A Novel Bayesian Similarity Measure for Recommender Systems

Guibing Guo, Jie Zhang, Neil Yorke-Smith

1. Motivation

Ratings are essentially important for collaborative filtering to identify similar users based on which recommendations are generated. However, traditional similarity measures (cosine similarity, Pearson correlation coefficient) suffer from issues:

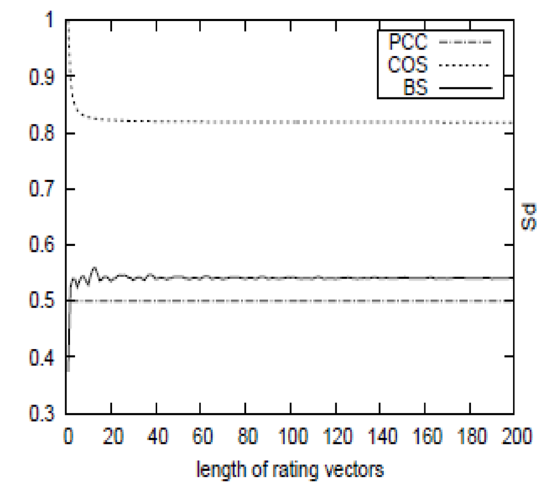
- Flat-value problem: COS=1, PCC non-computable
- Opposite-value problem: PCC=-1
- Single-value problem: COS=1, PCC non-computable
- Cross-value problem: PCC=-1 (crossing), 1 (otherwise)

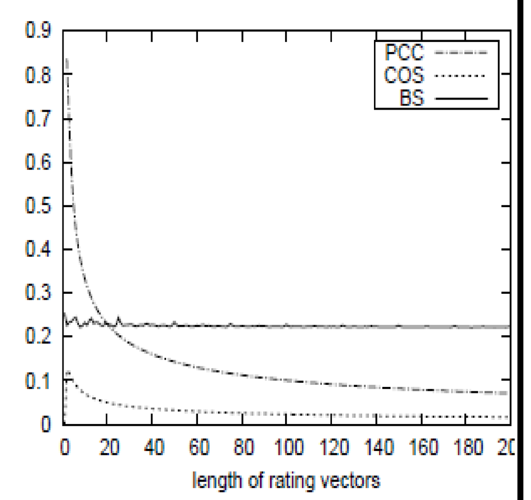
COS and PCC only consider the direction of rating vectors. Hence, we design a novel Bayesian approach by taking into account both the direction and length of rating vectors.

3. Experiments

(1) Examples of comparing with COS and PCC (left figure)
(2) The nature of similarity measures w.r.t. vector length

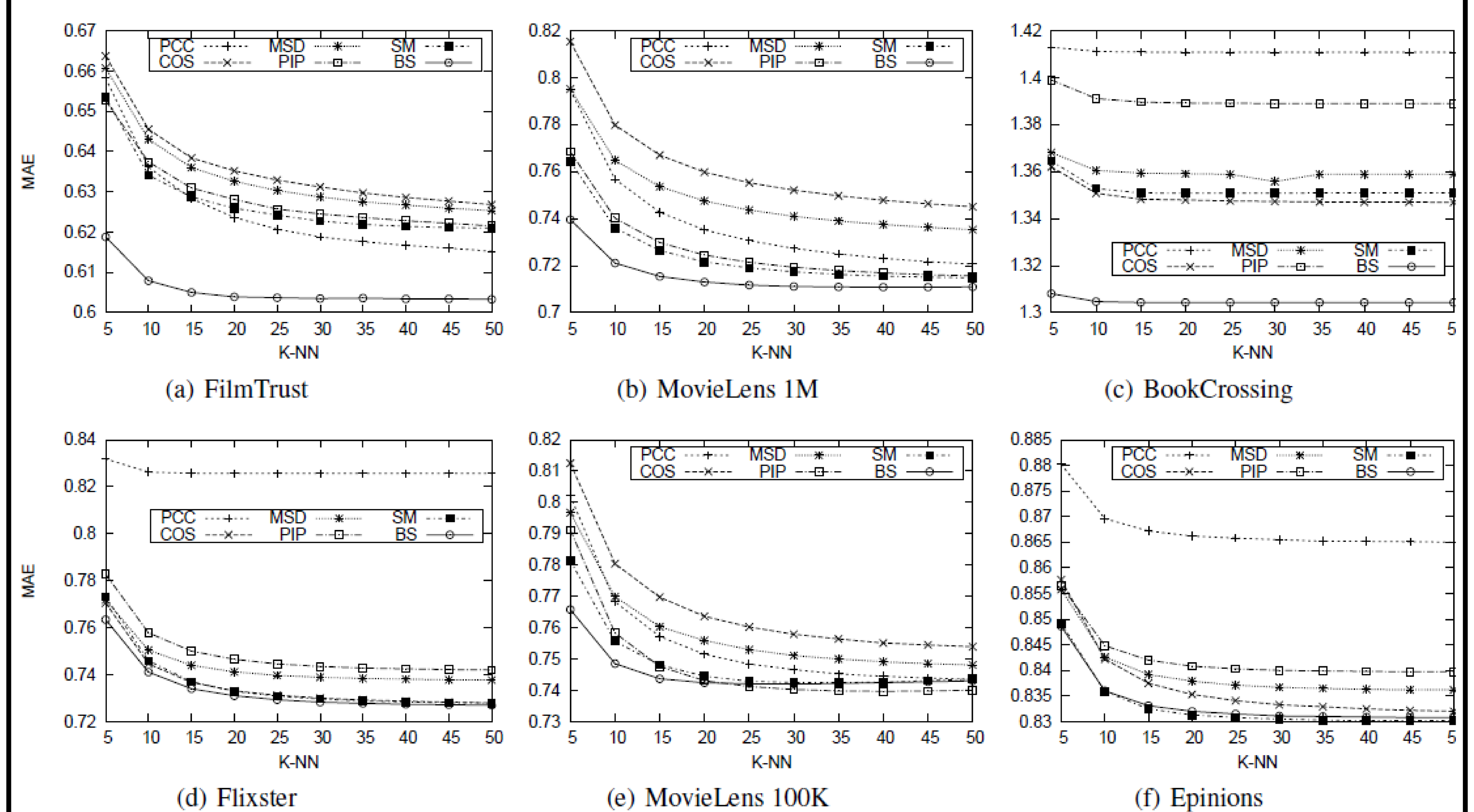
Problem	Examples			PCC	COS	BS	BS-1
	ID	Vector u	Vector v				
Flat-value	a_1	[1, 1, 1]	[1, 1, 1]	NaN	1.0	0.952	0.96
	a_2	[1, 1, 1]	[2, 2, 2]	NaN	1.0	0.677	0.71
	a_3	[1, 1, 1]	[5, 5, 5]	NaN	1.0	0.0	0.0
Opp.-value	a_4	[1, 5, 1]	[5, 1, 5]	-1.0	0.404	0.0	0.0
	a_5	[2, 4, 4]	[4, 2, 2]	-1.0	0.816	0.446	0.46
	a_6	[2, 4, 4, 1]	[4, 2, 2, 5]	-1.0	0.681	0.334	0.335
Single-value	a_7	[1]	[1]	NaN	1.0	0.76	0.96
	a_8	[1]	[2]	NaN	1.0	0.39	0.71
	a_9	[1]	[5]	NaN	1.0	0.0	0.0
Cross-value	a_{10}	[1, 5]	[5, 1]	-1.0	0.385	0.0	0.0
	a_{11}	[1, 3]	[4, 2]	-1.0	0.707	0.332	0.383
	a_{12}	[5, 1]	[5, 4]	1.0	0.888	0.530	0.5616
	a_{13}	[4, 3]	[3, 1]	1.0	0.949	0.485	0.5623





Conclusion: Bayesian similarity can solve the four issues of COS and PCC, and compute more reliable and distinguishable similarity measurements.

(3) Predictive accuracy on six real-world data sets:



(a) FilmTrust (b) MovieLens 1M (c) BookCrossing
(d) Flixster (e) MovieLens 100K (f) Epinions

Conclusion: Bayesian similarity achieves better predictive accuracy than others across data sets.

4. Conclusion and Acknowledgement

By incorporating both direction and length of rating vectors, a better and novel Bayesian similarity measure is developed.

This work is supported by the MOE AcRF Tier 2 Grant, M4020110.020, and the Institute for Media Innovation, NTU.

2. Bayesian Similarity Measure

Dirichlet distribution represents an unknown event by a prior distribution on the basis of initial beliefs. It suits similarity measure since similarity is updated when new ratings arrive.

- $(r_{u,k}, r_{v,k})$ is a pair of ratings given by users u, v on item k.
- $L = \{l_1, l_2, \dots, l_n\},\ l_j < l_{j+1}$ is a set of rating scales.
- $d = |r_{u,k} - r_{v,k}|$ is the rating distance we focus on.
- $\boldsymbol{D} = \{d_1, d_2, \dots, d_n\}, d_i < d_{i+1}$, D is a random distance variable whose probability distribution is $x = (x_1, \dots, x_n)$

The probability density of the Dirichlet distribution is:

$$p(x|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1} \Gamma(\alpha_i)} \prod_{i=1} x_i^{\alpha_i - 1}$$

where $\alpha_0 = \sum_{i=1}^{n} \alpha_i$ and $\alpha_i > 0$ represent the pseudo rating pairs observed in the prior. They are set by:

$$\alpha_i = \begin{cases} \sum_{j=1}^{n} n^2 p_j^2, & if\ i = 1 \\ 2\sum_{j=1}^{n-i+1} n^2 p_j p_{j+i-1}, & if\ 1 < i \le n \end{cases}$$

A rating pair can be represented by a vector $\gamma = (\gamma_1, \dots, \gamma_n)$, where only $\gamma_i = 1$ due to $d_i = |r_{u,k} - r_{v,k}|$ and others remain 0. The evidence weight is defined by:

$$e_i = \begin{cases} 1 & if\ c\sigma_k = 0 \\ 1 - \frac{d_i}{c\sigma_k} & if\ 0 \le d_i < 2c\sigma_k \\ -1 & otherwise \end{cases}$$

where σ is the standard deviation of all ratings, and c is a constant and set to be l_1/σ or 0 if rating info unknown.

Hence the updated posterior probability is given by:

$$E(x_i|\alpha_i + \gamma_i^0) = \frac{\alpha_i + \gamma_i^0}{\alpha_0 + \gamma^0}$$

where $\gamma_i^0 = \sum_{j=1}^{N} \gamma_i^j e_i^j$ and $\gamma^0 = \sum_{i=0}^{N} \gamma_i^0$. The user distance is defined as the weighted average of rating distances:

$$d_{u,v} = \frac{\sum_{i=1}^{n} w_i d_i}{\sum_{i=1}^{n} |w_i|}$$

where $w_i > 0$ is the importance of the rating distance d_i, defined as the difference between the prior and posterior probability:

$$w_i = E(x_i|\alpha_i + \gamma_i^0) - E(x_i|\alpha_i)$$

The 'raw' similarity is obtained by normalizing user distance:

$$s'_{u,v} = 1 - \frac{d_{u,v}}{d_n}$$

Another consideration is correlation due to chance:

$$s''_{u,v} = \prod_{i=1}^{n} \left(\frac{\alpha_i}{\alpha_0}\right)^{\gamma_i^0}$$

Hence, the user similarity is defined by:

$$s_{u,v} = \max(s'_{u,v} - s''_{u,v} - \delta, 0)$$

where $\delta = 0.04$ is a constant user bias.