Amazon Product Recommendations Using Collaborative Filtering and Natural Language Processing

Nikhil Aourpally
Nikhil.Aourpally@asu.edu

Anthony Helmstetter
anthony.helmstetter@asu.edu

Prajwal Paudyal
ppaudyal@asu.edu

Anirudh Som
Anirudh.Som@asu.edu

Abstract

In this project we will implement a Recommender System to provide suggestions based on Amazon user reviews across Amazon Instant Video, Music, and Book product categories. The Recommender System will be implemented using collaborative filtering based on product ratings. To increase recommendation accuracy, we will also perform natural language processing on user written reviews to better inform the recommender system of user preferences. We will assess the accuracy of the Recommender System with and without additional textual information containted in user reviews, and evaluate the sensitivity of our Recommender System by varying the number of reviews by user, and number of reviews per product.

1 Project Description

Using Amazon product reviews, we will implemenet a recommender system to suggest new products to Amazon users. Suggested products will be determined by highly rated products as rated by similar users. Similarity of users will de betermined by the degree to which users have rated identical products similarly. This type of recommendation system falls under the category of collaborative filtering, in which we filter based on similarity of users and not similarity of products. Product ratings by users are given as integers between 1 and 5, which is a rather coarse grained metric by which to determine the closeness of user preferences.

To increase the performance of our recommender system, we will also incorporate additional information dervied from text reviews written by users, to develop a more accurate metric of user similarity. We will explore various methods in natural language processing in order to better inform our recommender system. The performance of the recommender both with and without incorporating additional textual information will be assessed and compared. Specifically, we hope to compare the perfomance of the recommender in instances where users or products have too few reviews to be reliably useful without incorporating text reviews.

If time permits, we plan on exploring content based filtering wherein we make recommendations determined by product similarity, as opposed to user similarity. In the case of movie recommendations, this would be acheived by peforming movie "tag" analysis or movie description processing, scraped from sites like IMDB.com.

2 Dataset

The Dataset consists of three datasets: Amazon-Instant-Video.txt, Books.txt, and Music.txt. All datasets are publically available at https://snap.stanford.edu/data/web-Amazon.html. The Amazon-Instant-Video is 252 MB and consists of 717,651 reviews, the Books dataset is 4.4 GB and consists of 12,886,488 reviews, and the Music dataset is 2.1 GB and consists of 6,396,350 reviews. A typical entry is shown below.

```
product/productId: B002BNZ2XE
product/title: Amazon.com: Moving Target (1988): Jason Bateman, John Glover,
Jack Wagner, Chynna Phillips: Amazon Instant Video
product/price: 0.00
review/userId: AZVY9Y3A0YU1I
review/profileName: Willard R. Stephen
review/helpfulness: 1/1
review/score: 4.0
review/time: 1277424000
review/summary: Moving Target
review/text: I found this to be an intrieging suspense thriller with a
minimum of violence and a happy ending. This is a rare combination in
recent releases. I found it well excepted and enjoyed by teen age
audiences.
```

Each entry in the dataset consists of a productId, reviewrId, Score, and TextReview. The productId, reviewId, and Score will be used to implement a collaborative filter recommender system, further informed by the textual information contained in the TextReview.

3 Expected Roles

While we expect each member to contribute in each facet of the project, our team orginzational structure will consist of designated "team leads"for core features of the project. These core features and their leads are:

0. Project Management, led by Anirudh
1. Programming, led by Prajwal
2. Documentation, led by Anirudh
3. Mathematics / Theory, led by Anthony
4. Data Preprocessing, led by Nikhil
5. Data Visualization, led by Nikhil

Project management consists of coordinating meetings, managing project timelines, and managing development tools and environments. Programming will be done by all team members, with Prajwal prioritizing and delegating sub tasks. All code will be documented by the programmer, but will be managed and curated by Anirudh. Anthony will lead the mathematics research and theory behind the project, documenting throughout. While everyone will to some extent be involved in data preprocessing and visualization, Nikhil will take the lead and delegate as needed.

References

[1] McAuley, J. & Leskovec, J. (2013) Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text.*AProceedings of the 7th ACM conference on Recommender systems*, pp. 165-172. New York, NY: ACM.

[2] Recommender System (2015). https://en.wikipedia.org/wiki/Recommender_system

[3] SNAP Web data: Amazon Reviews. https://snap.stanford.edu/data/web-Amazon.html