Similarity Aggregation for Collaborative Filtering

Analysis of Images, Social Networks and Texts (AIST 2015)
April, 9-11th, Yekaterinburg

Authors

Sheikh Muhammad Sarwar

Institute of Information and Technology
University of Dhaka, Bangladesh

Mahamudul Hasan

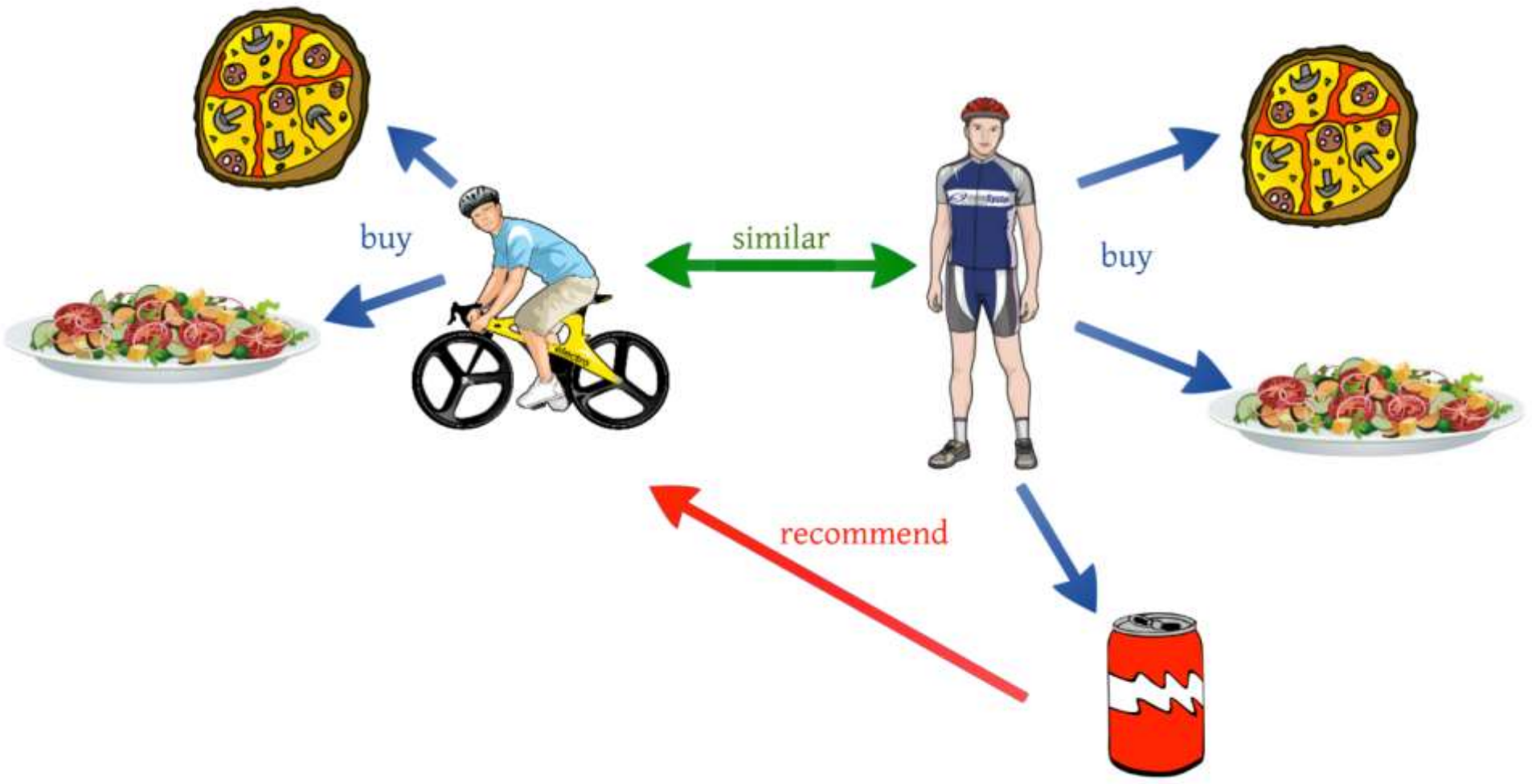
Masum Billal

Department of Computer Science and Engineering
University of Dhaka, Bangladesh

Dmitry Ignatov

National Research University Higher School of Economics,
Moscow, Russia

User-based Collaborative Filtering



Prime task is to determine similarity between users!!

Existing Metrics for Finding Similarity

user1 and user3 are very similar. Cosine gives 0.975 similarity score but PCC gives 0.0

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Pearson Correlation Coefficient (PCC)

	user2	user3	user4	user5
user1	0.707	0.0	0.707	0.707
user2		1.0	1.0	1.0
user3			1.0	1.0
user4				1.0

user2 and user4 are dissimilar. PCC gives 1 and cosine gives 0.997. Both are doing wrong!!

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Cosine similarity

	user2	user3	user4	user5
user1	0.612	0.975	0.606	0.605
user2		0.703	0.997	0.997
user3			0.696	0.695
user4				1.0

Existing Metrics for Finding Similarity (contd.)

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Mean Squared Deviation (MSD)

	user2	user3	user4	user5
user1	0.98	0.96	0.85	0.98
user2		0.98	0.74	0.96
user3			0.84	0.98
user4				0.9

Similarity between user1 and user2 is 0.98 and similarity between user1 and user3 is 0.96. But in reality user1 and user3 are much more similar than user2 and user3; they have 3 same ratings. However, JMSD understands this difference.

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Jaccard Mean Squared Deviation (JMSD)

	user2	user3	user4	user5
user1	0.49	0.96	0.42	0.49
user2		0.49	0.74	0.96
user3			0.42	0.49
user4				0.9

Our Hypotheses

- One single metric can't be trusted
- A user pair with higher similarity values from most of the metrics is super similar
- A User pair with lower similarity values from most of the metrics is super dissimilar
- Other user pairs are average similar
- Specifically we use votes or support of different metrics for analyzing a single user pair

Aggregated Similarity Metrics

- Pearson Correlation Coefficient (PCC)
- Constraint Pearson Correlation Coefficient (CPCC)
- Sigmoidal Pearson Correlation Coefficient (SPCC)
- Jaccard Similarity
- Mean Squared Deviation (MSD)
- Jaccard Mean Squared Deviation (JMSD)
- Cosine Based Similarity (CS)
- Adjusted Cosine Similarity (ACS)

Proposed Approach

- At first we compute support matrix. The support matrix may look like below:

	user2	user3	user4	user5
user1	0	5	8	4
user2		6	2	1
user3			0	0
user4				2

It can be seen from the table that support count between user1 and user2 is 0 and between user1 and user4 it is 8. Support count indicates how many metrics provide reasonably good similarity values for a specific user pair. As we are working with 8 metrics the highest value can be 8 and the lowest value is 0.

Process of Counting Support

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Jaccard Mean Squared Deviation (JMSD)

	user2	user3	user4	user5
user1	0.49	0.96	0.42	0.49
user2		0.49	0.74	0.96
user3			0.42	0.49
user4				0.9

Median = 0.49; everything above median =1 and below = 0

	user2	user3	user4	user5
user1	0	1	0	0
user2		0	1	1
user3			0	0
user4				1

Add with Support Matrix

All initial entries in support matrix is zero

	user2	user3	user4	user5
user1	0+0	0+1	0+0	0+0
user2		0+0	0+1	0+1
user3			0+0	0+0
user4				0+1

Process of Counting Support

	i1	i2	i3	i4
user1	4	3	5	4
user2	5	3	-	-
user3	4	3	3	4
user4	2	1	-	-
user5	4	2	-	-

Mean Squared Deviation (MSD)

	user2	user3	user4	user5
user1	0.98	0.96	0.85	0.98
user2		0.98	0.74	0.96
user3			0.84	0.98
user4				0.9

Median = 0.96; everything above median =1 and below = 0

	user2	user3	user4	user5
user1	1	0	0	1
user2		1	0	0
user3			0	1
user4				0

Add with Support Matrix

All initial entries in support matrix is zero

	user2	user3	user4	user5
user1	0+0+1	0+1+0	0+0+0	0+0+1
user2		0+0+1	0+1+0	0+1+0
user3			0+0+0	0+0+1
user4				0+1+0

Support Matrix After Applying Two Similarity Metrics

	user2	user3	user4	user5
user1	1	1	0	1
user2		1	1	1
user3			0	1
user4				1

If we look at the current support matrix, no user pair could get the support from the two metrics. However, we will keep this process until we have applied all the 8 metrics.

Super similar, Super dissimilar and Average Similar user pairs

Support Count in Support Matrix	User Type	Similarity Value
$S_{xy} \geq 5$	Super similar	1
$S_{xy} \leq 2$	Super dissimilar	-0.3
$S_{xy} = 3$ or $S_{xy} = 4$	Average similar	JMSD

We used JMSD for average similar users because, according to our implementation on Movielens data it performs the best among the 8 metrics we have considered. The performance is measured using MAE (Mean Absolute Error).

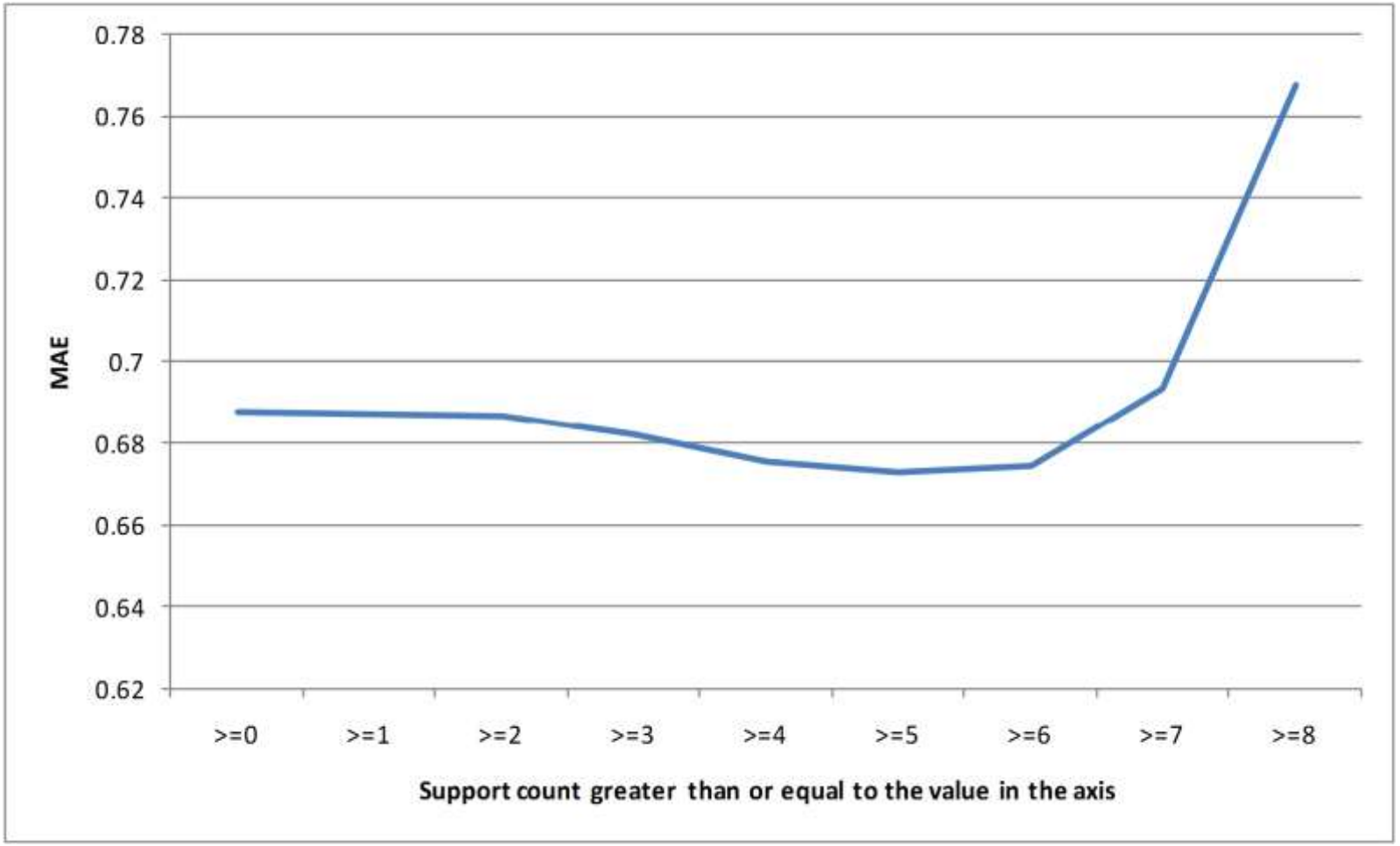
MAE values for 8 different existing metrics implemented by us

PCC	SPCC	CPCC	ACOS	COS	JMSD	MSD	JACCARD
0.688	0.687	0.685	0.687	0.687	0.680	0.688	0.682

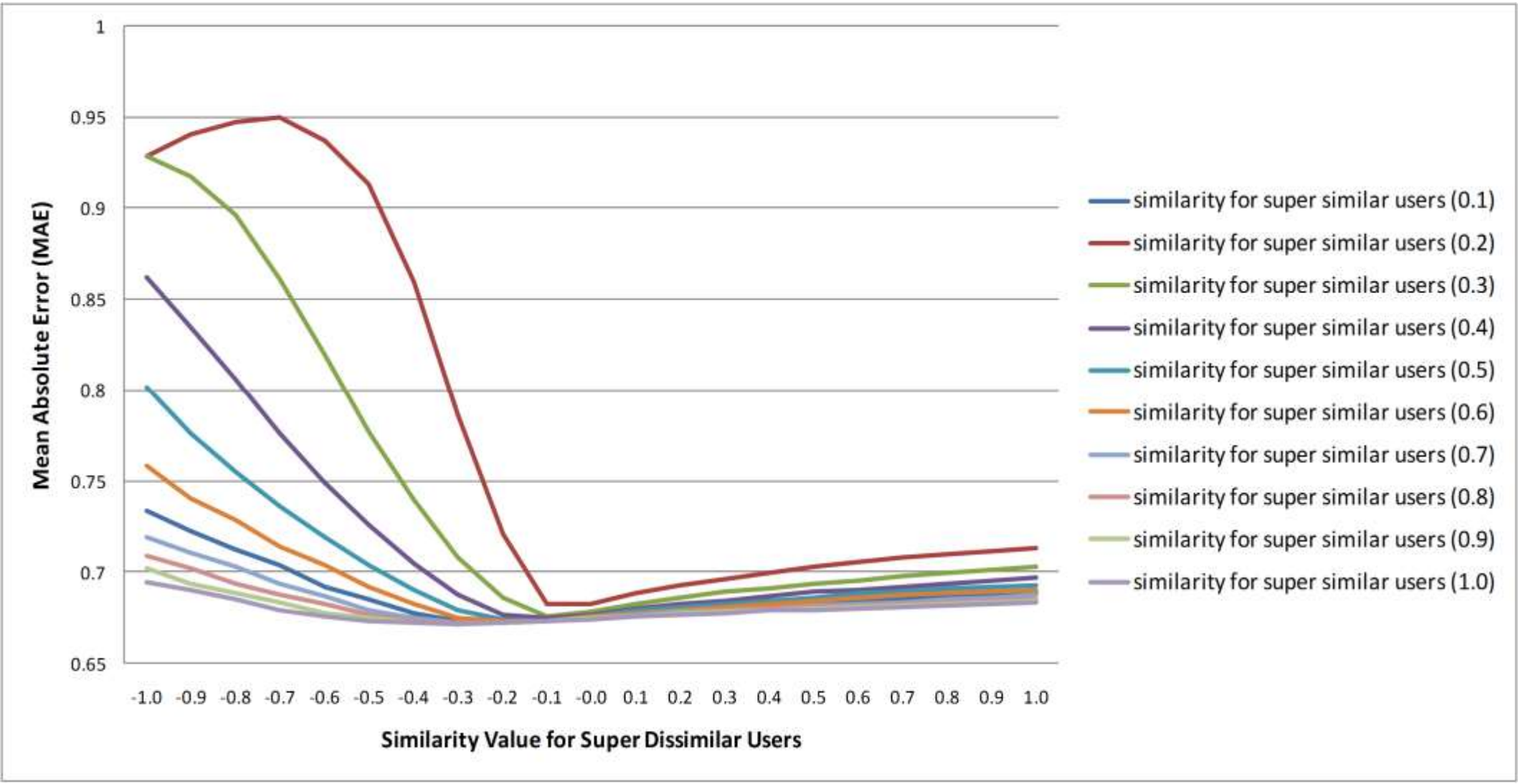
Mean Absolute Error

- $MAE = \frac{1}{U}\sum_{u \in U}\sum_{i \in 0}\frac{|\Pr(u,i) - Or(u,i)|}{I_u}$
- I_U represents the number of training items rated by the user *u*. *Pr*(*u*,*i*) and *Or*(*u*,*i*) respectively mean the predicted rating and original rating made by user *u* for item *i*

Reason for Setting $S_{xy} \geq 5$ in Support Matrix for Super Similar Users



Reason for Setting Similarity Value of Super Dissimilar User Pairs as -0.3



Experimental Dataset and Result

Dataset: Movielens

Users	Items	Ratings
6040	3,900	1,000,209

MAE values for different support counts of super similar users

super similar ($\geq$5) + super dissimilar ($<$5) (no average user here)	Super similar + JMSD	super dissimilar + JMSD	JMSD	super similar + average user + super dissimilar
0.673	0.675	0.735	0.680	0.668

Conclusions

- This is our initial footstep in metric aggregation. In future we hope to implement different similarity metrics for different user pairs.

- The aggregation can be performed in different other ways. As we proved through experimentation that our metrics using aggregation performs better than each of the metrics individually, we can expect that aggregating metrics more intelligently will yield better result.

References

1. Fernando Ortega, José-Luis SáNchez, JesúS Bobadilla, and Abraham GutiéRrez. Improving Collaborative Filtering-based Recommender Systems Results Using Pareto Dominance. *Inf. Sci.*, 239:50–61, August 2013.
2. Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, and Xuzhen Zhu. A new user similarity model to improve the accuracy of collaborative filtering. *Know.-Based Syst.*, 56:156–166, January 2014.
3. Hyung Jun Ahn. A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-starting Problem. *Inf. Sci.*, 178(1):37–51, January 2008.
4. Jesus Bobadilla, Fernando Ortega, Antonio Hernando, and Javier Alcalá. Improving collaborative filtering recommender system results and performance using genetic algorithms. *Knowledge-based systems*, 24(8):1310–1316, 2011.
5. Abhishek Kaleroun. *Hybrid Bee Colony Trust Mechanism in Recommender System.* PhD thesis, THAPAR UNIVERSITY, 2014.

Query or Suggestion?
You are most welcome