Collaborative Filtering based Recommendation System

The Data

The dataset that was used in the project is a free dataset titled Book Crossing dataset [1]. It was mined by Cai-Nicholas Ziegler in summer 2004 and consists of three csv files:

- BX_Users: It contains data from 278,858 users (User ID, Location, Age).
- BX_Books: It has information for 271,379 books (Title, Author, Year of Publication, Publisher and URL details).
- BX_Book_Ratings: It contains 1,149,780 ratings provided by the users (User ID, ISBN of the book and book rating).

Data Analysis

A data analysis was conducted to provide an assessment of data quality, identify the presence of outliers, duplicates, data inconsistencies and missing values.

Users Dataset

First of all, we confirmed that at the Users datasets all UserID entries are unique.

Then we noticed that Location is a combination of City, State and Country, which means that we can split the Location string into City, State and Country. Furthermore, since there are different capitalization styles, we used a standard capitalization rule for all strings. We also cleaned up the Location data by replacing some ASCII characters with appropriate characters (e.g. æ with ae, å wit aa), replacing some missing values that are “n/a” text strings or empty strings with NaN values, removing numbers and removing some special characters from the strings (e.g. ` or ? or *). This is a very crucial step, since a particular item may not be recognised as having the same value as another if one has a special character included and another does not.

The Age field is poorly populated, as it has 39.72% missing values. We also observed that the age range goes from 0 to 244 years old. Obviously this does not make much sense and, hence, all values below 5 and above 100 were replaced with NaNs.

Books Dataset

First of all, we confirmed that there are not any duplicate books based on the ISBN.

The image URLs columns do not seem to be required for analysis, and hence they were dropped off.

YearofPublication has values ranging from 0–2050. All years of zero were replaced with NaN values. Furthermore, as this dataset was built in 2004, and as there are only a few (18 in total) books with outliers in the YearofPublication, we removed them from the Books table, as they could potentially skew the model.

The Publisher, BookAuthor and BookTitle data were cleaned up, as we did with the Location strings from the Users dataset.

A very important observation in the Books dataset is that although the ISBN column has only unique values, there are many books with the same title. These may be books corresponding to a different version of the book. Since our goal is to build a Book Recommender, our recommendations will be for a book, not a specific edition of a book. So, we created a new field, where all these cases (i.e., same BookTitle- BookAuthor pair) were considered as the same entry.

Ratings Dataset

First of all, we discovered that there are many books at the Ratings dataset that do not exist at the Books dataset. So, we filtered these books out, since we cannot recommend them.

The ratings of each user are either implicit (0) meaning the result of observed behavior, or explicit (1-10) meaning the user rated the book. The vast majority of ratings are 0. For building our book recommendation system, we used only explicit ratings, and so 0 rating entry rows were removed.

After this processing step, the above datasets were merged to provide a single consolidated dataset.

Building a recommender system using collaborative filtering

In collaborative filtering approach, the model is built from a user’s past behavior as well as decisions made by similar users (i.e., users with similar preferences).

In order to build the recommendation system, we used the surprise package. The following algorithms were implemented:

- Matrix Factorization-based algorithms (Singular Value Decomposition or SVD and Non-Negative Matrix Factorization or NNMF)
- Co-clustering

Matrix Factorization techniques predict the unknown ratings based on a low rank approximation of the original ratings matrix. These techniques have the ability to discover the latent (hidden) features underlying the interactions between users and items (books) [2].

The co-clustering approach predicts the unknown ratings based on the average ratings of the co-clusters (user-item neighborhoods) while taking into account the individual biases of the users and items [2].

Note that we decided to not use Memory-based algorithms because they do not scale well, as they are stored in memory and tend to overfit when we have data with high sparsity level.

In order to evaluate the accuracy of predicted ratings we used the Root Mean Squared Error (RMSE). RMSE measures the distance between the predicted preferences and the true preferences over items (i.e. the ones given by users).

SVD has the best performance compared to the other two algorithms, so we optimized it with hyper parameter tuning, in order to improve the predictions even further. Using the optimized hyperparameters we see a slight improvement in the test RMSE compared with the unoptimized SVD algorithm.

Conclusion

In this project, we used the Book Crossing Dataset to create a recommendation system. We examined three different algorithms, with the Singular Value Decomposition (SVD) algorithm giving the best performance. A "grid search" method was used to optimize some of the model hyperparameters, resulting in a slight improvement in model performance.

References

[1] http://www2.informatik.uni-freiburg.de/~cziegler/BX/

[2] T. George, S. Merugu, “A scalable collaborative filtering framework based on co-clustering”. Proceedings of the Fifth IEEE International Conference on Data Mining. ICDM ’05. Washington, DC, USA. 2005. 625–628.