고객-제품 구매여부 데이터를 이용한 협동적 필터링에서의 유사성 척도의 사용
Use of Similarity Measures in Collaborative Filtering Based on Binary User-Item Matrix

이 종 석* 권 준 범** 전 치 혁**
{jongseok, samson, chjun}@postech.ac.kr
포항공과대학교 *정보통신대학원 **산업공학과
경상북도 포항시 남구 효자동 산31

Abstract

Collaborative filtering (CF) is originally based on the ratings of customers who vote on the items they used. When customers' votes are not available, user-item binary data set which represents choice and non-choice can also be used in this analysis. In this case the similarities between active user and the other users must be modified. Therefore we compare eight types of binary similarities by applying them in the modified CF Algorithm. Some experimental results will be reported.

1. Introduction

추천시스템 중 가장 대표적인 협동적 필터링을 이용한 추천 시스템은 여러 제품에 대한 고객들의 평가 데이터(voting history)를 기반으로 추천을 필요로 하는 고객과 타 고객과의 유사성을 바탕으로 관련 제품을 추천해 주는 기법이다[5]. 이처럼 제품의 추천을 위해서는 여러 고객들의 제품에 대한 평가치가 필수적인데 아무리 적극적인 고객이라도 업체가 다루는 제품의 1%정도도 평가하지 않을 정도로 고객 중에 제품의 사용 후 그 제품에 대한 평가를 하는 고객은 그리 많지 않다. 이는 협동적 필터링이 갖는 가장 큰 두 가지 문제, sparsity문제와 scalability 문제 중 sparsity문제에 해당하며 이를 해결하기 위해 기업은 고객들로부터 평가치를 얻기 위한 많은 노력을 하게 된다. 또한 고객 평가 데이터의 희소성을 해결하기 위한 관련 연구도 활발하기는 하나 대부분 결측치를 추정에 의해 대입하는 방안을 찾는 것으로서 실제 고객의 구매행태에 대한 사실성을 크게 반영한다고 할 수 없다[4].

이런 문제점을 극복하기 위한 한 방안으로써 협동적 필터링에 사용되는 데이터를 고객의 평가 데이터가 아닌 고객의 구매여부 데이터로 대체하는 시도가 있었다[1][2]. 일반적으로 구매여부를 나타내는 데이터는 평가 데이터에 비해 고객의 구매행태에 대한 더 많은 정보를 담고 있을 뿐만 아니라 고객으로부터 제품에 대한 평가를 얻기 위한 노력이 필요치 않다는 장점을 가진다. 즉, 평가 데이터에 의존하지 않음으로써 그 데이터를 사용했을 때 발생할 수 있는 문제점을 원천적으로 막을 수 있게 되는 것이다.

평가치로 구성된 데이터 형태에서 구매여부를 나타내는 1과 0으로 구성된 데이터로 그 분석 대상이 바뀌므로 협동적 필터링에서 고객간의 유사성을 나타내는 척도도 적절히 조정될 필요가 있다. 이에 본 논문에서는 각기 다른 8가지의 유사성을 제시하고 그를 이용한 추천하는 실험을 해 보았다. 그리고 각 유사성의 사용에 따른 정확도를 비교해 본다.

2. Binary Similarities

$n \times 1$크기의 1(구매)과 0(비구매)으로 구성된 서로 다른 두 고객의 구매행태를 나타내는 벡터 i와 j가 있다고 가정하자. 두 벡터의 각 원소에 대해 나타날 수 있는 4가지 경우에 대한 수를 [표 1]과 같이 세었을 때 벡터 i와 j의 8가지 유사성을 [표 2]에 정리 해 보았다[6]. [표 2]의 첫 번째 유사성인 covariance는 식(1)과 같이 근사 된 값이다.

$$\begin{aligned} Cov[i,j] &= E[ij] - E[i]E[j] \\ &= P\{i=1, j=1\} - P\{i=1\}P\{j=1\} \\ &\cong \frac{a}{n} - \frac{a+b}{n} \cdot \frac{a+c}{n} \end{aligned} \quad (1)$$

[표 1] Contingency table

		j	
		choice (1)	non-choice (0)
i	choice (1)	a = conjoint choice	b = mismatch
	non-choice (0)	c = mismatch	d = conjoint non-choice

[표 2] Similarity measures for binary variables

ID	Similarity	Formula	Properties
1	Covariance (Co)	a/n – ((a+b)/n×(a+c)/n)	-
2	Jaccard's coeff. Ⅰ (Ja)	a/(a+b+c)	Conjoint absence is ignored.
3	Dice's coeff. (Di)	2a/(2a+b+c)	Conjoint absence is ignored, conjoint presence is double weighted.
4	Russel&Rao's coeff. (RR)	a/(a+b+c+d)	Conjoint absence is not evaluated as similarity, but used in the denominator.
5	Sokal&Sneath's coeff. Ⅰ (SSⅠ)	a/(a+2(b+c))	Conjoint absence is ignored, mismatches are double weighted.
6	Simple matching coeff. (SM)	(a+d)/(a+b+c+d)	Absence and presence as well as matches and mismatches have equal weights.
7	Sokal&Sneath's coeff. Ⅱ (SSⅡ)	2(a+d)/(2(a+d)+b+c)	Matches (conjoint absence and presence) are double weighted.
8	Rogers&Tanimoto's coeff. (RT)	(a+d)/(a+d+2(b+c))	Mismatches are double weighted.

3. Experiments and Results

3.1 실험데이터

실험에서 사용된 데이터는 Collaborative Filtering 분야의 벤치마크 데이터로서 DEC Systems Research Center에서 제공하는 EachMovie Data Set이다[7]. 이 데이터는 72,916명의 사용자가 1,628개의 영화와 비디오에 대한 평가를 모은 것으로서 평가의 단계는 [0.0, 0.2, 0.4, 0.6, 0.8, 1.0] 6단계로 이루어져 있다. EachMovie Data Set은 데이터의 특성상 density가 약 5%정도로 희소성이 매우 높다.

본 실험에서는 구매여부가 나타나있는 데이터가 적합한데 구매여부의 데이터가 평가데이터보다는 더 조밀할 것이므로 이 데이터를 본 실험에 그대로 사용하기에는 적합하지 않다. 따라서 데이터내의 평가가 없는 셀은 비 구매로, 평가가 있는 셀은 구매로 간주하고 각각 구매를 나타내는 숫자 1과 비 구매를 나타내는 숫자 0으로 데이터를 변환하였다. 그리고 데이터 내의 density를 구매여부 데이터와 비슷하게 하기 위해 일정 수 이상의 사용자가 평가한 영화만을 그리고 일정 수 이상의 영화를 본 사용자만을 추출하여 데이터를 구성함으로써 density를 약 27~30%로 높이는 가공을 하였다. 실험에 사용된 데이터는 EachMovie Data set의 서로 다른 부분에서 추출한 모두 5개의 가공된 데이터를 사용하였으며 data description은 [표 3]과 같다.

3.2 수식

이분 행렬(Binary Matrix)을 이용한 협동적 필터링에서 특정 아이템 j에 대한 active user 즉, 추천의 대상이 되는 고객 a의 구매여부를 예측하는 예측 값 $P_{a,j}$는 식(2)를 통해 계산이 된다.

$$P_{a,j} = \kappa_a \sum_{i=1}^{n} w(a,i) \cdot v_{i,j} \qquad (2)$$

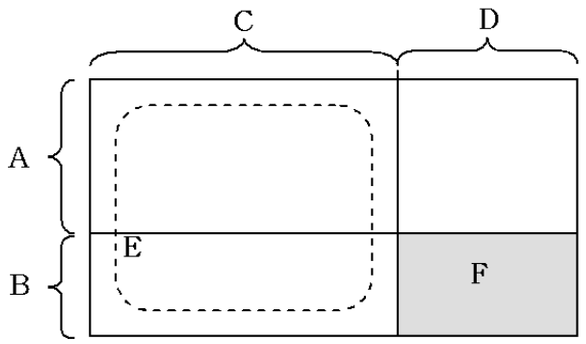
$$\kappa_a = 1 \Big/ \sum_{i=1}^{n} |w(a,i)| \qquad (3)$$

식(2)의 $w(a,i)$는 active user a와 다른 user i 간의 유사성을 의미하고 $v_{i,j}$는 user i의 j번째 아이템에 대한 구매여부를 나타내는 수치로서 0 또는 1의 값을 갖는다.

식(3)은 $P_{a,j}$를 0과 1사이의 값으로 예측하기 위한 normalizing factor로써 active user a와 그 외 n명의 i user 간의 유사성 절대값의 합의 역수로 표현된다. 이때 [표 2]에서 소개된 8가지 유사성 척도 중 1번 척도인 covariance만이 음의 값을 가질 수 있는 반면 나머지 7개의 척도들은 0과 1사이의 값을 가지므로 covariance를 사용하는 경우 외에는 절대값의 사용이 필요 없게 된다.

[표 3] Data description for experiments

	Non-zeros	Users	Items	Density (%)	A	B	C	D
Data 1	5444	105	168	30.86	70	35	100	68
Data 2	6902	160	150	28.76	100	60	100	50
Data 3	6641	161	147	28.06	100	61	100	47
Data 4	7145	155	168	27.44	100	55	100	68
Data 5	7055	159	146	30.39	100	59	100	46



A : Training users
B : Active users
C : Items for calculating the similarities
D : Items under recommendation
E : Data for calculating the similarities between A and B
F : Comparison data set

[그림 1] Experimental data split

3.3 실험방법

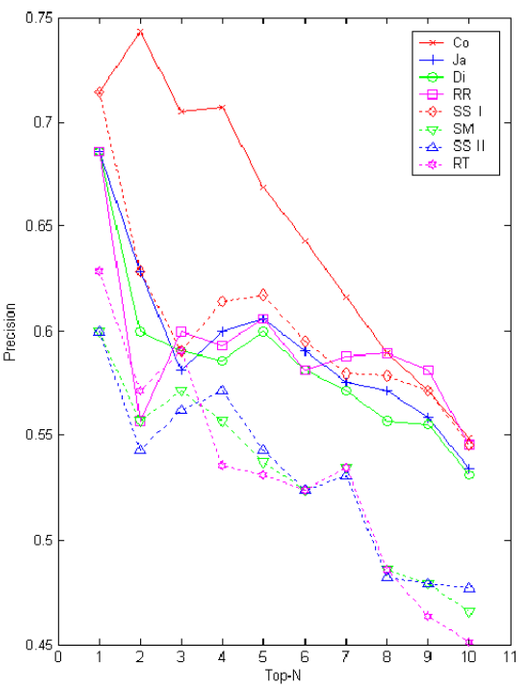
협동적 필터링의 공정한 평가를 위해 [표 3]에 기술된 5개의 데이터는 모두 [그림 1]과 같이 분할되게 된다. 점선으로 표시된 부분의 데이터 E는 오직 user A와 B사이의 유사성을 계산하는 데에만 사용하고 C부분에 해당하는 아이템에 대해서는 예측하지 않음으로써 유사성을 계산하기 위한 아이템과 추천 대상이 되는 아이템(D)을 구분하였다. 즉, E로부터 계산된 유사성과, A에 해당하는 사용자들의 D에 해당하는 아이템에 대한 구매행태로부터 B 사용자들의 아이템 D에 대한 구매여부를 예측한다. 결국 F부분에 해당하는 데이터에 대해서, 예측된 값으로부터 아이템을 추천했을 때 실제 구매여부와 일치하는지 여부를 정확도의 기준으로 본다. 이때 정확도의 척도로는 정보검색(Information Retrieval)에서 많이 사용되는 Precision을 사용하며 식(4)와 같다.

$$\text{Precision} = \frac{\text{hitting number}}{\text{Top - N}} \qquad (4)$$

식(4)에서 hitting number는 어느 한 사용자에 대해 예측된 값 중에서 가장 높은 N개의 아이템을 추천했을 때 실제 그 사용자가 추천된 아이템을 구매했는지 여부를 나타내는 값이다. 예를 들어 '갑'이라는 사용자에게 5개의 아이템을 추천했을 때 '갑'이 그 5개의 아이템 중 실제 4개를 구매한 적이 있다면 precision은 80%가 되는 것이다.

3.4 실험결과

5개의 데이터에 대해서 Top-N의 경우는 1~10까지 두었다. 즉 1개부터 10개까지 추천했을 때 각 데이터에 대해서 precision을 살펴보았으며 데이터 1에 대한 결과는 [그림 2]와 [표 4]와 같다.



[그림 2] Result for the data 1

8가지 유사성 모두 비슷한 precision을 보이나 데이터 1에 대해서는 covariance 유사성을 사용했을

때가 다른 유사성을 사용했을 때에 비해 조금 높은 precision을 보임을 확인할 수 있다. 특히 추천 아이템의 개수가 2~7인 경우는 다른 여느 유사성에 비해 높은 정확도를 보인다.

[표 4] Precision for the data 1

ID / N	Co	Ja	Di	RR	SS I	SM	SS II	RT
1	**0.71**	0.69	0.69	0.69	**0.71**	0.60	0.60	0.63
2	**0.74**	0.63	0.60	0.56	0.63	0.56	0.54	0.57
3	**0.70**	0.58	0.59	0.60	0.59	0.57	0.56	0.59
4	**0.71**	0.60	0.59	0.59	0.61	0.56	0.57	0.54
5	**0.67**	0.61	0.60	0.61	0.62	0.54	0.54	0.53
6	**0.64**	0.59	0.58	0.58	0.60	0.52	0.52	0.52
7	**0.62**	0.58	0.57	0.59	0.58	0.53	0.53	0.53
8	**0.59**	0.57	0.56	**0.59**	0.58	0.49	0.48	0.49
9	0.57	0.56	0.56	**0.58**	0.57	0.48	0.48	0.46
10	**0.55**	0.53	0.53	**0.55**	**0.55**	0.47	0.48	0.45
평균	**0.65**	0.59	0.59	0.59	0.60	0.53	0.53	0.53

5개 데이터에 대해서 모두 [그림 2] 그리고 [표 4]와 같은 결과를 얻었으며 5개의 결과를 모두 포괄하는 결과는 아래 [표 5]와 같다.

[표 5] Average precision for the all data

	Co	Ja	Di	RR	SS I	SM	SS II	RT
Data1	**0.65**	0.59	0.59	0.59	0.60	0.53	0.53	0.53
Data2	**0.77**	0.75	0.75	0.74	0.75	0.75	0.75	0.75
Data3	**0.77**	0.76	0.76	0.75	0.76	0.74	0.75	0.74
Data4	**0.76**	0.75	0.75	0.75	0.75	0.73	0.73	0.73
Data5	**0.75**	**0.75**	0.74	0.74	**0.75**	0.74	0.74	0.74
평균	**0.74**	0.72	0.72	0.72	0.72	0.70	0.70	0.70

8가지 유사성들이 비슷한 정확도를 보이기는 하지만 5개의 데이터 모두 covariance 유사성이 가장 높은 정확도를 보이고 있다.

4. Conclusion and Future Works

고객의 평가치를 기반으로 하는 협동적 필터링에서는 통상 Pearson correlation coefficient나 vector similarity를 사용하는데 데이터의 형태가 제품의 구매여부를 나타내는 이분행렬이라면 유사성도 데이터의 특성에 맞게 조절되어야 한다고 생각하고 본 실험을 하게 되었다. 그 유사성으로서 8가지를 제시하였고 각 유사성을 사용하여 제품의 구매여부를 예측하는 실험을 해 보았는데 8가지 모두 협동적 필터링에서는 만족할 만한 정확도를 보였다.

특히 본 논문의 실험에서는 covariance 유사성이 다른 여느 유사성에 비해 높은 정확도를 보였다. 이에 본 논문에서는 이와 같이 구매여부 데이터를 이용한 협동적 필터링에서의 covariance 유사성 사용을 추천한다. 또한 Jaccard's coeff. I 이나 Dice's coeff., Sokal&Sneath's coeff.같은 유사성을 사용할 때 두 고객의 유사성을 살펴본 결과 conjoint non-choice(d)의 경우 밖에 없다면 두 고객 사이의 유사성을 계산할 수 없는 위험한 경우에 빠지게 된다. 하지만 covariance 유사성은 어떤 경우에도 그 값이 계산가능 하므로 이것이 또 하나의 추천이유가 된다.

본 논문에서는 고객의 구매여부 데이터를 이용한 협동적 필터링에서 적절한 유사성 사용에 대한 제안을 하고자 하였으며 좀 더 정확성을 높이기 위한 방안, 예를 들어 Association Rule Mining과 같은 방법을 동시에 사용한 노력이 추후 계속되어야 할 일이라고 생각한다. 또한 본 논문과 같은 고객간의 유사성에 기반한 User-Based Approach외에 동일 데이터를 이용한 logistic regression과 같은 모델을 통한 Model-Based Approach에 대한 연구도 가능할 것이라고 생각한다.

5. References

[1] 이종석, 권준범, 전치혁, "고객-제품 구매여부 데이터를 이용한 제품 추천 방안", *한국경영과학회 학술대회 논문집*, 191-194, 2003. 11.,

[2] Andreas Mild, Thomas Reutterer, "An improved collaborative filtering approach for predicting cross-category purchase based on binary market basket data", *Journal of Retailing and Consumer Services,* Vol. 10, 123-133, 2003

[3] John S. Breese, David Heckerman, Carl Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", *Microsoft Research Technical Report*, MSR-TR-98-12, 1998

[4] Michael Pryor, "The effect of singular value decomposition on collaborative filtering", *Dartmouth College CS Technical Report*, 1998

[5] David Goldberg, David Nichols, Brian Oki, and Douglas Terry, "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, 35(12):61-70, 1992

[6] SPSS Inc., 2001: Cluster. (SPSS Statistical Algorithms, http://www.spss.com/tech/stat/Algorithms.htm).

[7] http://www.research.digital.com/SRC/eachmovie/