273

From Chunks to Clusters: Identifying Similarity Features in Social Discussion

Yunseon Choi
Associate Professor, Valdosta State University

Socially-discussed online reviews

As socially-discussed online reviews have been noticeably increased, the online reviews have received significant attention among researchers, particularly in business contexts, because users' online reviews play a vital role in understanding users' opinions and interests about the products. The book features, such as genre and textual content, are factors to consider in selecting books, especially for children. There is still a lack of studies discussing online book reviews in terms of their value for identifying book characteristics or features.

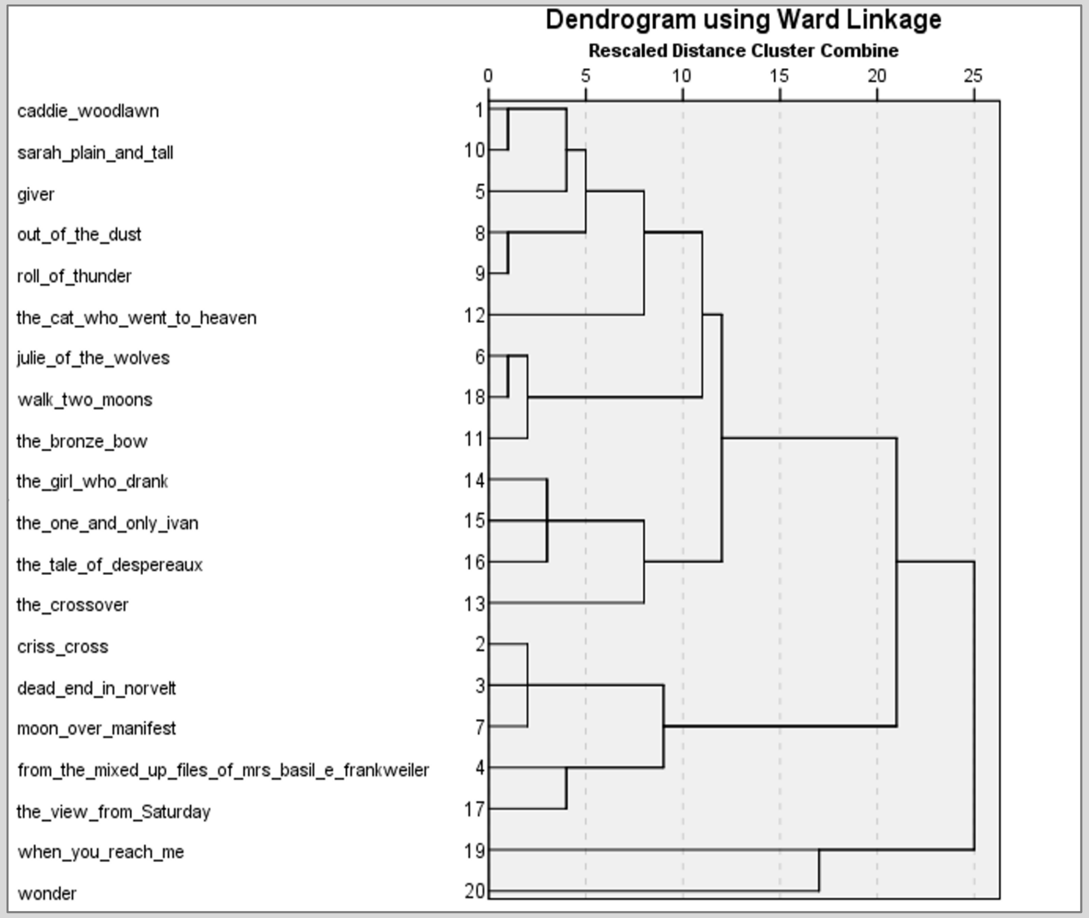
Meaningful similarities in Social Discussion

This study examined whether there are meaningful similarities in socially discussed online book reviews in terms of book features and characteristics. This study performed a hierarchical cluster analysis on the selected books to identify homogeneous clusters of cases (books) based on selected characteristics (frequencies of review words used in book reviews).

RQ 1. What homogenous clusters of books emerge based on the frequency of review words used in online book reviews?
RQ 2. Can the clustering of books be explained to identify book features with similar patterns?

The sample books were randomly collected from ALA Newbery Medal Winners list from 1922 to present. A total of 3,062 reviews (55,856 words) from the 20 sampled books.

Hierarchical Cluster Analysis



The dendrogram shows how far or close the distance is between books clustered using a 0 to 25 scale along the top of the chart. The smaller the distances before two clusters are joined, the smaller the differences in these clusters. For example, it is noticeable that three clusters, such as clusters 2 ("Criss Cross"), 3 ("Dead End in Norvelt"), and 7 ("Moon over Manifest"), are very similar. These three books show a similar pattern of word frequency in the Award category. Additionally, the dendrogram shows that three clusters, such as clusters 14 ("The Girl who drank"), 15 ("The one and only Ivan"), and 16 ("The Tale of Despereaux"), are more similar to each other than they are to others. It demonstrates that regarding the categories of Award, Audience, Emotion, and Evaluation, all those three books show very similar word frequency.

Frequency Pattern per Category

We discovered an interesting fact that the word frequency on the "award notes" tends to be associated with users' strong opinions about a book that can be either positive or negative. For example, when users highly recommend the book, they address "the award" in their reviews. On the contrary, when users have strong opposite preferences for the books, they also comment on "the award." The examples of these re-views are:

A review of the book titled "Moon over Manifest":
"A charming book, and I can see why it won the Newbery. It has the feel of a classic, like something that I would have read (and that would have won the award) when I was a kid. The writing is gorgeous, and the characters are instantly real and appealing....."

A review of the book titled "Criss Cross":
"Well now I better understand the fuss about disappointing Newbery Medal winners. This book was boring, not memorable, and the characters did not interest me one bit..I was trying to figure out why it won the Newbery. .."

Implications of the Study

The results of this study provided practical insights into the intrinsic values of users' social discussion in identifying similarities among books.

Acknowledgments

The author is deeply grateful to Dr. Soohyung Joo, Associate Professor at the University of Kentucky, for his valuable contribution to data analysis.