Harry Pines

Data Mining 315 Homework 2

Analytical Part (40 points)

Q1. Consider the following ratings matrix with three users and six items. Ratings are on a1-5 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 |

1. Treat missing values as 0. Compute the Jaccard similarity between each pair of users

J(1,2) = {4,5,1}/{1,2,3,4,5} = 3/5 = 0.6

J(1,3) = {4,1}/{1,2,3,4,5} = 2/5 = 0.4

J(2,3) = {4,3,1,2}/{1,2,3,4} = 4/4 = 1

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

Sim(1,2)=

Sim(1,3)=

Sim(2,3)=

|  |  |  |  |
| --- | --- | --- | --- |
|  | User 1 | User 2 | User 3 |
| User 1 | 1.0 | 0.60644 | 0.51301 |
| User 2 |  | 1.0 | 0.61394 |
| User 3 |  |  | 1.0 |

1. 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 |
| User 1 | 0.25 | 1.25 |  | 1.25 | -2.75 |  |
| User 2 |  | 0.4 | 1.4 | 0.4 | -1.6 | -0.6 |
| User 3 | -0.5 |  | -1.5 | 0.5 |  | 1.5 |

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

Sim(1,2)= 0.72225

Sim(1,3)= 0.10911

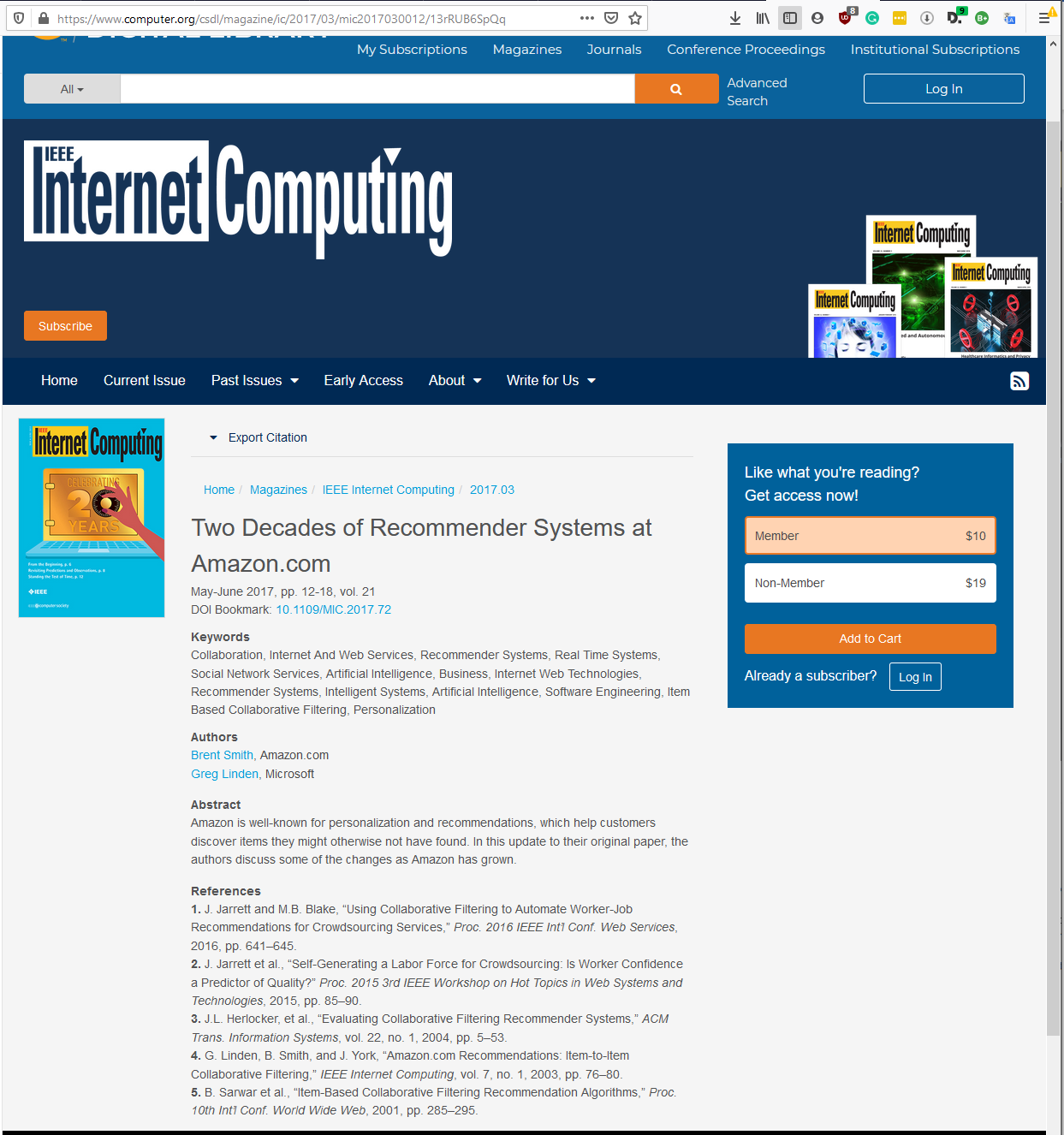
Sim(2,3)= -0.54912

|  |  |  |  |
| --- | --- | --- | --- |
|  | User 1 | User 2 | User 3 |
| User 1 | 1.0 | 0.72225 | 0.10911 |
| User 2 |  | 1.0 | -0.54912 |
| User 3 |  |  | 1.0 |

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. IEEEInternet Computing 21(3): 12-18 (2017) <https://www.computer.org/csdl/mags/ic/2017/03/mic2017030012.pdf>

The link unfortunately takes me to a paywall for the article:



I did however find the article elsewhere:

<https://pdfs.semanticscholar.org/0f06/d328f6deb44e5e67408e0c16a8c7356330d1.pdf>

Summary

The first article explains the basics of how amazon’s recommendation engine functions. Due to the sheer scale of data that amazon works with more and with than 29 million customers, it’s important that any system implemented functions promptly. This is where item-item collaborative filtering comes in. User-user filtering is computationally heavy and gets more so for any increase in dataset size. The problem is worsened when attempts are made to filter or reduce the dataset as it heavily impacts the quality of the recommendations made. Thus, user similarity filtering is impractical for real-time recommendations on such a large-scale platform. Clustering is equally inefficient as while it can assign new users to a cluster quickly, its recommendations aren’t specific to the user but rather the cluster that the user is a member of and thus delivers poorly related recommendations. Instead, Amazon uses item-item collaborative filtering where a matrix of probability is constructed to infer if item y should be recommended having bought item x. Much of the modern challenge with the implementation is considering both time and time directionality. Relationships between items x and y become far less relevant as the time between each one increases, and it will require far more intelligent algorithms to build user profiles that can better understand a user’s desire over longer spans. The order in which items are recommended is also important. Just because x infers y shouldn’t also necessarily mean that y infers x. You don’t typically buy the memory card before the camera. Because the algorithm scales independently of the size of amazon’s catalog, the runtime is only dependent on how active the user is on the platform.

In conclusion, the articles summarized the effectiveness of recommendation algorithms for targeted marketing application s for online retail. They elaborated on why some methods are less time effective than others and how scalable methods are the most practical for customer wide datasets. Standard collaborative filtering is inadequate as it struggles on larger data sets unless it's reduced by some means which results in reduced quality of recommendations. Both papers agree that there is a future to recommendation engines, whether that be making them adaptable to users changing tastes, recommending products in the future based on purchases now or understanding the differences between one-off short-term buys and longer hobby investments. Recommendation systems still have a long way to go.