Tag를 이용한 CBF방식의 컨텐츠 선호도 예측 방법

*엄태광, 최성환, 이재황
삼성전자 DM연구소 Application S/W Lab.
e-mail : *tk.um@samsung.com, sh96.choi@samsung.com, jaehwang.lee@samsung.com*

A Study on Contents Preference Prediction Method using Tags based on Content-based Filtering

*Tae-Kwang Um, Sung-Hwan Choi, Jae-Hwang Lee
Application S/W Lab., Digital Media R&D Center,
Digital Media Business, SAMSUNG Electronics Co.,Ltd.

Abstract

A content recommendation according to users preferences comes up in the Internet application due to contents overwhelming. This paper newly proposes a method to predict contents preference using tags in conjunction with Content-Based Filtering. By implementing this method, this paper cleans up the contents sparsity problem in Content-Based Filtering, and shows the outstanding improvements.

I. 서론

최근 사용자의 선호도를 추출하여 개인화 서비스를 제공하는 추천 시스템이 증가하고 있다. 이러한 추천 시스템은 주로 인터넷 쇼핑이나 컨텐츠 추천과 같은 분야에서 사용자의 지속적인 소비를 유도하거나 광고의 효율성을 높이는 효과가 있다.

사용자의 선호도에 맞는 컨텐츠를 추천하기 위해서는 컨텐츠 유사성 판단을 통해서 기존에 사용자가 선호하는 컨텐츠와 유사한 추천 컨텐츠를 선별하게 된다. 이러한 필터링 기법을 Content-based Filtering (CBF)이라한다[1]. CBF방식은 컨텐츠 특성인 Feature Vector를 정하고 각 Feature에 가중치를 두어 사용자의 선호도를 계산하게 된다. 하지만 컨텐츠 제공자에 따라 사용하는 Feature의 형태가 다르고 통합적으로 관리되지 못하고 있기 때문에 <표 1>과 같이 컨텐츠를 비교하기 위한 Feature 데이터에 공백이 생기게 되어 컨텐츠에 대한 사용자의 선호도를 구할 수 없는 문제점이 있다(Sparsity 문제).

	Feature_1	Feature_2
Movie_1	O	-
Movie_2	-	O

<표 1> Sparsity문제 발생

본 논문에서는 이러한 문제점을 해결하고자 Feature를 Tag Space로 확장하여 사용자의 선호도를 예측하는 새로운 방안을 제안하였다. 또한 사용자에게 유효한 Tag를 추출해내기 위해서 Tag의 유효성을 검증하기 위한 방법을 도출하였다.

II. 본론

본 논문에서는 컨텐츠의 Feature Vector를 Tag Space로 확장하여 Sparsity문제를 해결하고자 한다. 여기서 Tag는 해당 컨텐츠의 관련 키워드를 말한다. 하지만 Tag중에서는 컨텐츠 선호도와 상관없이 많은 컨텐츠에 공통적으로 등장하는 Tag가 존재한다. 이러한 무의미한 Tag는 유효성 검증을 통해 제거함으로써 사용자의 선호도를 추출할 때 사용자에게 유효한 Tag만을 추출할 수 있다.

Tag를 이용한 컨텐츠 선호도 예측 방법의 성능을 측정하고자 영화 컨텐츠에 대한 사용자의 평가 점수가 나타나있는 Netflix movie dataset을 사용하여 식(1)과 같이 RMSE(root mean square error)를 구하였다. 이 때, 컨텐츠의 Feature를 Tag Space로 확장하는데 필요한 키워드는 IMDB 사이트에서 추출하였다[2,3].

$$RMSE = \sqrt{\frac{(R_1 - P_1)^2 + \ldots + (R_n - P_n)^2}{n}} \quad (1)$$

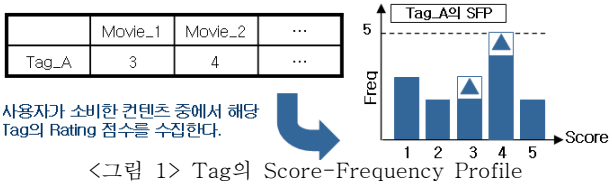
R : 실측값
P : 예측값
TS : Tag Score
n : 컨텐츠 갯수

관련된 기술의 성능 평가를 위해 Netfilx movie dataset을 Training set과 Test set으로 구분하여 실험을 진행하였다. Training set은 사용자의 선호도를 추출하기 위해 분석해야할 학습 데이터로 사용하였고, Test set은 단지 식(1)에서 추출된 사용자의 선호도를 이용해 컨텐츠의 평가 점수를 추정한 값인 예측값(P)과 비교하기 위한 실측값(R)으로만 사용하였다.

Ⅲ. 구현

1. Tag Space에서의 선호도 예측 방법

컨텐츠에 대한 사용자의 평가 점수를 해당 컨텐츠의 Tag 점수에 반영하면 <그림 1>와 같이 Tag의 Score-Frequency Profile(이하 SFP)로 나타낼 수 있다.



<그림 1> Tag의 Score-Frequency Profile

Tag Space에서의 예측값(P)은 기본적으로 식(2)와 같이 구할 수 있다. 여기서 Tag의 가중치(W)와 Tag를 대표하는 점수인 Tag Score(TS)는 SFP의 평균값을 사용하였으며 식(3)과 같이 표현할 수 있다.

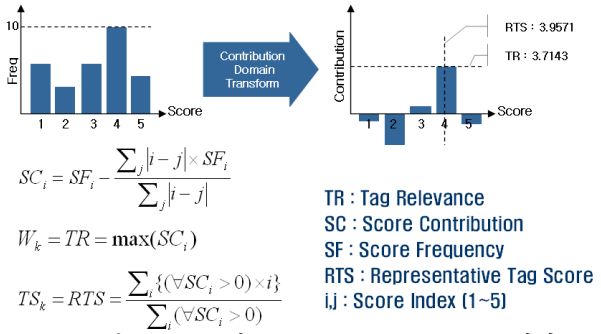
$$P = \frac{\sum_{k=0}^{m} (W_k \times TS_k)}{\sum_{k=0}^{m} W_k} \quad (2)$$

P : 예측값
W : 가중치
TS : Tag Score
m : Tag 갯수

$$W_k = \frac{\sum_{i=1}^{5} SF_i}{5} \qquad TS_k = \frac{\sum_{i=1}^{5} i \times SF_i}{\sum_{i=1}^{5} SF_i} \quad (3)$$

W : 가중치
TS : Tag Score
SF : Score Frequency

2. Tag의 Contribution Domain 변환 및 적용



$$SC_i = SF_i - \frac{\sum_j |i-j| \times SF_i}{\sum_j |i-j|}$$

$$W_k = TR = \max(SC_i)$$

$$TS_k = RTS = \frac{\sum_i \{(\forall SC_i > 0) \times i\}}{\sum_i (\forall SC_i > 0)}$$

TR : Tag Relevance
SC : Score Contribution
SF : Score Frequency
RTS : Representative Tag Score
i,j : Score Index (1~5)

<그림 2> Tag의 Contribution Domain 변환

Tag의 유효성을 판단하기 위해서 <그림 2>과 같이 Tag의 SFP를 각 점수의 Contribution Domain으로 변환하는 공식을 도출하였다. 즉, Tag의 Contribution Domain을 이용하면 1~5점 사이의 점수별 상관관계를 고려하기 때문에 Tag의 가중치(W)와 Tag Score(TS)를 분명히 알 수 있다. 이것은 점수별 편차가 거의 없는 무의미한 Tag를 제거할 수 있는 기준이 된다.

Ⅳ. 결론 및 향후 연구 방향

Tag의 Sparsity문제를 해결하기 위해서 <표 2>와 같이 컨텐츠의 Feature를 Tag Space로 확장하여 사용자의 선호도를 추출하는 방안을 제안하였다.

	Feature_1	Feature_2	Tag
Movie_1	○	-	○
Movie_2	-	○	○

<표 2> Sparsity문제 해결 방안

또한 실험결과는 <표 3>과 같이 나타났는데 Netflix movie dataset에서 영화 컨텐츠의 Feature Vector만으로는 Sparsity문제로 인해 CBF방식을 사용할 수 없었는데 Tag를 이용해 사용자의 선호도를 추출할 수 있는 것으로 나타났다.

컨텐츠 선호도 예측방법	RMSE 측정결과	
	Tag(①) 사용	확장된 Tag(①②③④⑤) 사용
Tag SFP 사용	0.9823	0.9760
Tag Contribution 사용	0.9998	0.9859

• Tag 종류 : ①Keyword ②Casting ③Director ④Genre ⑤Writer

<표 3> Tag Space를 적용한 실험결과

또한 사용되는 Tag Space를 확장하면 RMSE를 더 낮출 수 있는 것을 알 수 있다. 하지만 Tag의 Contribution Domain을 적용한 결과는 상대적으로 성능이 떨어지는 것으로 나타났는데 이것은 각 Tag의 의미를 Normalize하지 않아 Tag의 가중치(W)를 제대로 추출하지 못한 것이 그 원인으로 추정된다.

향후 이러한 문제점을 해결하기 위해서 추출된 Tag간의 Meaning Normalize 기술인 Ontology 기술을 활용한다면 사용자의 정확한 선호도를 추출하는데 더욱 좋은 결과를 기대할 수 있을 것이다.

참고문헌

[1] Matthew G., Gregory D., *Mixed Collaborative and Content-Based Filtering with User-Contributed Semantic Features*, 2006, In Proceedings of the Twenty-First National Conference on Artificial Intelligence, AAAI
[2] Netflix Prize, http://www.netflixprize.com
[3] IMDB, http://www.imdb.com