가정에서 출발하는 것이다. 실제로 사람들은 주위 친구들이나 인터넷을 통한 다른 사용자들과의 커뮤니케이션 등을 통해서 아이템이나 경험에 대한 추천을 받고 이를 아이템 선택에 참고한다. 협업 필터링은 특정 사용자의 선호도를 예측하기 위해 다른 사용자의 과거 선호도를 예측해내는 방식으로, 제록스 팔로알토 연구소의 Nichols등에 의해서 개발된 Tapsetry가 협업 필터링을 적용한 최초의 시스템이다. 협업 필터링 기법은 아이템의 속성 분석이나 사용자의 속성에 제약이 없고 이웃집단에 속한 사용자들의 다양한 정보를 활용할 수 있는 기회를 제공한다. 협업필터링은 사용자 기반 협업필터링(user-based collaborative filtering)과 아이템 기반 협업필터링(item-based collaborative filtering)이 있다.

사용자 기반 협업필터링은 다음의 과정으로 실행된다.

(ⅰ) 입력데이터 구성 : 사용자가 기존에 구매하거나 평점(Rating)을 매긴 표 1과 같은 데이터를 바탕으로 표 2과 같이 사용자-아이템 행렬(User-item matrix)을 구성한다.

Table 1. Examples of the Item Rating Data

Alice: Shrek(Rating:4), SnowWhite(Rating:5), Superman(Rating:3)
Bob : SnowWhite(Rating:3), Superman(Rating:5), spiderman(Rating:5)
Chris: Superman(Rating:5), spiderman(Rating:5)
Toby: Superman(Rating:3)

Table 2. User-Item Matrix

	Shrek	SnowWhite	Superman	spiderman
Alice	5	5	5	-
Bob	4	3	5	4
Chris	3	-	5	-
Toby	-	-	3	-

(ⅱ) 이웃집단 탐색 : 같은 아이템에 대한 평점을 기준으로 다른 사용자들의 평점과 비교하여 사용자간 유사도를 구하고 이를 바탕으로 이웃집단을 구성한다. 표 2의 경우 Toby를 대상으로 이웃집단을 선정하면 Alice, Bob, Chris 모두 Toby가 평가한 Superman을 동일한 점수로 평가하고 있어 세 사용자 모두 동일한 유사도 값을 가진 이웃으로 선정된다.

(ⅲ) 추천목록 생성 : 아이템들에 대한 예상 선호도를 이웃집단을 토대로 계산하여 추천 목록을 생성한다.

표 2를 기준으로 추천 목록을 생성하면 Alice, Bob, Chris가 평가한 모든 영화가 추천 대상이 되며 이들 중 평점평균이 4로 동일하므로 모든 영화가 추천대상 영화가 된다.

Toby의 예를 보면 이웃 집단 탐색과정에서 너무 많은 이웃이 유사 이웃으로 선정되는 문제가 있으며 그에 따라 너무 많은 영화가 추천 대상이 되는 문제를 갖는다.

아이템 기반 협업필터링은 사용자 기반 협업필터링 과정에서 (ⅱ), (ⅲ)이 다음과 같이 변경된다.

(ⅱ) 유사 아이템 탐색 : 서로 다른 사용자가 함께 구매하거나 유사한 평점을 매긴 정도에 따라 아이템 유사도를 계산한다.

(ⅲ) 추천목록 생성 : 사용자가 기존에 구매하거나 경험했던 아이템과의 유사도를 토대로 계산하여 추천 목록을 생성한다.

협업 필터링 기법은 가장먼저 도입된 성공기법이라고 평가

되고 있으며 k-means나 연관규칙 분석 등의 결합될 수 있는 기반 기법으로 활용된다[13][14].

그러나 협업필터링은 위의 Toby의 예와 같이 근본적으로 신규 고객 및 신규아이템에 대한 추천이 어렵다는 문제를 가지고 있다. 협업 필터링 기법에 필요한 값은 사용자가 기존에 구매하거나 평가했던 아이템에 대한 사용자-아이템 행렬 값인데 이때 아이템에 대한 선호도나 평점 정보가 없는 신규 아이템의 경우 다른 사용자로부터 선호되었다는 근거가 없기 때문에 추천할 수 없다는 문제가 있다. 이와 같이 신규 아이템에 대한 추천이 불가능한 문제를 Item-Side Cold Start 문제, 또는 새로운 아이템 문제(New item problem)라고 한다. 협업 필터링 기반 추천 시스템의 또 다른 문제는 신규 사용자에 대한 추천을 제공하지 못한다는 점이다. 이는 협업 필터링 알고리즘이 사용자간의 유사도를 기반으로 이웃집단을 도출한 후 아이템을 추천하기 때문인데, 신규 사용자는 아이템들에 대한 어떠한 취향이나 선호도 보이지 않은 상황이기 때문에 이웃집단을 계산할 수 없는 한계 때문에 발생한다. 이와 같이 신규 사용자에게 아이템 추천이 불가능한 문제를 User-Side Cold Start 문제 또는 새로운 사용자 문제(New user problem)라고 한다[2].

2. Related Works

새로운 사용자 문제는 협업 필터링 기반의 추천 시스템에서 중요한 이슈로 여러 연구에서 해당 문제를 다루고 있다. 새로운 사용자 문제는 사용자가 무엇을 좋아할지에 대한 정보가 부족한 상태에서 추천할 만한 아이템을 예측해야하기 때문에 근본적인 해결이 어렵다. 이런 이유로 대부분의 연구들은 새로운 사용자 문제를 해결하기 위해 사용자의 인구통계학적 정보를 사용하는 방법[3]이나, 경험에 대한 개방성이나 외향성 등 과 같은 개인정보를 이용하는 방법[4], 초기에 사용자들에게 아이템 집합에 평가하게 하거나 여러 질문의 대답을 유도하는 적극 학습(active learning)을 이용하는 방법[5, 6] 등 부가적인 정보를 사용한다. Almazro 등은 인구 통계 데이터를 이용하여 인구학 및 협력 필터링의 하이브리드 방식을 도입 하였다[7]. 이 연구에서는 인구 통계 학적 특성, 예를 들어, 사용자의 나이, 학생 여부, 자녀보유 여부, 성별 등을 기반으로 영화의 장르를 구분했다. Vozalis 등은 사용자 프로파일에 사용자 인구 벡터를 가산 및 유사도의 계산을 위해 협력 필터링 알고리즘을 포함하여, K 최근 접 이웃의 수정 된 버전을 보여 주었다[8]. 소셜 서브 커뮤니티(social sub-community)와 온톨로지 의사결정 모델을 추가 정보로 활용한 연구도 있다[9]. 소셜 서브 커뮤니티는 사용자의 히스토리 데이터와 서로 간의 관계에 따라 구분되었다. 그러나 이렇게 부가적인 정보를 활용하는 방안은 부가정보 획득 시 사용자의 불편을 초래할 수 있다.

III. A Recommender System Based on Word Embedding Learning

1. Word Embedding and Skip-gram Model

워드 임베딩이란 하나의 단어를 인공 신경망을 이용하여 벡터 공간상에 나타낼 수 있는 값으로 변환하는 것을 의미한다[10]. 워드 임베딩은 텍스트 마이닝이나 자연어 처리 분야에서 광범위하게 활용되는데 코퍼스(Corpus)의 라이클리후드(Likelihood)를 최대화하는 방향으로 학습한다[11].

Mikolov등은 Skip-gram 모델을 신경망 학습을 통한 효과적인 워드 임베딩 기법이라고 소개[12]한다. Skip-gram 모델은 하나의 주어진 단어로부터 주위 단어를 예측하기 위한 모델로 아키텍처는 그림1과 같다.

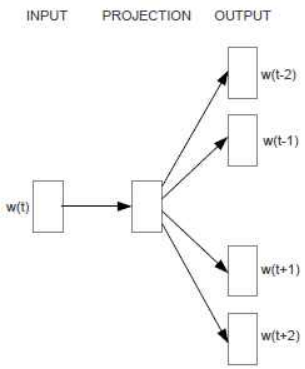

Fig. 1. Skip-gram Architecture

그림 1은 w(t-2), w(t-1), w(t), w(t+1), w(t+2)라는 워드 집합을 학습대상으로 w(t)이라는 단어로 w(t-2), w(t-1), w(t+1), w(t+2)를 예측하는 모델을 학습하는 예로 w(t)를 input으로 w(t-2), w(t-1)와 같이 input의 일정 범위의 앞과 w(t+1), w(t+2)와 같이 input의 일정 범위 뒤에 있는 단어를 output으로 학습한다. Skip-gram 모델을 수식으로 표현하면 식(1)과 같으며 학습단어 w1,w2,w3, . . . ,wT에 대해 확률을 최대화 하는 방향으로 학습한다. 이 식에서 c는 output의 범위로 컨텍스트(context)라고 한다. 컨텍스트는 보통 기준 단어의 주변 단어를 가리킨다.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c, j\neq 0}\log p(w_{t+j}|w_t) \qquad (1)$$

일반적으로 Skip-gram은 $p(w_{t+j}|w_t)$를 정의하기 위해 수식(2)와 같은 소프트맥스(softmax) 함수를 사용한다.

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^{\top} v_{w_I})}{\sum_{w=1}^{W} \exp({v'_w}^{\top} v_{w_I})} \quad (2)$$

이 식에서 v_w 는 워드 w의 input 벡터표현이고 v'_w는 워드 w의 output 벡터표현이다. 그리고 W는 Vocabulary내의 전체 단어의 개수이다. Mikolov는 소프트맥스를 개선한 기법으로 Hierarchical Softmax와 Negative Sampling기법을 적용한 Skip-gram 모델도 추가로 제안하고 있다[12].

협업필터링 기법이 가지고 있는 신규 사용자문제를 해결하려면 소비되거나 평가된 아이템간의 연관관계를 분석할 수 있는 새로운 분석방법이 필요하다.

본 논문에서는 사용자가 기존에 구매하거나 경험했던 아이템의 평점(Rating)간의 관계를 분석하여 신규 사용자에게 아이템을 추천하는 워드 임베딩 모델 기반의 추천 방법을 제안한다. 워드 임베딩 모델을 이용한 추천방법은 아이템의 속성이나 사용자의 속성 등 추가적인 정보가 필요 없으며 다른 사용자들의 아이템선호 정보를 다양하게 활용할 수 있다는 점에서 협업필터링 기법의 확장으로 볼 수 있다. 협업 필터링 중 아이템 기반 협업 필터링 방식은 서로 다른 사용자가 함께 구매하거나 유사한 평점을 매긴 수준을 기반으로 아이템간 유사도를 계산한다. 본 논문에서 제안하는 방식은 아이템간 유사도를 계산하기 위하여 근접한 신경망 학습을 적용하여, 하나의 단어를 인공 신경망을 이용하여 벡터 공간상에 나타낼 수 있는 값으로 변환하는 방식으로 동일한 사용자가 시간 순으로 인접한 아이템 평가간의 라이클리후드(Likelihood)를 최대화 하도록 학습한다.

2. A Recommender System Based on Word Embedding Learning

본 연구에서 제안하는 워드 임베딩 기법을 이용한 추천기법의 수행과정은 그림 2와 같으며 크게 3단계로 구분된다.

Fig. 2. The Precess of the proposed recommender system

2.1 Build Item Preference List

첫 번째 단계에서는 사용자가 기존에 구매하거나 경험했던 아이템의 평점(Rating)을 이용하여 사용자별 아이템 선호 목록을 학습 데이터로 생성한다. 학습데이터는 다음과 같은 두 가지 가정을 기반으로 한다.

(ⅰ) 아이템 평점을 사용자순, 시간 순으로 정렬하면 주위에 있는 아이템 평점은 서로 관련이 있다.

(ⅱ) 아이템에 대한 평점을 각 아이템의 선호도 속성 값으로 부여하고 속성 값이 다른 아이템을 각각 다른 객체로 정의하면 표 3와 같이 특성 명확하게 구분될 수 있다. 속성은 긍정, 부정으로 구분하거나 실제 평점 값을 사용한다.

Table 3. Items with distinctive Property

어벤져스.4 ≈ 어벤져스.5
우유.Positive ≠ 우유.Negative
어벤져스.Negative ≈ 어벤져스2.Negative
양들의 침묵.Negative ≈ 나 홀로 집에.Positive

평점을 긍정, 부정 속성으로 구분하는 방법은 평점의 평균, 분포 등을 고려하여 다양화 할 수 있다.

2.2 Train Item Preference Relation

두 번째 단계에서는 선호도를 속성으로 가진 각 아이템들이 어떻게 같이 사용되었는지를 학습하여 벡터를 생성한다. 표 4은과 2차원 벡터의 생성 예이다.

Table 4. Examples of Item Preference Vector

양들의 침묵.Positive = -0.099752476, 3.894467378
매트릭스.Negative = -0.482424537, 3.809729331

벡터생성은 사용자별 아이템 선호도 목록을 워드 임베딩 기법을 이용해 학습한다. 본 연구에서는 Skip-gram 모델을 적용한다.

Skip-gram 학습 예를 들어보자. 어떤 사용자의 아이템 선호도 목록을 “Cinderella.5, Time Tracers.4, Back to the Future.5, Meet Joe Black.3, Last Days of Disco.5”라고 가정한다. 이때 “Cinderella.5”는 선호도 값이 5인 Cinderella라는 아이템을 나타낸다. 컨텍스트를 1로 하여 기준 객체의 앞, 뒤 1개씩 2개의 객체가 기준 객체와 관련 있다고 학습한다면 맨 처음 “Cinderella.5”를 Input으로 “Time Tracers.4”를 Output으로 학습한다. 다음으로 “Time Tracers.4”를 Input으로 “Cinderella.5”를 Output으로 학습하고 다시 “Time Tracers.4”를 Input으로 “Back to the Future.5”를 Output으로 학습을 진행한다. 컨텍스트를 1로 할 때 전체 input과 output 구성은 표5와 같다.

Table 5. Examples of the Training Sets

Input	Output
Cinderella.5	Time Tracers.4
Time Tracers.4	Cinderella.5
Time Tracers.4	Back to the Future.5
Back to the Future.5	Time Tracers.4
Back to the Future.5	Meet Joe Black.3
Meet Joe Black.3	Back to the Future.5
Meet Joe Black.3	Last Days of Disco.5
Last Days of Disco.5	Meet Joe Black.3

Skip-gram은 Input 단어로부터 컨텍스트 단어를 예측한다. 즉, 우리는 “Cinderella.5”로부터 “Time Tracers.4“를 예측하고, “Time Tracers.4”로부터 “Cinderella.5”와 “Back to the Future.5”을 예측한다. 마찬가지로 “Back to the Future.5”으로부터 “Time Tracers.4”과 “Meet Joe Black.3”을 예측한다.

표 6는 세부적인 학습 알고리즘이다.

Table 6. Learning Algorithm

Input
c : context size
U : Set of users
$RItemList_a$: Item Preference List of User a
Output
θ : matrix of vector representation

For each $u \in U$ do
  For each $RItem_i \in RItemList_a$ do
    For each $RItem_j \in$
    $RItemList_u$[index($RItem_i$)-c : index($RItem_i$)+c] do
      $J(\theta) = -\log Pr(RItem_j | \theta(RItem_i))$
      $\theta = \theta - \alpha \times \frac{\delta J}{\delta \theta}$
    end for
  end for
end for

2.3 Build Recommendation Item Set

세 번째 단계에서는 임의의 아이템 평점을 가지고 있는 사용자에게 그 주변 아이템을 추천한다. 추천 아이템은 임의의 아이템 평점과 선호도 속성 값이 긍정적인 아이템에 대해 벡터 간 유사도가 높은 순서로 정한다. 표 7은 추천 아이템 리스트를 생성하는 알고리즘이다.

Table 7. Algorithm of Building Recommendation List

Input
$RItemList_a$: Item Preference List of User a
$TrainedModel_V$: Trained Vector model with Data set V

Output
RList : Recommendation List

For each $RItem_i \in RItemList_a$ do
  For each $RItem_j \in TrainedModel_V$
    if $RItem_i \neq RItem_j$
      s := similarity ($RItem_i, RItem_j$)
      RList($RItem_j$) += s
  end for
end for
sort(RList)

IV. Simulation and Evaluation

1. Test Design

본 연구에서는 실제 사용자를 통해 수집된 데이터를 기반으로 제안된 추천기법을 적용하여 그 성능을 측정하고자 하였으며 미네소타 대학의 GroupLens Research 프로젝트의 일환으로 수집된 MovieLens 데이터 중 100만 건의 평점이 포함된 1 million 데이터를 이용하였다. MovieLens 1 million 데이터는 6,040명의 사용자가 3,952편의 영화에 대해 평가한 1,000,209개의 평점으로 구성되어 있다.

본 연구에서 제안한 기법의 성능은 사용자기반 협업필터링, 아이템 기반협업 필터링과 비교하였으며 새로운 사용자 문제를 중심으로 평가하기 위하여 영화 평점이 1개인 경우부터 20개까지 사용자의 평점 개수를 늘려가며 성능을 측정하였다.

본 연구에서는 최종 생성된 추천 목록 리스트가 얼마나 실제 구매 아이템과 일치하는지를 확인할 수 있는 평가척도인 정확도를 사용하여 성능을 평가하였다. 정확도는 추천 리스트 중에서 몇 개의 아이템이 고객이 실제로 구매했는지를 나타내는 평가방법이다. 사용자 a가 실제 이용한 아이템 집합을 V_a, 사용자 a에게 추천된 아이템 집합을 R_a로 정의할 때 사용자 a에 대한 정확도 PR_a는 수식 3과 같다.

$$PR_a = \frac{(V_a \cap R_a)}{|R_a|} \quad (3)$$

2. Evaluation

Skip-gram 알고리즘을 기반으로 컨텍스트를 5로 고정하여 사용자별 영화 평점인 아이템 선호 목록을 학습한 2차원 벡터를 생성하였다. 표 8은 제안된 워드 임베딩 기법을 이용하여 생성된 벡터 예이다. 학습결과 16,912개의 아이템 선호에 대한 벡터가 생성되었다.

Table 8. Trained Results

Item Preference	Vector
American_Beauty.5	1.49815253496, 0.780713597517
Star_Wars_Episode IV.5	2.19466013429, 0.61229204485
E.T.4	1.84077218, 0.663774688
Jurassic_Park .3	0.941820096922, 1.16861520503
Twelve_Monkeys.4	1.861578974, 0.677268623
Forrest_Gump.4	1.20425350856m 0.894458481628

그림 3은 'Indiana Jones 3점'과 가까이에 위치한 영화 평점을 2차원 평면에 표현한 예이다. 'When Harry met Sally 3점', 'E.T 4점', 'Twelve Monkeys 4점' 등이 가까이 위치하는 것을 확인할 수 있다.

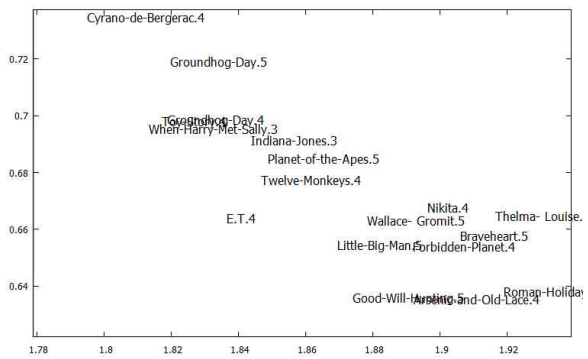

Fig. 3. Vector Representations of Movie Ratings

본 연구에서 제안하는 기법의 성능을 기존 추천 기법들과 비교 평가하기 위해 두 가지 실험을 시행하였다.

첫 번째 실험에서는 평점 정보가 하나 밖에 없는 신규 사용자를 가정하고 추천 성능을 비교하기 위하여 동일한 영화를 본 사용자 10명에 대하여 각 기법별로 50개의 영화 추천 목록을 생성한 후 정확도를 비교하였다.

Table 9. 1-rating based recommendation results

구분	사용자기반 협업필터링	아이템기반 협업필터링	제안기법
user1	0.24	0.14	0.62
user2	0.12	0.14	0.50
user3	0.18	0.14	0.20
user4	0.22	0.18	0.54
user5	0.06	0.10	0.14
user6	0.06	0.04	0.02
user7	0.10	0.04	0.08
user8	0.06	0.10	0.18
user9	0.14	0.20	0.36
user10	0.18	0.12	0.14
평균	0.136	0.109	0.278

표 9는 Indiana Jones를 5점으로 평가한 10명의 사용자에게 이 평점만을 바탕으로 50개씩의 영화를 추천한 결과이다. 제안된 기법의 경우 27.8%의 정확도 수준을 보이며 사용자 기반 협업필터링과 비교하면 104.4%, 아이템 기반 협업 필터링과 비교하면 155.0% 성능이 개선되었다.

Table 10. Recommendation Performances for New User

평가 아이템 수	사용자기반 협업필터링	아이템기반 협업필터링	제안기법
1	0.0610	0.0443	0.1030
2	0.0732	0.0820	0.1000
3	0.0891	0.1151	0.1013

두 번째 실험은 전체 데이터 중 10%의 테스트 데이터에 대하여 평점개수를 1개에서 3개까지 증가시키면서 50개의 영화 추천 목록을 생성한 후 정확도를 측정하였다. 표 10는 사용자의 평점개수를 증가시키면서 다른 기법과 정확도를 비교한 결과이다.

사용자 기반 협업필터링과 비교하면 평가 아이템이 한 개인 경우 68.8%, 평가 아이템이 두 개인 경우 36.6% 성능이 향상되었다. 아이템기반 협업필터링과 비교하면 평가 아이템이 한 개인 경우 132.5%, 평가 아이템이 두 개인 경우 21.6% 성능이 향상되었다. 다만 평가아이템 수가 3개 이상인 경우 제안기법보다 좋은 성능을 보여주고 있어 평가 아이템 수가 2개 이하인 새로운 사용자에게만 제안방법이 효과적임을 확인하였다.

V. Conclusion

본 연구에서는 협업 필터링 기법이 아이템 선호도 정보가 없는 신규 사용자나 신규아이템의 경우 추천할 수 없다는 단점을 극복하기 위하여 아이템의 연관성을 미리 학습하여 선호정보가 1개, 2개 수준의 신규사용자라 하더라도 신경망 학습결과에 따라 근접한 아이템을 추천할 수 있는 새로운 추천 시스템을 제

안하였다. 학습 모델은 아이템간의 연관관계에 관심을 두고 각 아이템의 선호도에 대한 근접도를 수치화하기 위해 아이템평가 정보를 n차원에 임베딩 하는 기법인 워드 임베딩 기법을 이용하였으며 기준 되는 아이템의 주위에 어떤 아이템이 나올지 예측하기 위해 사용하는 Skip-gram 알고리즘을 이용하여 관계를 학습하였다. 제안기법은 3개의 과정으로 구성된다. 첫 번째 단계에서 사용자가 기존에 구매하거나 경험했던 아이템의 평점을 이용하여 사용자별로 선호도 속성을 가진 아이템리스트인 아이템 선호도 목록을 생성한다. 두 번째 단계에서는 선호도를 속성으로 가진 각 아이템들이 어떻게 같이 사용되었는지를 학습하여 벡터로 표현한다. 세 번째 단계에서는 임의의 아이템 평점을 가지고 있는 사용자에게 그 주변 아이템을 추천한다.

제안한 기법은 두 가지의 실험을 통해 평가하였으며 사용자 기반 협업필터링 기법에 비해 새로운 사용자 문제 상황에서 정확도가 20%이상 향상되는 결과를 보였다.

본 연구의 향후 방향은 선호도를 긍정, 부정으로 구분하거나 학습 알고리즘을 다양화하여 추천 성능에 미치는 결과를 확인하여 평점이 3개 이상인 사용자에 대한 추천 성능을 현재보다 개선하는 것이다. Skip-gram 알고리즘을 변형하거나 다른 학습 알고리즘을 적용하고 특히 Skip-gram기법이 단순함을 개선하는 방안으로 Hidden Layer를 한정하지 않고 고차원 Layer를 적용한 딥러닝 기법의 학습모델을 적용하여 추천 정확도에 미치는 영향을 살펴본다.

REFERENCES

[1] Daniel Billsus and Michael J. Pazzani, “Learning collaborative information filters.” In Proceedings of the 15th International Conference on Machine Learning, ICML’98, pages 46-54, 1998.

[2] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel., "Probabilistic memory-based collaborative filtering", IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 56-69, JANUARY, 2004.

[3] Paolo Massa, Paolo Massa, and Bobby Bhattacharjee., “Using trust in recommender systems: An experimental analysis.” In Proceedings of ITRUST 2004, 2995, pp. 221-235, 2004.

[4] Tkalcic, M., Kunaver, M., Koˇsir, A., and Tasic, J., Addressing the new user problem with a personality based user similarity measure. In DEMRA, 2011.

[5] Mehdi Elahi, Matthias Braunhofer, Francesco Ricci, and Marko Tkalcic., Personality-based active learning for collaborative filtering recommender systems. AI*IA 2013: Advances in Artificial Intelligence, pp. 360-371, 2013.

[6] Chae Banseok, “Naver translator, where did you come from?”, 2015, Bloter.

[7] D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, W. Nzoukou., A Survey Paper on Recommender Systems, 2010, arXiv:1006.5278.

[8] M. Vozalis, K.G. Margaritis., Collaborative filtering enhanced by demographic correlation, in: Proceedings of the AIAI Symposium on Professional Practice in AI of the 18th World Computer Congress, 2004.

[9] M. Chen, C. Yang, J. Chen, P. Yi., A method to solve cold-start problemin recommendation system based on social network sub-community and ontology decision model, Proceedings of the 3rd International Conference on Multimedia Technology(ICMT2013),pp.159-166.2013.

[10] Li, Yitan, Xu, Linli., “Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective”, In Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean., “Efficient estimation of word representations in vector space.”, ICLR Workshop, 2013.

[12] Tomas Mikolov., Sutskever. Ilya, Chen. Kai, Corrado. Greg, "Distributed Representations of Words and Phrases and their Compositionality", 2013.

[13] Young Sung Cho, Mi Sug Gu, Keun Ho Ryu, “Development of Personalized Recommendation System using RFM method and k-means Clustering”, Journal of The Korea Society of Computer and Information, Vol. 17 No. 6 , June 2012.

[14] Jong-Hee Kim, Soon-Key Jung, “The Goods Recommendation System based on modified FP-Tree Algorithm”, Journal of The Korea Society of Computer and Information, Vol. 15 No. 11, November, 2012.

Authors

Su-Mi Shin received the B.S. and M.S. degrees in Computer Engineering from Hongik University, Korea, in 1997 and 2005, respectively. She is a Ph.D. student in Dept. of Computer Engineering, Hongik University, Korea.

And she is currently a senior researcher in the Department of Information Service, KISTI, Korea.

She is interested in recommender system, big data big data analysis, databases, data retrieval and Internet of Things.

Kyung-Chang Kim received the B.S. in Computer Science from Hongik University in 1978, M.S. in Computer Science from KAIST in 1980 and Ph.D in Computer Science from University of Texas at Austin in 1990.

Dr. Kim joined the faculty of the Department of Computer Engineering at Hongik University, Seoul, Korea, in 1991. He is currently a Professor in the Department of Computer Engineering, Hongik University. He is interested in main memory databases, sensor databases, web databases, Internet of Things and big data processing.