Visualizing Reader-Created Tags for Books on GoodReads

Oreoluwa Arowobusoye, Tina Thanh Huynh, Anthony Tang
University of Calgary

ABSTRACT

This project focuses on the question of how to design a visualization that reveals underlying relationships between books (i.e. texts) based on user-generated tags as they relate to these books. The core problem the project addresses is that main mechanisms to classify texts--either based on literary classifications or text-analytics-based classifications are either too high-level, or too low-level. Rather, a perhaps more valuable way to explore relationships between texts is to use reader-generated tags. Here, we make use of data scraped from GoodReads, a social media site where the principal artefact under discussion is a book, and accumulated book reviews. In particular, we visualize the "bookshelves" each book is classified with (these "bookshelves" are user-generated tags), and we present two sets of visualizations that allow prospective readers with the ability to compare the thematic differences between different books.

Keywords: GoodReads, user-generated tags, visualization.

1 INTRODUCTION

The challenge we are addressing in this work is the problem of identifying texts that are similar to one another. Conventional methods of classifying content of texts is either too coarse-grained or too fine-grained. For example, the typical classification scheme relies on either author-entered keywords, or are done by experts. Popular categories of books include, for example, mystery, sci-fi, romance, classics, young adult, and so forth. Yet, the fact that these classifications are made once (and typically, books can only have one classification) make it fundamentally clear that these are too high-level. Two sci-fi books, for example, may explore entirely distinct themes, connected only loosely by the "sci-fi" category. Instead, two texts may be more thematically connected—both in terms of story arc, structure, plot points and so forth, yet this would not be captured by the "sci-fi" tag.

Another approach to this problem might be to explore natural language proessing approaches, which classify streams of text based on the structure, and to some extent structure. The problem with these approaches, however, is that analysis is fundamentally too low-level. The human experience of reading of a text is entirely gone—that is, how one infers and creates meaning from the text is entirely lost with these approaches.

In our work, we consider using reader-generated tags as a mechanism for not only organizing, but also comparing between texts. The intention is to design methods that allow people to compare between texts in order to identify texts that one might be interested in based on related themes or tags that other *readers* have identified, rather than relying on high-level classifications from "experts", or on natural language processing techniques. We use a two-pronged approach in our work: first, we gathered data from a social media site called GoodReads, which collects reader reviews of books, as well as their classifications of these books; second, we developed a series of interactive visualizations that allow an interested reader (or researcher) to explore these different visualizations in order to discover texts similar to one another.

While our work is ongoing, the purpose of developing these visualizations is to explore how to design effective interactive visualizations of this type of data. Our early explorations into this space are promising, and they are beginning to reveal a rich design space for further exploration.

* email address

2 RELATED WORK

Ridenour and Wooseob [1] address the visualization and analysis of common books shared amongst readers. They bring up the idea that books are more similar in appeal to other readers if a co-read pair of books occurs. They offer a graph with colour coded group nodes using Gephi to visualize this idea of clusters of pairs. This visualization can contribute to finding a more efficient way for learning about new books and improving a recommendations system for Goodreads. In our visualizations, our ultimate aim is to focus on book genre clusters rather than books that are co-read.

Deal [2] focuses on information visualization and how it is important as it allows for users to explore digital collections and information more easily. There is a challenge for organizing and managing collections online and Laura offers multiple solutions for doing so by highlighting different methods of visualizing and interacting with information. This includes geographic browsing using a website, spreadsheets, and pie charts. Data such as dates important dates of the Cold War and countries affected are visualized for users where they can interact with these collections of data. With these varying visualizations, Deal is able to gain more insight on which type is easier to interact with and learn from and understand.

3 DESIGN APPROACH AND DESIGN GOALS

Our general design approach was to develop sketches based on the fundamental concerns of people making use of these visualizations, and then to iterate on these sketches over time based on discussions with one another and potential users of our system. At the same time, we worked with the GoodReads API to discover exactly what kinds of data could be collected without needing to request special, additional access, and without violating terms of service.

We ultimately arrived at several design goals that governed the remaining designs that we ultimately realized through implementation:

Employ simple aesthetic style. While it is possible to encode multiple variables on a per book basis, we observed early on that this would be difficult for most readers to make sense of. Instead, we rely on a simple aesthetic style, where only very simple data is represented in the visualizations.

Visualizations must facilitate comparison. While it was interesting to visualize data about single book texts, a fundamental issue was being able to compare between different book texts. Minimally, we wanted to allow comparison between at least two texts, but future iterations of our work may facilitate multiple comparisons.

Transitions for detailed exploration. We appropriated the common focus+context style of visualization. Here, our visualizations begin with an overview, and allow for more detailed views to support curiosity-based exploration.

Interaction for exploration. Our goal was to provide interactive means of exploring the data as revealed in the visualizations. Each of our visualizations allow for this.

4 BubbleSets Visualization

The objective of this visualization (Figure 1) is to discover how data can be represented using the popular social media website Goodreads. One of the primary functions of Goodreads is to give users recommendations on what books to read next. For curious users, this relation of books brings up a new question. How are the books we read similar to each other, and in what way? Could it even be that we can find relationships between two seemingly unrelated texts? In order to answer these questions, we created an interactive visualization that showcases the relationships between a set of books through a users mouse hover activity.

In this interaction, two bubble sets are compared. Each of these bubble sets represent books, and are composed of user curated tags that can be found on Goodreads. Each of the internal bubbles represents the frequency with which a user tag has been associated with the particular book. We have also included a "brushing" mechanism, whereby when a bubble is hovered over, tags that are shared between the two books are both highlighted.

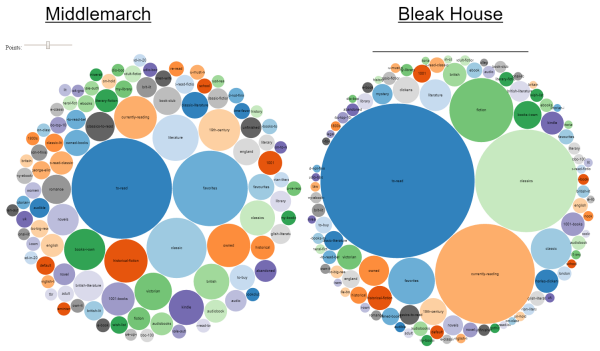

Figure 1: Each book is represented by a cloud of bubbles. Each of the internal bubbles represents a user-generated tag. Popular tags (i.e. multiple people tagged the book similarly) are represented with bigger internal bubbles.

5 Tag Popularity Visualization

The tag popularity visualization (Figure 2) compares the tags of a book a user has reviewed with the tags of the most recent book they have reviewed. The idea behind this is it allows for a user to see any differences or similarities between past and present books they've reviewed and if they've changed as a reader at all. In the graph, the bigger the sections, the more popular the shelf is for that book among users. The tags chosen were the top picks for each book. The user can switch views and the graph will change to a percentage perspective where the tags will take account the number of users that have read each book (as some books are more popular than others) and the sections will be scaled more proportionally. The user can also hover over each section to get detailed numbers and exact percentages for clarity. Currently, there are only a set number of users available selected randomly to choose from.

Here, we employ a far simpler aesthetic style to allow users to more easily compare and understand the differences between books (i.e. height comparison is easier than area comparison as in Figure 1).

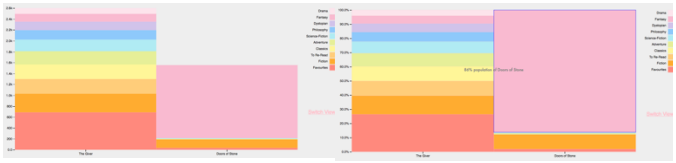

Figure 2: Each book here is represented by a vertical bar, with each of the stacks representing the number of people that had tagged the book with a given label (left), or the % of the tags for each label (right).

6 Conclusions

Our work is ongoing. At this point, we have developed a number of interactive visualizations based on our sketching process. Our aim is to evaluate these, and develop a set of guidelines to guide designers and researchers in the future to develop newer and more effective visualizations of user-generated tags for books.

References

[1] L. Ridenour, W. Jeong, "Leveraging the power of social reading and big data: an alaysis of co-read clusters of books on Goodreads," University of Wisconsin-Milwaukee, 2016.

[2] L. Deal, "Visualizing data collections," Woodrow Wilson International Center for Scholars, 2014.