Book Recommender System: Visual Recommender Supported by Visual Analytics

Vijaya Venkata Nischal Samji
Arizona State University
+1-480-469-5412
vsamji@asu.edu

ABSTRACT

Recommendation Systems are ubiquitous in the present day world. Most of the recommendation systems that are widely used use some form of identifiable personal information to deliver content that is specific to a particular person or entity. My project aims to design an effective but non personalized recommender system that focuses on general criterion rather than personalized parameters for recommending a particular book/ object. In addition to creating a recommender system my visualization prototype focuses on creates an interactive visualization that shows the analysis for different "objects/ books" and the trend that backs a successful object/ book.

Categories and Subject Descriptors

D.3.3 [**Programming Languages**]: Language Constructs and Features – *abstract data types, polymorphism, control structures.* This is just an example, please use the correct category and subject descriptors for your submission. The ACM Computing Classification Scheme: http://www.acm.org/class/1998/

General Terms

Algorithms, Design, Human Factors, Programming

Keywords

Visualization, Recommender Systems, Sentiment Analysis, Text Analysis, Analytics

1. INTRODUCTION

Many Websites that provide suggestions to user, like consumer electronics, gadgets, movies, books etc., rely highly on personal information. Websites like Netflix (www.netflix.com) Amazon (www.amazon.com), Google (www.google.com), Flipkart (www.flipkart.com) provides suggestions for user like products, search results, recommendations based on personally identifiable information like the age, previous search history to deliver content targeted to a particular individual/ entity. While the per-user recommendations and results increase the user's personal experience, it makes the user closed in a box and provide deeply personalized results thus increasing the "depth" for a user and limiting him on the breadth of the recommendations he choose. One particular, use-case where the depth and breadth of the recommendations matter in case of books (both digital and non-digital.) Book readers (amateur, hobbyist and avid) can be widely categorized into 40 types [1] based on Mark O Connell's article [2]. My analysis and Visualization serves the "elite" of the book readers. Readers who do not stick to a particular genre, author or type of the book but readers who enjoy reading. Most Book recommender systems prominently Goodreads (www.goodreads.com) and Amazon (www.amazon.com) provides suggestions based on a particular user's history – the books he liked and the books he previously read. But when a new user is trying to pick up a book, these recommendation systems cannot provide an accurate prediction. Moreover, users are shown the results that are based on immediate results i.e. books that are "trending" – books purchased by most people at a particular point of time and books that are "talked" about. As an example, when "Harry Potter and the Deathly Hallows" was released as a movie in 2009, non-personalized recommendation systems provided results that are related to Harry Potter over other books. While this is perfectly fitting for a recommendation scenario, it does not provide a generic recommendation for readers who chose to read "a" book over readers who chose to read "the" book. Based on these observations and personal experiences with these recommendation systems, I prototyped a visualization that helped users pick a book that he can enjoy without accessing any information that is bound to a specific user. The key attributes that went into building this system are "popularity" and "enjoyability" Unlike consumer goods like mobile phones, gadgets and other tools the quality of a book cannot be decided by an expiry date/ ease of use/ price but by user satisfaction which can be measured using the sentiment of the users. The recommendations in this use case can also be provided based on closely related books that have similar sentiment associated with them.

2. MOTIVATION

Based on the above mentioned experiences, I feel that a personalized recommendation system is very apt in many scenarios but at the same time, a recommender system that recommends books without any prior information is also needed. To recommend the books on a smaller scale, I chose to analyze a smaller but popular set of books – "Books Everyone should read in their life time" a book list curated and created by readers comprising of various demographics. Lack of simple recommender systems and lack of proper book recommender systems motivated me to design this visualization. The current prototype sets to analyze how users like a particular book i.e. discover the trend why some books are popular than other books and what drives the popularity of a book are visualized.

3. Evaluation Plan

3.1 Data Collection

The initial "seed" of the data was book lists – 100 books that everyone should read in their lifetime [3] and books extracted from

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference'10, Month 1–2, 2010, City, State, Country.
Copyright 2010 ACM 1-58113-000-0/00/0010 …$15.00.

Goodreads' famous booklists [4]. The initial data contained only the details about the books and their attributes like author, ratings, reviews and description. Using the fair usage policy of the Goodreads.com API[5], data related to these books – user comments, user ratings, user reviews was collected. To have a proper recommendation system that shows sufficient number of books and perform a proper analysis, a total of data about 557 books, 38340 user comments (approx. 300 - 900 per book.) 120000 rating details (approx.. 100 – 3000 per book were collected.) Using this data was slow as the Goodreads API[5] restricted only one request per second. The Data thus collected was stored as a .csv and .json that contained the following attributes book_id, book_name, author_name, reviews, ratings, comments etc., The data collection was done using python[6], because of relative use of use.

3.2 Data Cleaning

As the main analysis was sentiment analysis on comment data for each book, the collected data was to be cleaned before it was to be analyzed and processed. Bookreads (www.bookreads.com) has a huge user database spanning lot of geographies and this allowed the user to comments in his/ her choice. As sentiment analysis tools required the comment data in English, comments created in languages other than English were ignored. Any comment data that has characters outside of the standard English – ASCII character set was completely deleted and data that had a closer relationship like French was searched for related words and rest of the comments was deleted. Data Cleaning was done using python NLTK toolkit [7].

3.2.1 Removal of Stop words

Using python's inbuilt stop words for English corpora [8], all the stop words from the user comments were completely removed. The Resulting dataset still contained punctuation marks.

3.2.2 Removal of Punctuation

Using standard regex strings in python, punctuation marks such as '!', ',', '.' etc., was cleaned. The resultant text contained whitespaces and the keywords.

3.2.3 Single Word Comments

Comments that contains single words were completely removed as they did not contribute a signify sentiment or serious sentiment.

3.3 Data Analysis

Initial Data Analysis was conducted using Tableau, to find patterns from data using different classification parameters such as ratings, reviews, pages etc., using this technique, an interesting pattern was discovered in the book data. Advanced Data Analysis was conducted using python NLTK and TextBlob[10]. Using these tools the sentiment related to each comment was calculated. The process of analysis and classification of every sentiment is discussed in the next section. The Sentiment Analysis used on the data is a key input parameter for the recommendation system.

4. Sentiment Analysis on Comment Data

For each comment from each of the books, Sentiment Analysis was done using Text Blob's built-in Naïve Bayes Classifier. The key approach in classifying a comment as negative/ positive was done using this approach. As the books selected were highly popular using positive words and negative words from existing corpora could not give any results that showed the discovered pattern. So, a training set manually made was used to train the classifier.

4.1.1 Popularity Vs Critical Acclaim

To classify each book's popularity – Every comment that signified reader satisfaction was marked negative and any book that had high critical acclaim but low user enjoyment was marked negative. By using this type of classification and training data. Score related to each sentiment - Popularity and Critical Acclaim was calculated and normalized between 0 and 1. Books that had a high user popularity and high critical acclaim had both scores closer to 1. Books that had high user popularity but low critical acclaim had popularity score closer to 1 and critical acclaim closer to 0 and similarly for the other set of data. Using these scores each book was marked accordingly. The following table shows the details about the training data used in the classifier and the ranking system.

Table 1. Table Showing the Comments and associated Sentiment

Comment	Sentiment
I loved this book. Would totally read it again	pos
This book made me cry. Changed my world altogether.	pos
Very good attempt by the author. But his language is sub standard	neg
Read this book to lose your command in English	neg
The words used were too simple.	neg

4.1.2 Challenges in Sentiment Analysis

Unlike other products where a negative sentiment can be easily classified using existing methods and techniques, books cannot be easily classified whether a positive/ negative sentiment is conveyed. As an example, "Kite Runner made me cry, I had tough time picking myself up after reading this book" this comment conveyed a positive emotion about the book. Classification of these type of user comments was tough and challenging which would be discussed in the further sections.

4.1.3 Overall Score and Accuracy

Using the above mentioned techniques, an accuracy of 73% was achieved. Each books sentiment positive/ negative was eventually normalized to a maximum value of 1and to a minimum value of 0.The score thus obtained was concatenated to each book in the final data file.

5. Research Questions

5.1.1 What makes a book popular?

Even though there are lot of books that were published decades ago, highly critically acclaimed and had loads of users, most books could not become popular, significant disparity between popularity and critical acclaim always existed.

5.1.2 How close are books related to?

What is the pattern involved when users like a particular kind of book, if a person likes a book X, what is the chance that he likes another book Y.

6. Implementation of the Visualization

6.1 Prototyping

Using the processed data, preliminary analysis was carried out to prepare a draft form of the visualization, no. of ratings/reviews were used to find the existing patterns, During the prototyping phase, a clear pattern emerged for the books. Prototyping that comprised of data analysis was done using Tableau [11]. The results from the tableau data analysis are demonstrated in the below figure.

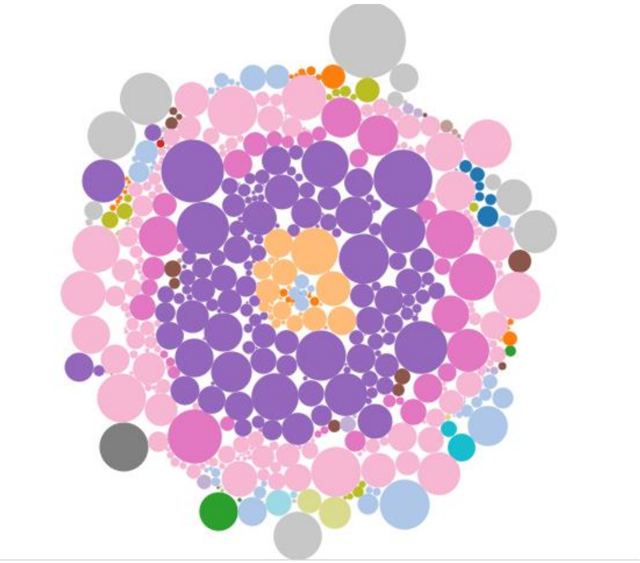

Figure 1. Prototype/ Data-Analysis of collected data using Tableau.

From, the above prototype of the data, results/ patterns clearly emerged that will be discussed in further sections of this paper. Each color in Figure 1, signified the genre of the book, the size represented the no. of ratings of each book. And the overall cluster represents all the data in the data set.

6.2 Visualization Design

Based on the above findings, a modified version of a bubble-scatter plot was used to implement the visualization. Each bubble represented a book collected and the size of the bubble reflected the filtering criteria chosen by the user for that particular set. The initial visualization when the user loads the page, shows the popularity index. Based on the user choices, the data is dynamically loaded and represented on the visualization area. Each filtering criteria, provides a kind of recommendation system for the user that prompts him to choose one of the bubbles and generate more recommendations based on the current choice. To accommodate high user engagement, proper coloring and classification techniques were used. The principles/ design methodologies and techniques used for developing the visualization are further discussed.

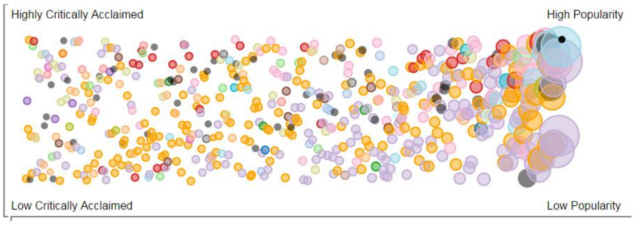

Figure 2. Initial View

In figure 2, the axes, books and the distribution of the data is shown. Following are the attributes that went into the Visualization and their significance.

6.2.1 Axes

The X-Axis of the system signifies increasing popularity as one moves from the right to left. The Y-Axis of the system signifies, increase in the critical acclaim as one moves from bottom to top. As an example, a book that is highly critically acclaimed and highly popular lies on the upper right corner of the system while a low popular and low critical acclaimed book lies in the left bottom of the grid.

6.2.2 Color of the Bubbles

Each color of the bubble signifies a particular genre (fiction, non-fiction, religion etc.)

6.2.3 Size of the Bubble

The Size of the bubble is something that is dynamically loaded every time the user interacts with the system. The user can view the visualization on the basis of no. of pages, no. of ratings or number of reviews for the set of books that are particularly loaded.

6.2.4 Interactions

On mouse-over of each bubble, the particular bubble is highlighted using a simple function drawing the user focus, also the detail about each book is shown as a hover on top of each bubble.

6.2.5 Index of Colors

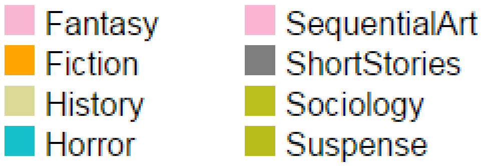

Figure 3. Index of the Colors

As shown in the above figure 5. The user is provided an index of colors that can be clicked upon to show the data about the particular genre.

7. Research Findings

Owing to the visual analytics used in the project, my research questions are answered. The most popular books have a movie associated with them. All the books that are popular i.e, which have high popularity have been made into a movie.

8. METHODOLOGY

Thus my visualization helped me in answering my questions of popularity in books and designing a visual recommender eventually. The clear right side popularity indicated all books that have a movie are more popular than other books in their particular genre and particular set of books.

9. FUTURE SCOPE AND LIMITATIONS

The recommender algorithm is still buggy and is not working as expected. Future Scope includes improving the algorithm and adding more filtering criteria for the existing system. Adding more books to improve the algorithm is part of the future scope.

10. REFERENCES

[1] http://www.thewire.com/entertainment/2012/08/many-more-types-book-readers-diagnostics-addendum/56425/

[2] http://www.newyorker.com/books/page-turner/promiscuous-reading.

[3] http://www.amazon.com/b?node=8192263011

[4] https://www.goodreads.com/list/show/9810.Books_To_Read_Before_You_Die

[5] https://www.goodreads.com/api

[6] https://www.python.org

[7] http://www.nltk.org/

[8] http://www.nltk.org/nltk_data/