Which Book Should I Pick?: Text Visualization Based on Readability and Genre

H. Kim1, J. W. Park2
1,2GSAIM, Chung-Ang University, Seoul, Rep. of Korea

Abstract - This paper proposes readability visualization, genre visualization, and combined visualization to provide unconventional information for book selection. Data visualization was initiated for the practical purpose of delivering information, as it efficiently links visual perception and data so that readers are able to instantly recognize patterns in overcrowded data. In this interdisciplinary research we used the strength of data visualization, and this paper suggests three possible textual visualizations of a book, which may help users to find a desirable book, with the use of intuitive information out of a large volume of book data.

Keywords: Data Visualization, Information Visualization, Information Aesthetics

1 Introduction

Today, internet bookstores sell almost every single book in the world, and they provide a convenient book-searching function using keyword entry. Moreover, they deliver purchased books to the customers’ home directly. Buying books over the internet has gradually become a part of our lives.

Internet bookstores provide information about books in order to help customers pick their desired books, information such as a cover image, table of contents, and book reviews, in addition to basic information: author, editor, number of pages, a size, ISBN, and price. Customers also get information from other customers’ opinions or statistical recommendations provided by internet bookstores, which are sometimes leading readers to follow other people’s preferences blindly. In other word, these opinions are not information centered on the individual user but on the provider. Ideally, to choose a book, customers should read one or more chapters in advance of purchase to decide if a book’s style and genre suit them and also to predict how much time and effort it will take from them to read it.

Many previous examples of text visualization including our recent projects[1][2] have generally focused on a book’s contents and on plot analysis [3][4][5] but these approaches hardly provide practical clues needed for a reader to pick the “right book” for them, out of the thousands of recommendations on an Internet book store’s web site.

Common readers’ questions are not academic but practical: How much time will I need to read this book? Is it easy enough for a third-grade kid? Is there a genre called philosophical thriller? Is Twilight a romance or fantasy? Where I can find a romantic philosophical science fiction like Bicentennial Man? This paper suggests that intuitive text visualization reflecting readability and a user-oriented customizable genre would complement available book information and support reader book selection. (Fig. 1)

Fig. 1. The images of text visualization with an algorithm suggested in this paper: from the left, Harry Potter and the Prisoner of Azkaban (3rd book), Harry Potter and the Goblet of Fire (4th book), both by J. K. Rowling; The Critique of Pure Reason (English edition) by Immanuel Kant; and The Ethics (English edition) by Baruch de Spinoza.

2 Readability Visualization

2.1 Readability

A book’s readability is one of the important considerations in book selection. “Readability” is the ease in which text can be read and understood, and it directly affects the difficulty of a book. The factors of readability measurement have been studied by various researchers [6][7][8][9][10][11][12][13][14] since Sherman L. A. (1888)[15] (Table 1). The length of a sentence, number of sentences per paragraph, and amount of characters per word; these factors are closely related to the physical print space and the eye movements readability issue of “return sweep to next line[16]”. 

Fig. 1. The images of text visualization with an algorithm suggested in this paper: from the left, Harry Potter and the Prisoner of Azkaban (3rd book), Harry Potter and the Goblet of Fire (4th book), both by J. K. Rowling; The Critique of Pure Reason (English edition) by Immanuel Kant; and The Ethics (English edition) by Baruch de Spinoza.
Table 1. Readability Factors

<table>
<thead>
<tr>
<th>Author</th>
<th>Factor Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitsfignon (1921)</td>
<td>syllables per word and length of a sentence</td>
</tr>
<tr>
<td>Lively &amp; Pressey (1923)</td>
<td>level of difficulty and frequency of words</td>
</tr>
<tr>
<td>Vogel &amp; Washburne (1928)</td>
<td>number of different words in a paragraph and number of prepositions</td>
</tr>
<tr>
<td>Gray &amp; Leary (1935)</td>
<td>grammatical complexity with various factors including average sentence length, percentage of easy words, and number of sentences per paragraph</td>
</tr>
<tr>
<td>Gunning Fog (1952)</td>
<td>number of complex words, length of a sentence, and percentage of three or more syllables words</td>
</tr>
<tr>
<td>Dale &amp; Chall(1948)</td>
<td>length of a sentence with word count</td>
</tr>
<tr>
<td>Flesch (1948)</td>
<td>total syllables in 100 words and average number of words in a sentence</td>
</tr>
<tr>
<td>Bornmuth (1975)</td>
<td>average characters in a word and average number of words in a paragraph</td>
</tr>
<tr>
<td>Harris &amp; Jacobson (1975)</td>
<td>percentage of difficult words and average sentence length with word count</td>
</tr>
</tbody>
</table>

2.2 Readability visualization with length factors

We visually simulated a human reading process using the length related readability factors. The longer word, sentence, and paragraph are less readable. Each character was converted to a visible point of which the brightness decreases gradually during reading process. If a character is punctuation, the brightness of the point slightly increases. As a result, a book with longer sentences/paragraphs presents a dark and static image, but a book with concise sentences/paragraphs shows bright and dynamic patterns (Table 2, Fig. 2).

Table 2. Algorithm 1: Readability Visualization

```plaintext
set the base hue, saturation, brightness (0.0, 0.0, Max)
while(not end of the book) {
    get a next paragraph
    while(not end of the paragraph) {
        get a next character
        if (the character is punctuation or comma)
            increase brightness (sentence formula)
        else {
            decrease brightness (paragraph formula)
        }
        if (three or more syllables) // optional
            decrease brightness (word formula)
        set a pixel on next location with assigned hue, saturation, brightness and draw
    }
    reset brightness to Max
}
```

'"So you did, old fellow!" said the others.

'We must burn the house down!' said the Rabbit's voice; and Alice called out as loud as she could, 'If you do, I'll set Dinah at you!'

Fig. 2. A sample readability visualization for Alice's Adventures in Wonderland by Lewis Carroll. To explain the algorithm, we accelerated the brightness transition ten times.

Fig. 3 shows a clear difference between two books. My Sister's Keeper has clear and concise sentence and paragraph style, as is often the case with most of the contemporary best sellers, but The Critique of Pure Reason is extremely hard to understand, even for well educated people not only because of the difficulty of contents but also because of the lengthy complex sentences and paragraphs. With this algorithm a reader can get an impression of overall readability instantly without reading the entire book. For example, the readability visualization of most children's books that have more dialogic contents will appear extremely bright and dynamic. Classic novels are darker than contemporary novels due to the lengthy writing style. Most of philosophical, scientific, or academic books are much darker than books of others genres. The visualization matches with common knowledge of readability and comprehensibility. One of the rare exceptions to this is poetry. For example an image of a poem by T. S. Elliot is bright but the words in the poem are highly connotative and implicative. Sometimes readability does not accompany comprehensibility.

Fig. 3. Result of the algorithm 1; left is My Sister's Keeper by Jodi Picoult, right image is The Critique of Pure Reason (English edition) by Immanuel Kant

3 Genre Visualization

The readability visualization of previous chapter is helpful but an even more important function for book selection would be to visualize a book based on each individual’s preference. This chapter suggests an intuitive
visualization algorithm that visualizes a book based on genre; either conventional or custom genre.

The most of previous text clustering studies focused on finding dimensional closeness with various method such as digital signal processing with wavelet transform [17][18][19][20]. In this paper we used text analytic method with viewpoint of normal customers.

Usually genres are defined by a librarian with a historical categorization method or by a publisher with market preferences, but this algorithm provides unconventional categorization with a user’s definition. The following is a brief explanation of each step:

a) Analyze as many books as possible to count the summed frequency of each word and to build an “Overall Word-Frequency Dictionary (OWFD)”-We processed four-thousand randomly selected digital books in our college academic library-.

b) Select books that represent a specific genre and make a “Genre Word-Frequency Dictionary (GWFD)”,

c) Compare the “Genre Word-Frequency Dictionary (GWFD)” with the Overall Word-Frequency Dictionary (OWFD) and find extraordinary words of more frequency in the Genre Word-Frequency Dictionary (GWFD), and with the selected extraordinary words make a Genre-Identity Dictionary (GID).

TheGenre-Identity Dictionary will be used for a similarity indicator on each genre, including a ‘user-defined unconventional custom genre’. What follows is a detailed explanation of this process.

3.1 Overall Word-Frequency Dictionary (OWFD)

Table 3. Overall Word Frequency Dictionary(OWFD)

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>11,583,621</td>
<td>1</td>
</tr>
<tr>
<td>and</td>
<td>5,885,183</td>
<td>2</td>
</tr>
<tr>
<td>to</td>
<td>5,320,279</td>
<td>3</td>
</tr>
<tr>
<td>of</td>
<td>4,923,773</td>
<td>4</td>
</tr>
<tr>
<td>a</td>
<td>4,669,669</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zombie &amp; leet</td>
<td>18</td>
<td>98,770</td>
</tr>
<tr>
<td>naf</td>
<td>18</td>
<td>98,771</td>
</tr>
<tr>
<td>zwn</td>
<td>18</td>
<td>98,772</td>
</tr>
</tbody>
</table>

We collected four-thousand digital books to make one OWFD (Overall Word-Frequency Dictionary). The dictionary contains frequency and rank of each word in all of the books in our sample. More than half a million words are in this dictionary, but the frequency distribution is not linear, so the 20% of high-ranked words covers more than 98% of the whole word-frequency. The lowest-ranked word group, which makes the database enormously inflated is filled with invented onomatopoeia, mimetic words, proper nouns (e.g. character names), and misspelled words. To reduce processing cost we selected the top 20% words (10,000 words) to build the final OWFD (Table 3).

3.2 Genre Word-Frequency Dictionary (GWFD)

We made a GWFD (Genre Word-Frequency Dictionary) the same way we made OWFD, but we only sampled books that are representative of each genre not whole collection. We chose some of conventional genres, which are fantasy, philosophy, and science fiction, and we also created a user-defined custom genre. To represent each genre, it is important to select the most suitable representatives of each genre; we chose several important books for each (Table 4).

Table 4. Selected genres and representative books

<table>
<thead>
<tr>
<th>Genre</th>
<th>Representing books</th>
</tr>
</thead>
</table>
| Fantasy | - The Lord of the Rings, by J. R. R. Tolkien  
- The Lion, the Witch and the Wardrobe, by C. S. Lewis  
- Harry Potter and the Sorcerer’s Stone, by J. K. Rowling |
| Philosophy | - The Apology, Crito and Phaedo of Socrates by Plato  
- Critique of Pure Reason by Immanuel Kant  
- The Ethics by Benedict de Spinoza |
| Science Fiction | - 2001: A Space Odyssey by Arthur C. Clarke  
- Childhood’s End by Arthur C. Clarke  
- Double Star by Robert A. Heinlein  
- Starship Troopers by Robert A. Heinlein  
- The Currents of Space by Isaac Asimov  
- The Naked Sun by Isaac Asimov |
| Custom Genre | - Jane Eyre by Charlotte Bronte  
- Sense and Sensibility by Jane Austen  
- My Sister’s Keeper by Jodi Picoult |

For the fantasy genre, we selected two books from two classic masters, Tolkien and Lewis, in addition to the first volume of the Harry Potter books, which is a best seller in the contemporary fantasy novel genre. For the philosophy genre, we selected three books: by Plato, Kant and Spinoza; and for the science fiction genre, we chose six books by the so called big three writers (Clarke, Heinlein, Asimov). The custom genre is a user-defined genre based on one user’s preference, who is one of co-authors of this paper. We named the genre ‘Kimyo’s selection’ after her nickname. She loved Jane Eyre, Sense and Sensibility, and My Sister’s Keeper. These books show women writers’ style, and they also speak out about social issues such as social class problems, child abuse,
women’s rights, and medical self-determination. These books are romantic and philosophical novels that cannot be described as a single conventional genre.

Table 5. Genre Word Frequency Dictionary(GWFD)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Fantasy</th>
<th>Philosophy</th>
<th>Science Fiction</th>
<th>Kimyo’s Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the 17,529</td>
<td>the 23,407</td>
<td>the 21,024</td>
<td>the 14,587</td>
</tr>
<tr>
<td>2</td>
<td>and 11,163</td>
<td>of 15,009</td>
<td>to 10,193</td>
<td>to 10,886</td>
</tr>
<tr>
<td>3</td>
<td>of 7,202</td>
<td>to 10,544</td>
<td>of 10,032</td>
<td>1 10,791</td>
</tr>
<tr>
<td>4</td>
<td>to 8,573</td>
<td>and 8,829</td>
<td>and 8,632</td>
<td>and 10,332</td>
</tr>
<tr>
<td>5</td>
<td>a 6,243</td>
<td>in 8,842</td>
<td>a 9,472</td>
<td>of 8,486</td>
</tr>
<tr>
<td>6</td>
<td>be 5,078</td>
<td>is 8,097</td>
<td>1 7,102</td>
<td>a 7,915</td>
</tr>
</tbody>
</table>

As we can see in Table 5, the words ranked by frequency are similar between the genre word databases because the basic and required words to build a sentence are same in each. Therefore it is not easy to find a difference with these databases.

3.3 Genre Identity Dictionary(GID)

The GWFD (Genre Word-Frequency Dictionary) has to be refined to represent a genre identity. We suggest an algorithm that compares between an OWFD (Overall Word-Frequency Dictionary) and a GWFD (Genre Word-Frequency Dictionary to extract extraordinary words for each genre.

We measured a rank difference of a word that occurs in both between two dictionaries. If a word had a higher rank in the GWFD than OWFD, we recorded the word into a database with a rank distance.

Fig. 4. The higher ranked words in the GWFD (Genre Word-Frequency Dictionary of fantasy genre); they form the Genre-Identity Dictionary

As Fig. 4 shows, there are words that are frequently found in a certain genres (e.g. wizard and wand in a fantasy genre). We call these words “genre-identity words” and the database a “Genre-Identity Dictionary (GID)”.

Table 6 shows genre-identity words in the GIDs (Genre-Identity Dictionaries) sorted by rank distance. Each genre has a special set of extraordinary words. Some of the words in the highest ranked group are proper nouns such as characters’ names -except philosophy genre- but very soon we can find the unique genre-oriented words; wizard, journey (fantasy); bugs, discovery (science fiction); behavior, obliged, engagement (Kimyo’s selection).

Table 6. Genre Identity Dictionaries(GIDs)

3.4 Genre closeness calculation

We can calculate genre closeness of a book with GIDs (Genre-Identity Dictionaries). First, we made a frequency-ranked word dictionary of the subject book that a customer wants to test, and we compared it against various GIDs. If a word in the dictionary of subject book is also found in a GID, we call it a “word hit”. We also calculated the average rank distance of each paired word, and we named it “Word Average Distance”. Naturally, more word hit and less word average distance implies that the subject book is closer to the genre.

Table 7 shows the results of the genre closeness between seven exemplary books and four predefined genres, including a custom genre ‘Kimyo’s selection’. Some of the subject books are not easy to describe as a single conventional genre. Alice's Adventures in Wonderland (by Lewis Carroll) is a children’s book, but it also has profound symbolic messages. Solaris by (Stanislaw Lem) appears as clearly science fiction genre, but the book’s message is philosophic; man’s anthropomorphic limitations. In Jodi Picoult’s first bestseller, Nineteen Minutes, she mixed family, morality, and many social controversies into a complex and twisted plot. We also tested Harry Potter and the Prisoner of Azkaban (by J.K. Rowling), The Analysis of Mind(by Bertrand Russell), An Inquiry into the Nature (by Adam Smith), and Little Women (by Louisa May Alcott).
Table 7. Genre Closeness for Subject Books

<table>
<thead>
<tr>
<th>Book</th>
<th>Fantasy</th>
<th>Philosophy</th>
<th>SF</th>
<th>Kimyo's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice's Adventures in Wonderland</td>
<td>274**</td>
<td>188</td>
<td>108</td>
<td>229**</td>
</tr>
<tr>
<td>Solaris</td>
<td>127</td>
<td>374**</td>
<td>307*</td>
<td>248</td>
</tr>
<tr>
<td>Harry Potter 3rd Book</td>
<td>350**</td>
<td>128</td>
<td>120</td>
<td>268**</td>
</tr>
<tr>
<td>Nineteen Minutes</td>
<td>86</td>
<td>145</td>
<td>115</td>
<td>270**</td>
</tr>
<tr>
<td>The Analysis of Mind</td>
<td>82</td>
<td>880**</td>
<td>355</td>
<td>467</td>
</tr>
<tr>
<td>Inquiry Into the Nature</td>
<td>125</td>
<td>687***</td>
<td>288</td>
<td>606**</td>
</tr>
<tr>
<td>Little Women</td>
<td>190</td>
<td>267</td>
<td>133</td>
<td>525**</td>
</tr>
</tbody>
</table>

The genre closeness of selected books is not different from our expectation. A primary genre, the closest genre of each subject book, fits into the conventional genre classification of Internet bookstores, but what we have to pay attention to is the multi-genre closeness of mixed-genre books. Solaris shows closeness in two genres; science fiction and philosophy, as we suspected earlier. Alice's Adventures in Wonderland shows complex characteristic of fantasy (word hit), philosophy (word average distance), and the custom genre (word hit). Although The Analysis of Mind and An inquiry into the Nature both are conventionally classified as philosophical publications, only An Inquiry into the Nature shows closeness in the custom genre Kimyo’s selection, which is related to social issues.

3.5 Genre closeness visualization

The result of genre closeness is accurate and clear, but showing many numbers, as seen Table 7, is not perceivable or practical. Moreover by increasing the number of subject books or predefined genres, the difficulty of similarity detection will be intensified. To solve this problem we suggest data visualization, which helps to find patterns intuitively based on genre closeness (word hit and word closeness), and we call the image “Figure of Genre Closeness”. (Table 8, Fig. 5)

Table 8. Algorithm 2: Figure of Genre Closeness

place a ‘word-frequency dictionary’ of subject book left side (sorted with frequency)
place a GID of a genre right side (sorted with rank distance) while(not end of ‘word-frequency dictionary’ of subject book)
get a next word from ‘word-frequency dictionary’ of subject book (left side)
find the word from opposite GID (right side) calculate the distance between two words set the thickness of a line and transparency based on the distance (closer is thicker and brighter) draw a line between two point based on the thickness and transparency

Fig. 5. Figure of genre closeness with a subject book, Harry Potter 3rd book (Harry Potter and the Prisoner of Azkaban) and a fantasy-genre-identity dictionary by connecting two same words in both dictionaries; the line thickness and transparency (alpha channel) are proportional to distance of each line.

Fig. 6 shows results of this visualization. The most complex and bright image implies the closest genre of the subject book, so users are able to instantly describe the genre closeness; The Analysis of Mind as philosophy, Harry Potter and the Prisoner of Azkaban as fantasy, and Little Women as a custom genre. Moreover, this visualization presents mixed multi-genre closeness as well as the primary genre.
4 Combined Visualization of Readability and Genre

Previously we suggested readability visualization (Algorithm 1) and genre visualization (Algorithm 2). In this chapter, we will propose a practical visualization that combines the information of ‘Readability Visualization’ chapter and ‘Genre Visualization’ chapter, as each has own strengths. The suggested steps are as follows:

a) Define genre-hue based on color symbols,

b) Set the genre-hue as base hue in the readability visualization with hue rotation.

There is no right answer for the color-genre connection. We defined genre-hue based on the Table 9 below; based on symbol of color. This definition is temporary, to show the visual clearness of the algorithm.

Table 9. Defined Genre Hue(bold words are closely related to each genre)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Color</th>
<th>Representing books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fantasy</td>
<td>Red</td>
<td>challenge, passionate, age, active, exciting, dangerous, courage, happiness, love, loyalty, delight, power, sin, passion, vivid</td>
</tr>
<tr>
<td>Philosophy</td>
<td>Blue</td>
<td>dignity, sorrow, truth, harmony, happiness, sincerity, holy, intelligence, spirit, cold, stability, calm, hope, cold, hearted, melancholy, intelligent</td>
</tr>
<tr>
<td>Science Fiction</td>
<td>Purple</td>
<td>victory, modesty, tragedy, mystery, sacred, Exotic, splendid, sorrow, dignity, complicated, nice, noble</td>
</tr>
<tr>
<td>Kinyo’s Selection</td>
<td>Green</td>
<td>relaxation, happiness, beauty, hope, nature, safe, peace, calmness, Mild, longing, cozy, pleasant, abundant</td>
</tr>
</tbody>
</table>

Table 10 (Algorithm 3) is mostly the same as the previously suggested readability algorithm (Algorithm 1) except for the addition of a base genre-hue and a hue rotation. The base genre-hue sets a color tone of book visualization, and the hue rotation makes rhythmical color patterns, which emphasize book readability and genre with juxtaposition of complimentary color shadow. Additionally, saturation is decided by the length of paragraph; it also shows paragraph readability. The results of Algorithm 3 are shown in Fig. 7.

5 Conclusion and Future Work

This alternative method of book selection has the strength of conventional data visualization that finds patterns instantly over huge data with human perception. We tested our visualization and the results prove the effectiveness of the suggested algorithm. All testers (n=20) easily categorized the same genre based on hue. The subjects were also asked to evaluate readability visualization on 5-point scale (1=very easy; 5=very hard) of 10 books, and the result shows that most of them predicted readability correctly (r = 0.79, p<0.01), so we believe this research will help readers to choose a right book.

With customers’ point of view this visualization could be very useful in the current online bookstore interface by showing essential information of the custom favorite genre and readability as the figure shows Fig. 8.
We only suggested one example of possible combined visualizations in this paper, but with readability and genre data there will be more visualization methods and interfaces, especially to support custom-genre applications. Those series of research will be our next task.

6 Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government MEST, Basic Research Promotion Fund (NRF-2010-013-H00001).

7 References


[15] L. A. Sherman, " Some observations upon the sentence-length in English prose", University studies (University of Nebraska (Lincoln campus)), Vol. 1, No. 2, pp. 119-130, 1888.


