**CptS 315: Introduction to Data Mining**

**Homework 2 (HW2)**

**Instructions**

* Please use a word processing software (e.g. Microsoft word) to write your answers.
* You will need to submit your answers as a pdf file on Blackboard.
* This assignment is due by the date stated on Blackboard.
* All homeworks should be done individually.

**Q1. (50 points)** 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 | (0) | 5 | 1 | (0) |
| User 2 | (0) | 3 | 4 | 3 | 1 | 2 |
| User 3 | 2 | (0) | 1 | 3 | (0) | 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: J(

J(User1, User2) =

J(User 1, User3) =

Jaccard(User2, User3) =

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

Cosine Similarity: sim(

sim(User1, User2) =

sim(User1, User3) =

sim(User2, User3) =

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

User 1 Average:

User 2 Average:

User 3 Average:

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

**Normalized Matrix:**

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

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

(Centered) Cosine:

sim(User1, User2) =

sim(User1, User3) =

sim(User2, User3) =

**Q2. (50 points)** Please read the following two papers and write a brief summary of the main points in at most TWO pages.

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>

There is a reason why Amamzon.com is one of the leading companies in the world. How they manipulate their data from customers to benefit their business needs is always improving. Their recommendation algorithms have become so successful and is a huge reason why customers continue to shop at Amazon.com. They believe every interaction the customer makes, should be a recommendation. That includes all navigation through the website and what their current and past behavior context is as well.

Amazon.com came out with its first item-based collaborative filtering in 1998. It had an advantage of viewing all millions of users without having to use a sampling or other technique. This helped keep the quality of the recommendations more accurate. It began by finding related items for each item in the catalog, so for every item i1, we want every item i2 that was purchased with unusually high frequency by people who bought i1. Once the related items table is built, we generate recommendations quickly as a series of lookups. For each item that’s part of this customer’s current context and previous interests, we look up the related items, combine them to yield the most likely items of interest, filter out items already seen or purchased, and then we are left with the items to recommend (Brent Smith).

In order for the quality of recommendations to work, it is important to also understand the role of time. The context of time is many things according the Amazon.com, including what a customer’s next moves are, new items that are added to the website, the lifecycle of customers, and the diversity in recommendations. Most significantly, the importance of diversity in recommendations is well defined. From a business perspective, it is better to show the customer multiple related items to what they are searching for than a narrow list. This gives the customer the satisfaction that they have can find what they are looking for but also have the satisfaction of exploring new things. This also comes with the territory that when a customer’s search is unclear, discovery and serendipity should be the ultimate goal (Brent Smith). “Finding the right balance in the diversity of recommendations requires experimentation along with a willingness to optimize for the long term” (Brent Smith).

Amazon.com’s recommendation algorithms are always facing new problems with the mass amounts of traffic and data always reaching their site. Some examples are; huge amounts of data with millions of distinct catalog items, applications require the results set to be returned in real-time, new customers typically have extremely limited information, Older customers can have a glut of information, and Customer data is volatile (Greg Linden, Brent Smith, Jeremy York). Contrary to these issues, their Item-to-Item collaborative filtering works because of its real-time, scalability to massive data sets, and generates high-quality recommendations.

Amazon.com’s Item-to-Item collaborative filtering works by matching each of the user’s purchased and rated items to similar items, then combines those similar items into a recommendation list (Greg Linden, Brent Smith, Jeremy York). To determine the most-similar match for a given item, the algorithm builds a similar-items table by finding items that customers tend to purchase together. It is a better approach to calculating the similarity between a single and related product (Greg Linden, Brent Smith, Jeremy York). The scalability and performance of their algorithm creates the expensive similar-items table offline and also is fast for the mass amounts of data they have.

Overall, Amazon.com’s algorithm has a high-quality recommendation system and can produce recommendations based on as few as two or three items. It is extremely important for a company like Amazon.com to personalize customer’s shopping experiences. This is the reason why they have such loyal customers and the company has been evolving as the years go on.