확장된 태그 기반 협력적 필터링

신동민○ 이재원 이경종 이상구
서울대학교 대학원 전기전자 컴퓨터 공학부
{beatlifedm, lyonking, kjlee, sglee}@europa.snu.ac.kr

Expanded Tag-based Collaborative Filtering Approach

Dongmin Shin○ Jae-won Lee Kyeong-Jong Lee Sang-goo Lee
School of Computer Science & Engineering, Seoul National University

요 약

정보 기술의 발전으로 인해 이용할 수 있는 정보가 기하급수적으로 늘어남에 따라, 사용자는 원하는 정보를 얻는 데 어려움을 겪게 되고, 양질의 정보를 찾기 위해 많은 시간을 들이고 있다. 이에 사용자의 의도를 정확하고 명백하게 드러내는 태그 정보에 기반한 협력적 필터링 기법을 이용하여 사용자가 원하는 적절한 음악을 추천하는 시스템을 제안하며, 태그의 확장을 통한 협력적 필터링 기법의 성능 향상을 제안한다.

1. 서 론

정보 기술이 발전하고 인터넷 사용이 증가하면서 컨텐츠의 홍수라 불릴 정도로 정보 과잉 현상이 발생하고 있다. 이로 인해 사용자는 원하는 정보를 얻는데 어려움을 겪게 되고, 양질의 정보를 찾기 위해 많은 시간을 들이고 있다. 추천 시스템은 사용자의 정보 요구에 대하여 적합한 정보를 찾아주는 역할을 수행한다. 현재까지 머신 러닝[1], 데이터 마이닝[2] 등과 같은 다양한 기법이 추천 시스템에 적용되어 연구되었으나 함축적인 사용자 의도를 분석해 내야 한다는 점이 한계이자 도전 과제로 남아있다. 이러한 점에서 최근 인터넷의 가장 큰 화두인 web 2.0은 추천 시스템의 새로운 가능성을 열어주고 있다. web 2.0의 가장 중요한 키워드 중 하나인 태그(Tag)는 기존 소수의 관리자가 정보에 관련 키워드를 제공하던 것과는 달리 사용자가 직접 컨텐츠에 관련 키워드를 입력하여 또 하나의 새로운 정보를 제공하고 있다. 이전 추천 시스템이 사용자의 함축적인 의도를 분석해야 했다면 태그를 이용하여 보다 정확하고 명백한 사용자 의도를 사용할 수 있게 된 것이다.

이에 본 논문[1]에서는 태그를 활용하여 추천의 가장 대표적인 방법인 협력적 필터링(collaborative filtering) 기법의 성능을 향상시키도록 한다. 태그 정보를 이용하여 취향이 비슷한 사용자를 찾아내고 그를 통해 음악을 추천한다. 성능의 향상을 위해 태그가 사용자에게 가지는 중요도에 따라 가중치를 부여하고, 온톨로지(Ontology)를 이용한 태그의 확장으로 단순 키워드 중심인 태그 이용의 한계를 극복했다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 태그 기반 협력적 필터링과 태그를 분석하여 키워드로 분리하는 협력적 필터링에 대한 기존의 연구들에 대하여 살펴보며, 3장에서는 태그를 확장하여 활용하기 위한 기법을 제안하고 4장에서는 결론 및 향후 과제를 제시한다.

2. 관련 연구

2.1 태그 기반 협력적 필터링

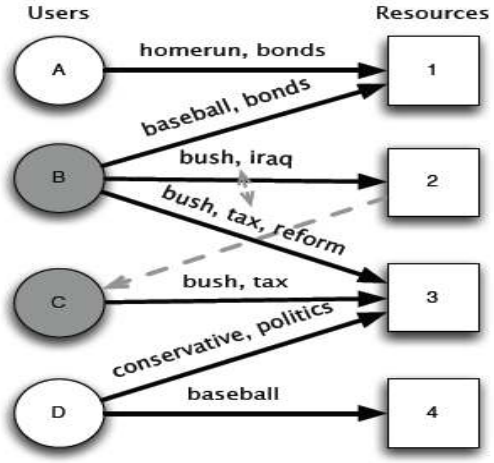
협력적 필터링으로 비슷한 성향의 사용자를 그룹화할 때, 사용자에 대한 표현에 태그의 집합을 포함시킨다. 협력적 필터링 기법에 있어 사용자에게 중요한 태그는 사용자가 선호하는 성향을 나타내고 있는 태그이므로, 태그의 신뢰도 및 중요도, 대중성 등을 계산해 가중치를 계산한다[3]. 같은 성향의 사용자라면 특정 음악에 같은 태그를 태깅할 것이라는 가정 하에 태그를 협력적 필터링에 이용하고 있다. 태그를 기반으로 사용자의 유사도 계산을 도식화한 것이 [그림1] 이다.

2.2 태그 분석 기반 협력적 필터링

단순히 태그 자체에 가중치를 두는 것을 넘어서 구문(phrase)으로 이루어진 태그를 단어 단위로 파싱하여, 어간 추출(stemming)의 단계를 거쳐 협력적 필터링에 사

1) 1) 본 연구는 지식경제부 및정보통신연구진흥원의대학 IT연구센터 육성·지원사업 (IITA-2008-C1090-0801-0031)의 연구결과로 수행되었음.

용한다[4]. [그림2]는 태그에서 추출된 키워드에 가중치를 하는 과정의 예시이다. 이 같은 태그 분석 단계를 거침으로써 처리 전에는 알 수 없었던 것에 비해서 추가적인 사용자간 연관관계를 파악할 수 있다.



[그림 1] 협력적 필터링을 이용한 사용자 유사도 모델

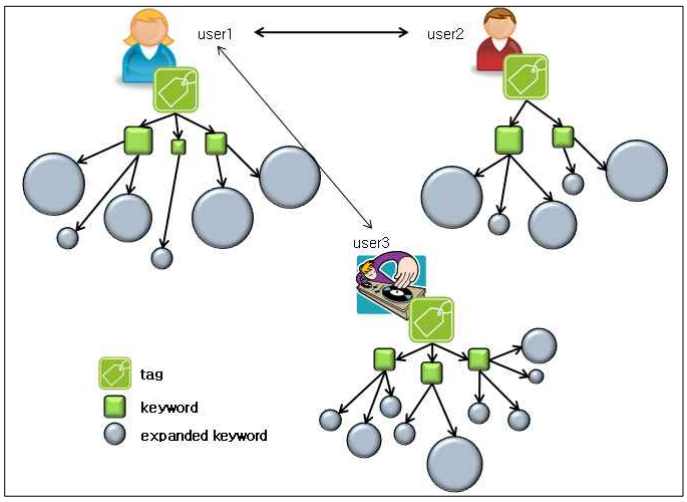
[그림 2] 태그로부터 추출된 키워드에 가중치 부여 과정

3. 태그 확장 기법

3.1 시스템 개요

구문 형태로 되어 있는 태그를 분석하여 단어 단위로 파싱한 후 어간 추출의 단계를 거쳐 키워드로 추출하는 기법은 기존의 단순 태그 이용 방식에 비해 향상된 결과를 도출한다. 이는 기본적으로 대부분의 태깅 시스템이 태그의 입력에 있어 형식 제한을 두지 않으므로, 서로 다른 태그가 같은 의미를 지니는지 여부를 판단하기 어렵기 때문이다[5]. 예를 들어, 'ALLTIMES FAVORITES'이라는 태그를 작성한 사용자의 경우, 본래의 태그가 'ALLTIME'과 'FAVORITE'으로 분리되므로 각각의 태그를 붙인 사용자들과 연관 관계를 계산하는 것이 가능하게 된다.

하지만 태그 분석 기반 협력적 필터링의 경우 분명한 한계를 지니는데, 이는 단어의 유사도를 계산하는 방식에서 오는 것이다. 태그 분석 기반 협력적 필터링 방법 역시 단어 자체의 1:1 매칭으로 사용자간의 연관도를 계산하기 때문에 이음동의어인 태그를 붙인 사용자 간의 유사도는 계산하지 못한다. 즉, 'ALLTIME FAVORITES'을 'ALLTIME', 'FAVORITE'과 연관 지을 수는 있어도 'GOLDEN MUSIC'과 연결할 수는 없다.



[그림 3] 시스템 구조

분석된 태그를 비슷한 의미를 지닌 단어들로 확장한다면 이러한 문제를 해결할 수 있다. 온톨로지를 이용하여 비슷한 의미를 지닌 연관 단어를 찾아낼 수 있는데, 이는 같은 문서에 출현하는 빈도가 높은 단어일수록 의미적으로 연결 되어있을 가능성이 높다는 가정을 전제로 한다[5]. 즉, 어간 추출된 키워드를 상호 연관 지수(Correlation Value)가 높은 다른 키워드들로 확장한 후, 이 키워드들을 이용하여 비슷한 성향을 지닌 사용자를 보다 정확하게 찾아낼 수 있다[그림3]. 주어진 태그를 분석한 후 키워드들로 확장하는 예는 [표 1]에서 확인할 수 있다.

[표 1] 태그의 확장 예시

User	Song	Tag	Keyword	Stem	Extended
User1	Why don't you get a job?	Indie Band	Indie	Indie	Funk
					Ska
					Rock
					Jam
					I know what you wanted
			Band	Band	...
User2	Wherever you will go	Love Actually	Love	Love	Ballad
					Queen
			Actually	-	
User3	...	...	...	...	...

온톨로지로부터 모든 키워드들을 추출해내고 각 키워드들이 출현하는 문서에서 다른 키워드들이 동시에 출현하는 빈도를 계산[6]하여 '키워드 동시 출현 빈도표'(keyword co-occurrence ratio matrix)[표2]를 만들

수 있다. 이와 같은 표를 이용하여 한 키워드에 대해 상위 n개의 키워드를 확장 대상으로 정한 후 협력적 필터링에 사용한다.

correlation(k,l) = 키워드 *k*와 키워드 *l*간의 상호연관지수

$$= \frac{n_{k,l}}{n_k + n_l - n_{k,l}}$$

n_k = the number of documents which has term k
$n_{k,l}$ = the number of documents which has term k and l

[표 2] 키워드 동시 출현 빈도표

	rock	indie	favorite	OST	love	party	...
alltime	0.75	0.32	0.86	0.42	0.76	0.53	
favorite	0.86	0.53	1	0.32	0.69	0.21	
guitar	0.91	0.84	0.44	0.02	0.05	0.02	
funk	0.74	0.59	0.77	0.21	0.07	0.02	
love	0.32	0.07	0.69	0.68	1	0.73	
rock	1	0.56	0.86	0.23	0.32	0.67	
top	0.87	0.68	0.92	0.33	0.45	0.56	
sad	0.62	0.31	0.87	0.29	0.79	0.04	
Christmas	0.25	0.01	0.95	0.23	0.90	0.65	
...							

3.2 시스템 모델

확장된 태그 기반 협력적 필터링 시스템은 다음과 같은 데이터 구조를 갖는다.

T	시스템 내의 태그 집합 $T = \{t_1, t_2, \cdots, t_k\}$ k : the number of tags
U	시스템 내의 사용자의 집합 $U = \{u_1, u_2, \cdots, u_i, \cdots, u_m\}$, ui : user i $u_i = <(d_{i1}, d_{i2}, \cdots, d_{ij}, \cdots, d_{il}),$ $(rs_{i1}, rs_{i2}, \cdots, rs_{ij}, \cdots, rs_{in}),$ $(wk_{i1}, wk_{i2}, \cdots, wk_{ij}, \cdots, wk_{ip}),$ $(wek_{i1}, wek_{i2}, \cdots, wek_{ij}, \cdots, wek_{ip})>$ 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 wek_{ij} = Extended keyword j's weight of user i = WF(i,t,k,l) m : the number of users n : the number of music l : the number of user profile's attributes p : the number of keywords q : the number of expanded keywords
	= the number of keywords in music ODP
M	시스템 내의 음악 집합 $M = \{m_1, m_2, \cdots, m_n\}$
K	태그 집합 T에서 추출한 키워드 집합 $K = \{k_1, k_2, \cdots, k_p\}$
EK	사용자 키워드와 온톨로지를 이용한 확장 키워드간의 상호연관도 집합 $EK = \{ek_{11}, ek_{12}, \cdots, ek_{ij}, \cdots, ek_{pr}\}$ ek_{ij} : co-occurence ratio of expanded keyword j with keyword i r : the number of expanded keywords

T는 사용자 집합 U로부터 뽑아낸 모든 태그 집합이다. 사용자 집합은 각각의 사용자 u_i의 모임이고, 각 사용자 u_i는 d, rs, wk, wek로 구성되며 각각은 사회통계학적 특성, 음악에 부여한 평점, 키워드 가중치, 확장된 키워드 가중치를 의미한다. EK는 온톨로지의 음악 카테고리 내에 있는 중복되지 않는 키워드 간의 상호연관도 집합이다.

WF	확장된 키워드의 사용자에 대한 가중치를 계산하기 위한 함수 wf_{il} = WF(i,t,k,l)
SFk	사용자와 사용자간 유사도를 계산하기 위한 함수

위에서 주어진 데이터를 이용해 사용자간 유사도를 계산하기 위해 WF, SFk와 같은 함수를 사용한다. WF는 태깅한 음악의 평점 평균과 태그의 신뢰도, 키워드의 출현 빈도를 이용해 확장된 키워드의 가중치를 계산한다.

WF(i,t,k,l) = 태그 *t* 에서 추출한 키워드 *k*를 확장한 키워드 *l* 의 사용자 *i* 에 대한 가중치를 산출

= 사용자 *i*가 태깅한 음악의 평점 평균

* 태그 *t*의 중요도 * term frequency of keyword *k*

* 키워드 *k*와 키워드 *l*간의 상호연관지수

= ***meanRating(i,t) * frequency(i,t) * tf(i,k) * correlation(k,l)***

$$meanRating(i,t) = \frac{\sum \text{태그}t\text{가 포함된 사용자}i\text{의 음악의 선호도 점수}}{\text{태그}t\text{가 포함된 사용자}i\text{의 음악의 수}}$$

$$frequency(i,t) = \frac{\text{사용자}i\text{가 태그}t\text{를 태깅한 아이템의 수}}{\text{사용자}i\text{가 태깅한 음악의 총 수}}$$

$$tf(i,k) = \frac{\text{키워드}k\text{가 사용자}i\text{의 태그 집합 내에서 출현한 빈도수}}{\text{사용자}i\text{의 태그 집합에서 추출한 모든 키워드의 총수}}$$

이 때 각각의 키워드가 아닌 태그의 신뢰도를 계산하는 이유는 태그가 여러 개의 단어로 이루어 졌다고 하더라도 하나의 구로써 가지는 태그의 의미를 유지하기 위함이다. 앞에서 주어진 데이터와 함수를 이용하여 사용자간 유사도를 측정하기 위해 코사인 유사도(Cosine similarity)를 사용하며 식은 다음과 같다. 여기서 α,β,γ 는 사회통계학적 유사도, 평점 유사도, 확장된 키워드 벡터의 유사도의 반영 비율을 나타낸다. 이번 연구에서는 α의 값을 0으로 둔 채, β와 γ의 값을 조정하여 좋은 결과값을 택하도록 한다.

$$SFt(i,j) = similarity\,function\,between\,user\,i\,\text{and}\,user\,j$$

$$= \alpha\left(\frac{\sum_{x=1}^{l}|d_{ix}|*|d_{jx}|}{\sqrt{\sum_{x=1}^{l}|d_{ix}|^2}*\sqrt{\sum_{x=1}^{l}|d_{jx}|^2}}\right)+\beta\left(\frac{\sum_{x=1}^{m}|rs_{ix}|*|rs_{jx}|}{\sqrt{\sum_{x=1}^{m}|rs_{ix}|^2}*\sqrt{\sum_{x=1}^{m}|rs_{jx}|^2}}\right)$$

$$+\gamma\left(\frac{\sum_{x=1}^{r}|wf_{ix}|*|wf_{jx}|}{\sqrt{\sum_{x=1}^{r}|wf_{ix}|^2}*\sqrt{\sum_{x=1}^{r}|wf_{jx}|^2}}\right)$$

$$,\alpha+\beta+\gamma=1$$

4. 결론 및 향후 과제

Web 2.0의 핵심 키워드인 사용자 태그 및 협력적 필터링를 결합함으로써, 개별 사용자의 요구 사항에 좀 더 부합되는 컨텐츠를 제공할 수 있는 가능성을 제시하였다. 기존의 연구는 단순 태그를 이용하여 유사한 사용자를 찾았다. 하지만 실제로 유사한 사용자가 다른 태그(이음동의어)를 사용한 경우, 그들간의 유사도를 계산하는데 한계가 있었다. 이에 본 논문은 온톨로지를 이용하여, 의미적으로 상호 연관 지수가 높은 키워드로 주어진 태그를 확장하였다. 태그 확장을 통해 비록 서로 다른 태그를 붙인 사용자이지만, 유사도를 계산할 수 있었다. 현재는 사용자들이 확장된 태그 집합으로 표현되지만, 다양한 도메인 온톨로지 컨셉 집합으로 표현하기 위한 연구를 진행하고자 한다. 사용자를 온톨로지 컨셉 집합으로 표현함으로, 데이터 희소성(Data Sparsity) 문제를 해결 할 수 있을 것으로 기대한다.

7. 참고 문헌

[1] B. Marlin, “Collaborative filtering: A machine learning perspective”. Master's thesis, University of Toront, 2004

[2]M.L. Shyu, C. Haruechaiyasak, S.C. Chen, N. Zhao, "Collaborative Filtering by Mining Association Rules from User Access Sequence", Proceedings of Web Information Retrieval and Integration, 2005

[3] Reyn NAKAMOTO, Shinsuke NAKAJIMA, Jun MIYAZAKI, Shunsuke UEMURA, “Tag-Based Contextual Collaborative Filtering”, Proceedings of Data Engineering Workshop 2007, 25-30, 2007

[4]이경종, 공기현, 이상구, 사용자 선호도와 태그 간 상관도 분석을 통한 태그 기반 협력적 필터링 기법에 관한 연구, 한국정보과학회 2007 가을 학술발표 논문집 제34권 제2호(C), 72-77, 2007

[5] T. Geodean, L. Koczy, “A Model of Intelligent Information Retrieval Using Fuzzy Tolerance Relations Based on Hierarchical Co-Occurrence of Words”, Soft Computing in Information Retrieval Techniques and Applications, 48-76, 2000

[6] Haruechaiyasak, C., Mei-Ling Shyu, Shu-Ching Chen, “Web document classification based on fuzzy association”, Computer Software and Applications Conference, 2002. COMPSAC 2002. Proceedings. 26th Annual International, 487-492, 2002