사용자 선호도와 태그 간 상관도 분석을 통한 태그 기반 협력적 필터링 기법[1)]

이경종° 공기현 이상구

서울대학교

{kjlee, rawtilt, sglee}@europa.snu.ac.kr

Tag-Based Collaborative Filtering Approach Using Analysis of the Correlation Between User's Preference and Tags

요 약

웹의 성장에 따른 기하급수적인 정보의 축적으로 인한 정보과다(Information Overload) 현상의 심화를 해결하기 위해 이루어져 온 많은 연구 중 하나인 추천 시스템은 사용자에게 고수준의 편의성을 제공하기 위한 시스템으로써 발전해 왔다. 그러나 과거에 고도로 집중화되어 관리, 구축되어 오던 정보와는 달리 Web2.0라는 새로운 웹 환경의 도래와 함께 태그, 블로그 등 새로운 형태와 특성을 가지는 정보들이 등장하게 되었다. 웹의 컨텐츠에 대한 메타정보를 사용자가 직접 입력한 Web2.0 기반의 태그 데이터를 활용해서 추천 시스템의 성능을 향상시킬 수 있는 기법을 연구하였다. 추천 기법 중 가장 대표적이고 기초적인 협업 필터링 기법에 태그를 활용하며 태그에 사용자에 대한 중요도를 감안한 가중치 부여 기법에 연구한다. 유사한 성향을 가진 사용자를 식별하는데 있어 태그 집합 간의 유사도를 비교하는 방법을 사용하며 사용자의 성향을 반영하기 위해서 태그와 사용자의 선호도 점수와의 연관성을 분석해서 이를 태그의 가중치로 환산하는 기법을 제안한다.

1. 서 론

인터넷과 웹의 폭발적인 성장 등과 함께 사용자가 이용할 수 있는 정보의 양은 기하급수적으로 증가하였다. 이러한 상황에서 사용자의 정보 요구에 적합한 정보를 찾아주는 역할을 수행하는 시스템인 추천 시스템(Recommendation System)의 역할이 중요시되고 있다. 추천 시스템이란 사용자에게 사용자의 잠재적인 정보 요구에 부합하는 자료를 자동으로 검색, 제공하는 시스템을 말한다[1]. 사용자가 자신이 필요로 하는 정보를 검색 키워드의 형태로 명시하는 일반적인 검색 시스템에 비해, 추천 시스템은 키워드를 입력받지 않으며 사용자의 액션, 패턴 등을 통해서 묵시적으로 유추해 내야하며 정보 검색 시스템보다 더 고수준의 분석 작업이 요구된다[1, 2].

이러한 상황에서 최근 Web2.0의 흐름을 주목해 볼 만하다. Web2.0의 흐름에는 웹상의 시스템 및 애플리케이션을 제작하는데 있어 적용될 수 있는 1) 플랫폼으로서의 네트워크(Network as Platform), 2) 집단 지성(Collective Intelligence)의 활용 등 의 패러다임을 포함한다[3]. 이 중 집단 지성이란 소수의 전문가 집단에 의해서 구축된 데이터가 아니라 다수의 사용자들에 의해서 구성된 데이터를 의미한다. 이러한 데이터에는 사용자의 태그(Tag), 사용자의 평점(User's Rating) 정보들이 포함되며 이러한 데이터는 추천 시스템에 있어서 유용한 정보를 제공하게 된다.

추천 시스템의 성능을 향상시키기 위해서는 사용자의 성향 및 미묘한 감성 등을 잘 표현할 수 있는 데이터가 필수적이다. 일반적으로 태그에는 해당 컨텐츠에 대한 사용자의 성향, 미묘한 감성, 상황 정보들을 포함하고 있으며 이를 활용해서 추천 시스템의 성능에 향상을 시킬 수 있다.

이에 본 논문에서는 태그를 활용해서 추천 기법 중 가장 대표적인 협업 필터링 기법의 성능을 향상시키기 위한 기법에 대한 연구를 수행하였다. 멀티미디어 중에서도 가장 사용자의 성향이 다양한 것으로 알려져 있는 음악을 도메인으로 해서 사용자의 프로파일 정보와 태그 정보를 바탕으로 사용자에게 음악을 추천하는 추천 시스템을 제안한다. 또한, 추천의 정확도를 향상시키기 위해 태그가 사용자에게 가지는 중요도에 따라 가중치를 부여하기 위한 기법을 제안한다.

2. 관련 연구

2.1 추천 시스템과 협업 필터링 기법

추천 시스템이란 특정 사용자에게 사용자가 선호할 것이라 예상되는 아이템들을 사용자에게 제공하는 시스템이다. 추천 시스템은 정보 검색 시스템과 여러 가지 면에서 공통점을 가진다. 그러나, 사용자가 사용자의 정보 요구를 키워드의 형태로 명시적으로 입력하지 않는다는

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성 지원 사업(IITA-2007-C1090-0701-0031)의 연구결과로 수행되었음

점에서 정보 검색 시스템과 차이가 있다. 즉, 추천 시스템은 미리 입력된 사용자의 프로파일 및 행동 패턴 그 이외에 시스템에 축적된 여러 데이터를 이용해서 사용자의 정보 요구를 추출해 내야 한다[2].

[1] 에서는 추천 과정을 정형화해서 정의하고 있다. 간단히 요약하면 해당 사용자가 기존에 선호하는 아이템들과 다른 사용자들에 대한 정보 등을 이용. 사용자가 전에 경험해보지 못한 아이템들에 대한 선호도 수치를 산출해 내는 과정이 된다.

	음악1	음악2	음악3	음악4	음악5
사용자1	9	7	3	?	?
사용자2	2	3	10	3	9
사용자3	10	6	4	10	3
사용자4	7	6	3	9	3

표 1 사용자-아이템 매트릭스

표1은 추천 기법에 기본이 되는 사용자-아이템 매트릭스이다. 각 셀은 각 사용자가 해당 아이템에 대한 선호도 수치(Rating Score)에 해당하며 이는 사용자가 직접 입력, 평가한 점수이다. 표1의 예에서 보면 사용자1은 음악4, 음악5를 이전에 들어보지 못하였고 평가하지 않았다. 추천 시스템의 목적은 음악4, 음악5에 대한 선호도 수치를 산출해 내는 것이다.

추천 시스템에는 추천하는 방식에 따라서 컨텐트 기반 기법, 협업 필터링 기반 기법, 하이브리드 기법의 3가지 타입이 존재한다.

컨텐츠 기반 기법은 추천의 대상이 되는 컨텐츠를 중심으로 사용자가 선호하는 품목과 비슷한 컨텐츠를 검색 그 비슷한 정도에 따라서 컨텐츠를 사용자에게 추천해주게 되는 형태이다. 협업 필터링 기반 기법은 해당 사용자와 비슷한 성향을 가진 사용자를 먼저 선택한 후 비슷한 사용자 그룹이 선호하는 품목을 해당 사용자에게 제공해 주게 된다. 하이브리드 기법의 경우 앞의 두 가지 형태의 장점을 결합한 형태가 된다.

컨텐츠 기반 기법과 협업 필터링 기반 기법의 가장 큰 차이점은 협업 필터링 기반 기법이 다른 사용자의 의견을 참조한다는 점이다. 일반적으로 추천 결과에 대한 사용자의 만족도가 협업 필터링 기법이 우수한 것으로 알려져 있다.

2.2 Web2.0과 태그

Web2.0의 흐름은 최근 산업계에서 크게 화두가 되고 있다. 웹 문서를 HTML파일로 작성하고 이를 HTTP 프로토콜에 의해서 주고 받는 등의 규약을 기반으로 하는 일련의 기법을 의미하는 Web1.0과는 달리 Web2.0은 특정 프로토콜이나 플랫폼을 의미하지 않는다. Web2.0은 시스템을 구축하거나 서비스를 구성하는 데 있어 다음과 같은 패러다임들을 고려하게 되는 하나의 흐름이라고 할 수 있다[3]. 이러한 Web2.0의 대표적인 사이트는 온라인 백과사전인 위키피디아[9], 태그를 이용한 사진 공개 사이트인 플리커[10] 등을 들 수 있다.

Web2.0의 흐름과 같이 나타난 것이 태그의 개념이다. 태그란 웹 상의 컨텐츠들에 대해서 사용자들이 추가하는 컨텐츠에 대한 키워드 등을 의미한다[6]. 일반적으로 태그 입력에 대한 제약 사항이 없으며 사용자는 컨텐츠에 자체에 대한 메타 정보, 감성 정보 등을 자유롭게 입력한다.

태그를 이용한 대표적인 사이트로 플리커[10], 딜리셔스[11] 등이 있으며 큰 성공을 거두었다. 이러한 성공에는 태깅 시스템이 가지는 우수성에 있다고 할 수 있다. 사용자가 컨텐츠에 대해서 태깅하는 과정을 인지과학적으로 살펴보면 인간이 특정 사물을 대할 때 발생하는 과정과 매우 잘 맞는다[5].

그러나, 태그를 분류하거나 태그에 대한 부가 정보를 입력하는 것은 사용자에게 큰 부담이 된다. 특히, 자신에게 익숙하지 않은 품목에 대해서 분류하는 것은 어려운 작업이다.

여기서 주목해야 할 점은 본 논문에서 태그를 활용하기 위해서는 태그가 사용자가 선호하는 성향을 포함하는지 그렇지 않은지의 여부를 결정해야 한다. 그러나, 이러한 정보는 태그 자체에는 표현되지 않으며 사용자에게 이러한 정보를 추가로 요구하는 것은 태그의 장점을 감소시킨다.

3. 태그 기반 협업 필터링 기법

3.1 개요

태그는 개별 단어일수도 있지만 여러 개의 단어로 이루어진 구의 형태도 존재한다. 태그를 활용하기 위해서는 태그가 가지는 자연 언어적인 특성을 고려해야만 한다. 기본적으로 대부분의 태깅 시스템에서 태그의 입력에 있어 형식에 제한이 없으므로 같은 의미를 가지는지를 판단하기 어려운 경우가 많다.

사용자	음악 제목	사용자의 태그
사용자1	SUGARCOAT	ALLTIME FAVORITES
사용자2	SUGARCOAT	FAVORITES SONG
사용자3	SUGARCOAT	FAVORITE

표 2 같은 의미를 가지지만 다른 형태의 태그들의 예

위의 표2의 예시는 여러 사용자가 같은 곡에 대해서 태깅한 예시이다. 사용자는 음악 "SUGARCOAT"에 대해서 같은 의미로 태깅을 한 것이라 판단할 수 있다. 그런데, 입력형식이 자유롭기 때문에 위와 같이 다른 형태로 입력을 하게 되는 경우가 매우 많다. 이러한 경우 다음과 같이 태그를 분리하고 어간추출의 단계를 거치면 같은 의미를 사용자가 표현했다는 것을 판단할 수 있다.

사용자	음악 제목	사용자의 태그	키워드	STEM
사용자1	SUGARCOAT	ALLTIME FAVORITES	ALLTIME	ALLTIME
			FAVORITES	FAVORITE
사용자2	SUGARCOAT	FAVORITES SONG	FAVORITES	FAVORITE
			SONG	SONG
사용자3	SUGARCOAT	FAVORITE	FAVORITE	FAVORITE

표 3 태그를 키워드 단위로 분리하고 어간 추출한 예

사용자1이 음악 'SUGARCOAT'에 대해 추가한 태그 'ALLTIME FAVORITES'는 2개의 단어로 이루어진 구의 형태로 되어 있으며 'ALLTIME', 'FAVORITES' 두 개의 개별 단어로 분리해 낼 수 있다. 또한, 복수형인 'FAVORITES'의 경우 복수형 어미인 ' -s'를 제거하고 'FAVORITE' 라는 어간을 추출할 수 있다. 마찬가지 방식으로 나머지 사용자2, 사용자3의 태그를 처리하면 사용자1, 사용자2, 사용자3 모두 'SUGARCOAT'에 대해서 'FAVORITE'라는 단어를 태깅한 것을 알 수 있으며 이러한 처리 전에 사용자 간의 연관성을 알 수 없었던 것에 비해서 추가적인 사용자간 연관관계를 파악할 수 있다.

또한, 위와 같은 과정에 의해 정제된 키워드들에 대해서 가중치 부여 과정이 필요하며 이 때 가중치를 계산할 때 평점정보와 태그 간의 연관성을 고려해야 한다. 각 사용자 별로 키워드에 대한 term frequency를 계산해서 키워드를 추출한 태그의 가중치에 곱해서 최종적인 키워드에 대한 가중치를 산출해 낸다.

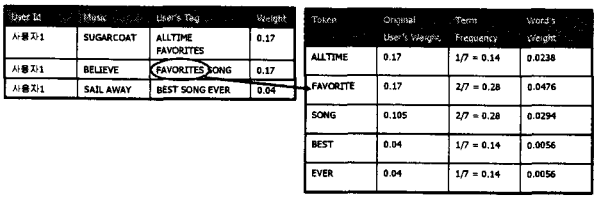

그림 1 태그로부터 추출된 키워드에 가중치 부여 과정

위의 그림 1은 태그로부터 추출된 키워드에 대해서 가중치를 부여하는 과정을 보여주고 있다. 여기서 주목할 점은 키워드에 대한 가중치 부여시 태그 단위로 계산했었던 태그와 사용자 평점과의 연관성으로부터 산출된 가중치를 사용했다는 것이다. 이는 태그를 키워드로 분리하더라도 원래 태그가 가지는 의미를 최대한 유지하기 위해서이다. 추출된 키워드에 대한 가중치를 계산할 때 사용된 Term Frequency는 일반적으로 정보 검색에서 사용되는 해당 키워드가 그 문서 내에서 나타나는 빈도를 의미한다. 본 논문에서는 문서의 개념이 아니라 한 사용자를 단위로 빈도수를 계산하게 되므로 Term Frequency를 다음과 같이 정의할 수 있다.

$$tf(i,j) = \frac{\text{키워드 } j\text{가 사용자 } i\text{의 태그집합내에서 출현한 빈도수}}{\text{사용자 } i\text{의 태그집합에서 추출한 모든 키워드의 총수}}$$

정보 검색 시스템에서 주로 term frequency와 같이 사용하는 inverted document frequency에 해당하는 inverted user frequency의 경우에는 일반적으로 협업 필터링 기법에 적합하지 않다는 연구결과가 있으므로 사용하지 않았다[7].

3.2 시스템 모델

사용자가 입력한 태그의 자연언어적인 모호성을 해결하기 위해 제안하는 추천 모델은 다음과 같은 요소들을 포함한다.

T	시스템 내의 태그의 집합 $T=\{t_1,t_2,\ldots,t_k\}$ k : the number of tags
U	시스템 내의 사용자의 집합
M	시스템 내의 음악의 집합
K	사용자의 태그 셋 T에서 추출한 키워드 집합 $K=\{k_1,k_2,\ldots,k_p\}$ p : the number of keywords
WF	태그의 가중치를 계산하기 위한 함수 집합 WF = {*WFt*, *WFk*} ① *WFt* : 태그가 사용자에 대해서 가지는 가중치를 계산하기 위한 함수 ② *WFk* : 태그에서 추출한 키워드가 사용자에 대해서 가지는 가중치를 계산하기 위한 함수
SFk	사용자와 사용자간 유사도를 계산하기 위한 함수

제안하는 추천 시스템에는 태그의 집합 T, 사용자의 집합 U, 음악의 집합 M 3가지의 집합에 사용자 태그에서 추출한 키워드 집합 K가 추가된다. 또한, 역시 태그의 가중치를 계산하기 위한 함수와 사용자간 유사도를 계산하기 위한 함수 두 종류의 함수들이 포함되어야 한다. 태그의 집합 T는 시스템 존재하는 모든 태그로 구성이 되며 그 개수가 k개이다. 시스템 에는 두 가지 종류의 가중치 함수가 존재하며 사용자에 대한 태그의 가중치를

계산하기 위한 함수와 음악에 대한 태그의 가중치를 계산하기 위한 함수로 나누어진다.

확장된 태그 기반 협업 필터링 모델의 User의 집합 U은 다음과 같이 정의할 수 있다.

$$U = \{u_1, u_2, .., u_i, .., u_m\}\ ,\ u_i : \textit{user } i$$

$$u_i = < (d_{i1}, d_{i2}, ..., d_{ij}, ..d_{il}), (rs_{i1}, rs_{i2}, ..., rs_{ij}, ...rs_{i\in}), (wk_{i1}, wk_{i2}, .., wk_{ij}, ..wk_{ip}) >$$

d_{ij} = *User profile of user i*
rs_{ij} = *User i's Rating Score of Music j*
wk_{ij} = *Keyword j's weight of user i = WFk(i, j)*

m : the number of users
n : the number of music
l : the number of user profile's attributes
p : the number of user's keywords

다음은 태그에서 추출된 키워드에 대한 가중치 부여 함수이다.

WFk(i,j) = **태그 k에서 추출한 키워드 j의 사용자 i에 대한 가중치를 산출**
= ***WFt(i,k) * tf(i,j)***
= ***correlation(i,j) * frequency(i,j) * tf(i,j)***

correlation(i,j) = **사용자J가 태깅한 음악의 평점 평균**
= $\frac{\sum \text{태그}i\text{가 포함된 사용자}j\text{의 음악의 선호도 점수}}{\text{태그}i\text{가 포함된 사용자}j\text{의 음악의 수}}$

frequency(i,j) = **태그i의 신뢰도 및 중요도 지수**
= **태그의 popularity = 태그i가 태깅된 음악비율**
= $\frac{\text{사용자}j\text{가 태그}i\text{를 태깅한 아이템의 수}}{\text{사용자}j\text{가 태깅한 음악의 총 수}}$

tf(i,j) = term freqyency of keyword j
= $\frac{\text{키워드 }j\text{가 사용자 }i\text{의 태그집합내에서 출현한 빈도수}}{\text{사용자 }i\text{의 태그집합에서 추출한 모든 키워드의 총수}}$

가중치 계산 시 키워드 각각과 태그 간 연관도를 산출하는 대신 키워드를 추출하기 전 태그의 가중치를 대신 사용하는 이유는 태그가 여러 개의 단어로 이루어 졌다고 하더라도 하나의 구로써 가지는 태그의 의미를 유지하기 위함이다. 또한, 각 태그에서 추출된 키워드들의 신뢰성을 측정하기 위해서 각 키워드의 빈도수 정보, Term Frequency를 사용하였다.

그리고 사용자 벡터 u_i간의 유사도 비교 함수의 정의는 다음과 같다.

$$SFt(i,j) = \alpha\left(\frac{\sum_{x=1}^{l}|d_{ix}|*|d_{jx}|}{\sqrt{\sum_{x=1}^{l}|d_{ix}|^2}*\sqrt{\sum_{x=1}^{l}|d_{jx}|^2}}\right) + \beta\left(\frac{\sum_{x=1}^{m}|rs_{ix}|*|rs_{jx}|}{\sqrt{\sum_{x=1}^{m}|rs_{ix}|^2}*\sqrt{\sum_{x=1}^{m}|rs_{jx}|^2}}\right) + \gamma\left(\frac{\sum_{x=1}^{k}|wu_{ix}|*|wu_{jx}|}{\sqrt{\sum_{x=1}^{k}|wu_{ix}|^2}*\sqrt{\sum_{x=1}^{k}|wu_{jx}|^2}}\right)$$
$$,\ \alpha + \beta + \gamma = 1$$

사용자에 포함된 3개의 벡터에 대해서 각각 코사인 유사도(Cosine Similarity)를 계산해서 가중 평균을 낸다. α , β, γ 파라미터는 유사도를 산출하는 데 있어서 사회통계학적 정보, 사용자 평점 정보, 가중치가 부여된 태그 벡터의 유사도를 얼마나 반영할 것인가의 파라미터이며 α , β, γ 는 대상이 되는 데이터의 품질에 따라 실험적으로 결정해야 한다.

4. 실험 방법

4.1 데이터 셋

본 논문에서 사용한 데이터 셋은 현재 운영되고 있는 음악 사이트인 last.fm [21] 에서 웹 크롤링을 통해 수집된 자료이다. 실제로 웹 크롤링한 데이터의 내역은 다음과 같다.

데이터 항목	건수
사용자	11140 명
사용자가 태깅한 전체 건수	159106 개
사용자별 평균 태그의 수	14.3 개
전체 사용자 태그 수	14302 개
전체 음악의 수	29354 곡
학습 셋에 포함된 음악 수	14849 건
테스트 셋에 포함된 음악 수	14505 건
전체 태그에서 추출한 키워드의 수	101042 개
중복 제거된 키워드의 수	9994 개

표 4 데이터 셋의 크기

4.2 추천의 정확도 측정 방법

추천의 정확도를 평가하기 위해서 사용자가 직접 입력한 선호도 수치와 태그 기반 협업 필터링 기법을 이용한 추천 엔진에 의해서 예측된 선호도 수치를 비교, 차이값을 측정하는 방식을 택하였다. 이를 위해서 먼저 기 구

축된 전체 사용자-음악 평점 매트릭스에서 추천 엔진 구축을 위한 학습용 데이터 셋과 테스트 대상이 될 셋으로 랜덤하게 분할하여 두 개의 데이터 셋을 생성하였다.

여기서 학습 셋이란 음악 추천 과정에서 사용자간 유사도를 측정하는데 활용되는 데이터를 의미하며 테스트 셋이란 음악 추천 엔진이 추천한 결과와 비교를 통해 추천의 정확도를 측정하기 위한 데이터 셋이다. 표 8에서와 같이 전체 29354 건 중에 학습 셋으로 14849 건, 테스트 셋으로 14505 건으로 분할하였다.

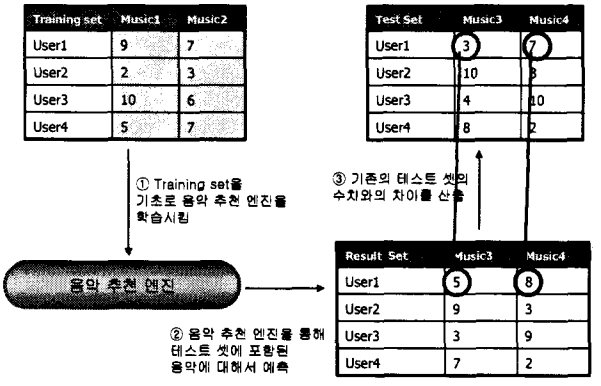

그림 2 추천의 정확도 측정 과정

위의 그림 2에서 추천의 정확도를 측정하는 과정을 보여준다. 위의 예는 음악 추천 엔진이 Music1, Music2, Music3, Music4 중에서 Music1, Music2에 관한 사용자들의 평점 벡터들을 학습 셋으로 각각 Music3, Music4에 대한 평점을 예측한다. 출력된 셋에 대해서 일치하는 음악에 대해서 수치를 계산한다. 위의 예에 대한 최종 에러 수치는 다음과 같다.

Result Set	음악3	음악4
User1	|5-3| = 2	|7-8| = 1
User2	|9-10| = 1	|3-3| = 0
User3	|4-3| = 1	|10-9| = 1
User4	|8-7| = 1	|2-2| = 0

표 5 테스트 셋에 대해서 측정된 에러의 예

이 예시의 평균 에러는 (2 + 1 + 1 + 1 + 1 + 0 + 1 + 0 = 7) / 8 = 0.875가 된다.

5. 실험결과 및 분석

정확도 측정을 위한 실험은 총 4가지의 실험을 진행하였다. 1) 첫 번째는 태그를 고려하지 않고 사용자의 음악에 대한 평점 벡터만을 활용한 협업 필터링 기법이며 2) 두 번째는 사용자의 음악에 대한 평점 벡터와 가중치를 부여하지 않은 태그 벡터를 같이 활용한 실험이다. 3) 세 번째는 태그를 파싱하지 않고 태그 자체에 가중치를 부여한 태그 기반 협업 필터링 기법에 관한 실험이며 4) 네 번째는 태그를 파싱해서 키워드를 추출 후 가중치를 부여한 확장된 태그 기반 협업 필터링 기법에 관한 실험이다. 2), 3), 4)번 실험에서 유사도 비교 함수에서 사용한 파라미터는 α = 0, β = 0.6, γ = 0.4 이다.

1) User's Ratings	2) User's Ratings+ User's Tag Vector (non-weight)	3) User's Ratings+ User's Weighted Tag Vector	**4) User's Ratings+ User's Keyword Vector**
11.74	10.32	9.96	5.72

표 6 파라미터 β = 0.6, γ = 0.4인 정확도 측정 실험결과

5) User's Rating	6) User's Rating+ User's Tag Vector (non-weight)	7) User's Rating+ User's Weighted Tag Vector	**8) User's Rating+ User's Keyword Vector**
11.74	10.97	10.88	8.42

표 7 파라미터 β = 0.8, γ = 0.2인 정확도 측정 실험결과

협업 필터링 기법에 있어 태그를 이용한 2), 3)번 실험이 1)번 실험에 비해 평균 에러수치가 감소된 것을 관찰할 수 있다. 이러한 결과를 통해 태그가 사용자의 성향을 잘 반영하고 있고 협업 필터링 과정에서 유사한 성향을 가지는 사용자를 찾아내는 데 효과가 있다는 것을 알 수 있다. 1)번 실험에 비해서 에러수치 감소의 폭은 각각 (11.74-10.32) / 11.74 = 12%, (11.74-9.96) / 9.96 = 15%이다. 또한, 태그에 사용자 평점과의 연관도에 의해 가중치를 부여한 3)번 실험의 평균 에러수치가 2)번 실험에 비해서 더 작다는 것을 알 수 있다. 이를 통해 태그에 가중치를 부여하지 않은 태그에 비해서 가중치가 부여된 태그가 협업 필터링 성능 향상에 도움이 된다는 것을 알 수 있다.

또한, 여기서 주목해야 할 점은 4)번 실험의 경우 1)번 실험보다 평균 에러 수치가 (11.74 - 5.72) / 11.74 = 50% 감소하였고 2)번,3)번 실험에 비교해서 비약적으로 감소하였음을 알 수 있다. 이는 태그를 파싱해서 키워드의 집합으로 표현하지 않은 경우 유사한 성향의 사용자를 검색하는 데 있어서 검색의 폭이 좁아진다는 것을 알 수 있다. 실험 5), 6), 7), 8)의 경우에도 추천의 정확도가 향상된 정도가 작지만 1), 2), 3), 4)번 실험의 결과와 동일하게 판단할 수 있다.

이러한 결과들을 종합해볼 때 태그의 모호성을 극복하는 것 또한 협업 필터링 기법에 도움이 된다는 것을 알

수 있다.

6. 결론 및 향후과제

6.1 결론

웹과 관련한 산업이 계속 증가하면서 추천 시스템의 중요성도 계속 증가하고 있다. Web2.0에 관련한 데이터들이 계속해서 웹 상에 축적되어 가고 있으나 기존 추천 기법들은 이를 충분히 사용하고 있지 못하였다. 이에 본 논문에서는 Web2.0 관련 데이터 중에 대표적인 것인 태그(tag)를 추천 기법의 정확도를 향상시키기 위한 기법을 제안하였다.

먼저 태그에 대한 가중치를 부여하는 기법을 제안하였다. 태그는 컨텐츠에 대한 메타 정보들을 포함하고 있으나 사용자의 선호하는 성향인지는 알 수 없으므로 사용자가 선호하는 성향을 직접적으로 나타내는 사용자 평점과의 연관도를 분석, 가중치로 환산하는 기법을 연구하였다. 또한, 태그의 모호성을 해소하기 위해 구의 형태인 태그의 경우 단어단위로 분리하고 어간 추출을 통해 태그에서 키워드들을 추출하였고 태그에 대한 가중치를 활용하여 이 키워드들에 가중치를 부여하는 기법을 제안하였다. 이러한 기법들이 태그를 사용하지 않은 기존 협업 필터링 기법과 가중치를 부여하지 않고 태그를 사용한 기법에 비해 추천의 정확도가 향상됨을 실험을 통해 보였다.

또한, 태그에서 추출한 키워드의 전체 수가 오히려 태그의 수보다 적다는 것을 관찰할 수 있었다. 이는 태그의 의미가 동일하더라도 표현 형식이 다양하게 입력된다는 사실을 알 수 있다. 이를 파싱 및 Stemming 과정을 거쳐 키워드를 추출하면 중복된 태그를 어느 정도 찾아낼 수 있고 이는 추천의 정확도뿐만 아니라 대상이 되는 추천 기법에 필요한 데이터의 양도 감소하기 때문에 추천의 성능 또한 태그 자체를 사용하는 것에 비해 향상되었다.

6.2. 향후 연구 과제

기본적인 협업 필터링 기법 뿐만 아니라 상업적으로 가장 성공한 기법인 아이템-기반 협업 필터링(Item-Based Colloaborative Filtering) 기법[4]이나 혼합형 기법 등에 대해서도 태그를 활용할 수 있도록 확장할 수 있다고 예상한다. 또한, 가중치 산출시 [8] 등에서 연구된 다양한 가중치 부여 기법과의 결합도 의미있는 연구가 될 것이다.

최근 유저 컨텍스트 정보를 고려하기 위해 온톨로지를 활용하는 추천 시스템에 대한 연구가 이루어지고 있는데 유사한 성향의 사용자 집합을 검색 시에 태그를 단순 매칭이 아닌 동의어 사전 또는 온톨로지를 이용해서 비교를 수행하게 하면 보다 더 정확하고 다양한 추천 결과를 얻을 수 있을 것이라 예상한다.

7. 참고문헌

[1] Gediminas Adomavicius, Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions". IEEE Transactions on Knowledge and Data Engineering, Vol 17, No6, June 2005, pages 734-749, 2005

[2] Ricardo Baesz-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval", Addion Wesley, pages 54-74, 1999

[3] Tim O'Reilly, "What Is Web 2.0", http://www.oreillynet.com , 2005

[4] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001

[5] Rashmi Sinha, "A Cognitive Analysis of Tagging", http://www.rashmisinha.com/ , 2005

[6] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F.Maxwell Harper, John Riedl, "tagging, communities, vocabulary, evolution", ACM Conference on Computer Supported Cooperative Work 2006, pages 181-190, 2006

[7] Kai Yu, Zhong Wen, Xiaowei Xu, Martin Ester, "Feature Weighting and Instance Selection for Collaborative Filtering", 2nd International Workshop on Management of Information on the Web - Web Data and Text Mining, pages 201-224, 2001

[8] Yi Ding, Xue Li, "Time Weight Collaborative Filtering", Proceedings of the 14th ACM Conference on Information and Knowledge Management 2005, pages 485-492, 2005

[9] http://www.flickr.com, 온라인 사진 공개 사이트

[10] http://del.icio.us, Social Bookmarking 사이트

[11] http://www.last.fm, 공개 음악 사이트 last.fm