CptS 315: Introduction to Data Mining

Homework 2

(Due date: Mar 5, midnight PST)

Instructions

• Please use a word processing software (e.g., Microsoft word) to write your answers and submit a printed copy to me at the beginning of the class on Feb 8. The rationale is that it is sometimes hard to read and understand the hand-written answers.

• All homeworks should be done individually.

Analytical Part (40 points)

Q1. 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: (20 points)

	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.

JSim (user 1, user 2) = (0 +3 +0 + 3 + 1 + 0) / (4 + 5 + 4 + 5 + 1 + 2) = 7/21 = 1/3

JSim (user 1, user 3) = (2 + 0 + 0 + 3 + 0 + 0) / (4 + 5 + 1 + 5 + 1 + 4) = 5 / 20 = 1/4

JSim (user 2, user 3) = (0 + 0 + 1 + 3 + 0 + 2) / (2 + 3 + 4 + 3 + 1 + 4) = 6 / 17

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

CSim (user 1, user 2) = (4*0 + 5*3 + 0 * 4 + 5 * 3 + 1 * 1 + 0 * 2) / (sqrt(16 + 25 + 0 + 25 + 1 + 0) * sqrt(0 + 9 + 16 + 9 + 1 + 4)) = 31 / (sqrt(67) * sqrt(39)) = 0.6064

CSim (user 1, user 3) = (4 * 2 + 5 * 0 + 0 * 1 + 5 * 3 + 1 * 0 + 0 * 4) / (sqrt(16 + 25 + 0 + 25 + 1 + 0) * sqrt(4 + 0 + 1 + 9 + 0 + 16)) = 23 / (sqrt(67) * sqrt(30)) = 0.5130

CSim (user 2, user 3) = (0 * 2 + 3 * 0 + 4 * 1 + 3 * 3 + 1 * 0 + 2 * 4) / (sqrt(0 + 9 + 16 + 9 + 1 + 4) * sqrt(4 + 0 + 1 + 9 + 0 + 16)) = 21 / (sqrt(39) * sqrt(30)) = 0.6139

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

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
average	(4 + 2) / 2 = 6 / 2 = 3	(5 + 3) / 2 = 8 / 2 = 4	(4 + 1) / 2 = 5 / 2 = 2.5	(5 + 3 + 3) / 3 = 11 / 3 = 3.667	(1 + 1) / 2 = 2 / 2 = 1	(2 + 4) / 2 = 6 / 2 = 3

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	4 – 3 = 1	5 – 4 = 1	0	5 – 11 / 3 = 4 / 3 = 1.333	1 – 1 = 0	0
User 2	0	3 - 4 = -1	4 – 2.5 = 1.5	3 – 11 / 3 = -2 / 3 = -0.667	1 – 1 = 0	2 – 3 = -1
User 3	2 – 4 = -1	0	1 – 2.5 = -1.5	3 – 11 / 3 = -2 / 3 = -0.667	0	4 – 3 = 1

d) Compute the (centered) cosine similarity between each pair of users using the above normalized matrix.

CCSim(user 1, user 2) = (1 * 0 + 1 * (-1) + 0 * (1.5) + (4/3) * (-2/3) + 0 * 0 + 0 * (-1)) / (sqrt(1 + 1 + 0 + (16/9) + 0 + 0) * sqrt(0 + 1 + (9/4) + (9/4) + 0 + 1)) = -0.4485

CCSim(user 1, user 2) = (1 * (-1) + (-1) * 0 + 0 * (-1.5) + (4/3) * (-2/3) + 0 * 0 + 0 * (-1)) / (sqrt(1 + 1 + 0 + (16/9) + 0 + 0) * sqrt(1 + 0 + (9/4) + (4/9) + 0 + 1)) = -0.4485

CCSim(user 1, user 2) = (0 * (-1) + (-1) * 0 + (1.5) * (-1.5) + (-2/3) * (-2/3) + 0 * 0 + 1 * (-1)) / (sqrt(1 + 0 + (9/4) + (4/9) + 0 + 1) * sqrt (1 + 0 + (9/4 + ()9/4) + 0 + 1)) = -0.5976

Q2. Please read the following two papers and write a brief summary of the main points in at most TWO pages. (20 points)

Brent Smith, Greg Linden: Two Decades of Recommender Systems at Amazon.com. IEEE Internet Computing 21(3): 12-18 (2017)

https://www.computer.org/csdl/mags/ic/2017/03/mic2017030012.pdf

Greg Linden, Brent Smith, Jeremy York: Industry Report: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Distributed Systems Online 4(1) (2003)
https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

Summary

The Article “Two Decades of Recommender Systems at Amazon.com”, shows us that the recommendation algorithm first used by Amazon has begun to spread since 20 years ago. Since then, the algorithm has been continuously updated to become more complete. Recommendation algorithms can be widely used in daily life. After users purchase a product and find a similar recommended product, they will consider continuing to consume it. Whether the customer is interested is the purpose of the recommendation algorithm. This article provides many examples, but the main purpose is to infer what customers are interested in and drive customer consumption.

In the article “Amazon.com Recommendations: Item-to-Item Collaborative Filtering”, the article mainly introduces the recommendation algorithm, and takes Amazon website as an example. While praising the recommendation algorithm, explained the advantages and disadvantages of several recommendation algorithms. First of all, the article introduces us to the recommendation algorithm, just like its name, based on collecting and understanding the interests of customers to generate a list of recommended items. Customers who frequently purchase and evaluate high-scoring products can be included in the list of recommended items. Not only that, the recommendation algorithm should also be able to make recommendations based on statistical data that customers frequently search for. Secondly, this article shows how large retailers use recommendation algorithms to use large amounts of data to provide customers with recommendations in a short period of time. This is similar to a personal service store, which can make it easier for customers to find products of interest through data accumulation. This article shows us three common methods for solving recommendation problems: traditional collaborative filtering, clustering models, and search-based methods. The author also compared the algorithm called item-by-item collaborative filtering they researched. Let us understand the advantages of this algorithm: it can produce recommendations in real time, scale to massive data sets, and generate high-quality recommendations. And it focuses on finding similar projects, not similar customers.

The traditional collaborative filtering uses a cosine metric, but when the data is very large, it will encounter serious performance and scaling problems. These scaling issues can be partially addressed by reducing the data size, but the quality of recommendations is also reduced. Cluster models are better than collaborative filtering because they are segmented before they are compared. Instead of comparing the entire data complex and expensive cluster computing can run offline. however, the quality of recommendations is very low. At the same time, because the similar customers found by the cluster model are not the most similar customers, the advice they produce is not very relevant. It's not worth it if you organize your data through many segments. Next is item by item collaborative filtering. By calculating the similarity between two items, such as the cosine metric. Sampling can reduced the running time, and the quality was hardly reduced. Given a similar product table, find products that are similar to each product in the user's purchases and reviews, aggregate those products, and then recommend the most popular or relevant products. The calculation speed of the algorithm is very fast, and it only depends on the number of items purchased or evaluated by the user.

Conclusion,the use of recommended algorithms should be targeted, there is no perfect algorithm, only suitable algorithms.