Washington State University
School of Electrical Engineering and Computer Science
CptS 315 – Introduction to Data Mining
Online

Ananth Jillepalli

Homework 2

Name: Nam Jun Lee

Student Number: 11606459

Question 1. Consider the following ratings matrix with three users and six items. Ratings are on a 1-5 star scale. Compute the following from data of this matrix:

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	4	5		5	1	
User 2		3	4	3	1	2
User 3	2		1	3		4

Table 1: Data of ratings from three users for six items.

a) Treat missing values as 0. Compute the jaccard similarity between each pair of users.

Jaccard similarity measure ignores the value of the rating:

User 1 = {Item 1, Item 2, Item 4, Item 5}

User 2 = {Item 2, Item 3, Item 4, Item 5, Item 6}

User 3 = {Item 1, Item 3, Item 4, Item 6}

Jaccard similarity between each pair of users: Number of observations in both / Number of observations in either

J(User1, User2) = |{Item2, Item4, Item5}| / |{Item1, Item2, Item3, Item4, Item5, Item6}|

= 3 / 6 = 0.5

J(User1, User3) = |{Item1, Item4}| / |{Item1, Item2, Item3, Item4, Item5, Item6}|

= 2 / 6 = 0.33

J(User2, User3) = |{Item3, Item4, Item6} / {Item1, Item2, Item3, Item4, Item5, Item6}|

= 3/ 6 = 0.5

b) Treat missing values as 0. Compute the cosine similarity between each pair of users.

Treat missing values as 0 for each users:

User 1 = {4, 5, 0, 5, 1, 0}

User 2 = {0, 3, 4, 3, 1, 2}

User 3 = {2, 0, 1, 3, 0, 4}

Cosine similarity between each pair of users: $sim(x,y) = cos(r_x,r_y) = r_x * r_y / ||r_x|| * ||r_y||$

Cos(User1, User2):

User1 * User2 = 4 * 0 + 5 * 3 + 0 * 4 + 5 * 3 + 1 * 1 + 0 * 2 = 31

||User1|| * ||User2|| = $\sqrt{4^2 + 5^2 + 0^2 + 5^2 + 1^2 + 0^2} * \sqrt{0^2 + 3^2 + 4^2 + 3^2 + 1^2 + 2^2}$

= 51.1175116765

= 31 / 51.1175116765

= 0.606

Cos(User1, User3):

User1 * User3 = 4 * 2 + 5 * 0 + 0 * 1 + 5 * 3 + 1 * 0 + 0 * 4 = 23

||User1|| * ||User3|| = $\sqrt{4^2 + 5^2 + 0^2 + 5^2 + 1^2 + 0^2} * \sqrt{2^2 + 0^2 + 1^2 + 3^2 + 0^2 + 4^2}$

= 44.8330235429

= 23 / 44.8330235429

= 0.513

Cos(User2, User3):

User2 * User3 = 0 * 2 + 3 * 0 + 4 * 1 + 3 * 3 + 1 * 0 + 2 * 4 = 21

||User2|| * ||User3|| = $\sqrt{0^2 + 3^2 + 4^2 + 3^2 + 1^2 + 2^2} * \sqrt{2^2 + 0^2 + 1^2 + 3^2 + 0^2 + 4^2}$

= 34.205262753

= 21 / 34.205262753

= 0.614

c) Normalize the matrix by subtracting from each non-zero rating, the average value for its user. Show the normalized matrix.

AVG(User1) = (4+5+5+1) / 4 = 15 / 4 = 3.75

AVG(User2) = (3+4+3+1+2) / 5 = 13 / 5 = 2.6

AVG(User3) = (2+1+3+4) / 4 = 10 / 4 = 2.5

Normalized Matrix:

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	0.25	1.25	0	1.25	-2.75	0
User 2	0	0.4	1.4	0.4	-1.6	-0.6
User 3	-0.5	0	-1.5	0.5	0	1.5

d) Compute the (centered) cosine similarity between each pair of users using the above nor- malized matrix.

Cosine similarity between each pair of users using normalized matrix:

$$sim(x,y) = cos(r_x,r_y) = r_x * r_y / ||r_x|| * ||r_y||$$

Cos(User1, User2):

User1 * User2 = 0.25 * 0 + 1.25 * 0.4 + 0 * 1.4 + 1.25 * 0.4 + (-2.75) * (-1.6) + 0 * (-0.6)

= 5.4

$||User1|| * ||User2|| = \sqrt{0.25^2 + 1.25^2 + 0^2 + 1.25^2 + (-2.75)^2 + 0^2} * \sqrt{0^2 + 0.4^2 + 1.4^2 + 0.4^2 + (-1.6)^2 + (-0.6)^2}$

= 7.47663025701

= 5.4 / 7.47663025701

= 0.722

Cos(User1, User3):

User1 * User3 = 0.25 * (-0.5) + 1.25 * 0 + 0 * (-1.5) + 1.25 * 0.5 + (-2.75) * 0 + 0 * 1.5

= 0.5

$||User1|| * ||User3|| = \sqrt{0.25^2 + 1.25^2 + 0^2 + 1.25^2 + (-2.75)^2 + 0^2} * \sqrt{(-0.5)^2 + 0^2 + (-1.5)^2 + 0.5^2 + 0^2 + 1.5^2}$

= 4.48305298651

= 0.5 / 4.48305298651

= 0.112

Cos(User2, User3):

User2 * User3 = 0 * (-0.5) + 0.4 * 0 + 1.4 * (-1.5) + 0.4 * 0.5 + (-1.6) * 0 + (-0.6) * 1.5

= -2.8

||User2|| * ||User3|| = $\sqrt{0^2 + 0.4^2 + 1.4^2 + 0.4^2 + (-1.6)^2 + (-0.6)^2} * \sqrt{(-0.5)^2 + 0^2 + (-1.5)^2 + 0.5^2 + 0^2 + 1.5^2}$

= 5.09901951359

= -2.8 / 5.09901951359

= -0.549

Question 2. Please read the following two papers and write a brief summary of the main points in at most TWO pages.

Amazon.com Recommendations Item-to-Item Collaborative Filtering

This article explains the customer recommendation algorithm used on the Amazon homepage. Recommendation algorithms are technologies that help consumers purchase by generating a list of recommended items using inputs to customer interests. Currently, customer recommendation algorithms are popularized on many sites, and each customer's consumption habits are customized by identifying customer purchase patterns and consumption tendencies. Large retailers have huge amounts of customer data, and there are three ways to address recommendations to these customer recommendation algorithms: traditional collaborative filtering, cluster models, and search-based methods.

Existing collaborative filtering algorithms represent customers as vectors of items and multiply vector components by inverse frequencies to make fewer known items more relevant. However, since this algorithm is computationally expensive, it is necessary to reduce the size of the data or to reduce the number of items by dividing the item space according to the product category or classification. The cluster model is a method of finding users-like customers and classifying the customer base into several sectors, which has excellent online scalability and performance, but is expensive. And search-based methods are search queries to find other popular items with keywords or themes like those purchased by users, which show good performance when there is less data, but poor performance within many data.

For Amazon, they are using inter-item collaborative filtering algorithms to generate real-time high-quality recommendations. This algorithm uses cosine measures to calculate similarities between two items in various ways. It has a very fast calculation

speed and, unlike traditional collaborative filtering, shows excellent performance with l imited user data.

As such, the recommendation algorithm is an effective way to create a custo mized shopping experience for each customer, and the distribution industry will furthe r improve and apply the recommendation algorithm for online and offline target mark eting in the future.

Two Decades of Recommender Systems at Amazon.com

This article describes Amazon's recommended algorithms and describes recom mended terms. Amazon has long constructed an algorithm for millions of customers to discover items that each customer has not found. The algorithm began by finding r elated items for each item in the catalog and expands many items without any other technology by producing the customer's current situation and previously purchased ite ms to remove items already purchased and recommend items remaining. By being use d in so many fields, Amazon has discovered a way to improve its recommendation al gorithm to a better one.

Machine learning allows computers to learn what customers prefer to see whic h parameters are best suited for the specific use of recommendations. Also, time play s an important role. This helps improve the quality of recommendations and has a sig nificant impact on the quality of recommendations.

As data grows larger and these customizations become more and more active in more areas, a new way of thinking about recommendations is needed, and in the f uture, intelligent computer algorithms that continue to use human intelligence to help people will need to be improved and adjusted.