http://dx.doi.org/10.5392/JKCA.2012.12.05.052

장르 판별 알고리즘을 이용한 책 장르 시각화

Book Genre Visualization based on Genre Identification Algorithm

김효영, 박진완
중앙대학교 첨단영상대학원

Hyoyoung Kim(greatinno@naver.com), Jin Wan Park(jinpark@cau.ac.kr)

요약

텍스트 시각화는 데이터 시각화의 한 분야로, 방대한 텍스트 데이터에 대한 다양한 분석 기법을 바탕으로 텍스트의 내용적 측면은 물론 구조적, 형식적 측면을 시각적으로 재현(represent)해내는 방법에 관한 연구이다. 본 연구에서는 이러한 텍스트 시각화 연구의 일환으로, 서적이 갖는 장르적 특성을 서적 본문에 직접 사용된 단어들을 바탕으로 파악해낼 수 있는 방법에 대해 고찰하고, 실험을 통한 검증을 바탕으로 서적 장르 시각화의 요소를 도출한 후, 이를 직관적이고 효율적으로 시각화하는 방법에 대해 서술하였다. 본 연구에서 제안하는 시각화는 첫째, 책에 직접 사용된 단어를 토대로 책의 실질적 장르를 파악할 수 있으며, 둘째, 시각화 결과 이미지를 통해 해당 서적이 어떤 장르와 가장 가까운지 한 눈에 파악할 수 있을 뿐 아니라, 한 책이 갖는 복합 장르적 특성을 알 수 있도록 해주고, 이미지 내의 점(dot)의 개수와 곡선의 곡률, 밝기 등을 통해 대표 장르로 파악된 장르의 근접도(유사도)를 짐작할 수 있다는 점에서 그 의의를 갖는다. 나아가 개별 소비자 자신이 선호하는 서적들에 대한 적용을 통해 개인별 선호 서적(또는 장르) 이미지를 제공하는 등 서적 추천 시스템과 같은 북 커스터마이징(book customizing)과 같은 분야에도 다양하게 활용될 수 있다.

■ 중심어 : | 텍스트 시각화 | 데이터 시각화 |

Abstract

Text visualization is one of sectors in data visualization. This study is on methods to visually represent text's contents, structure, and form aspects based on various analytic techniques about wide range of text data. In this study -as a text visualization study-, 1) a method to find out the characteristics of a book's genre using words in the text of the book was looked into, 2) elements of visualization of a book's genre based on verification through an experiment were drew, and 3) the ways to intuitionally and efficiently visualize this were explained. According to visualization suggested by this study, first, actual genre of a book can be understood based on words used in the book. Second, with which genre is closed to the book can be found out with one glance through images of visualization. Moreover, the characteristics of complicated genres included in a book can be understood. Furthermore, the level of closeness (similarity) of a genre -which is found to be a representative genre using the number of dots, curvature of a curve, and brightness in the image- can be assumed. Finally, the outcome of this study can be used for a variety of fields including book customizing service such as a book recommendation system that provides images of personal preference books or genres through application of books favored by individual customers.

■ keyword : | Text Visualization | Data Visualization |

* 본 연구는 한국연구재단의 지원을 받아 수행되었음(No. 2011-0018616).
접수번호 : #120221-001
접수일자 : 2012년 02월 21일
심사완료일 : 2012년 04월 26일
교신저자 : 박진완, e-mail : jinpark@cau.ac.kr

I. 서 론

1. 연구 목적

텍스트 시각화는 데이터 시각화의 한 분야로, 방대한 텍스트 데이터에 대한 다양한 분석 기법을 바탕으로 텍스트의 내용적 측면은 물론 구조적, 형식적 측면을 시각적으로 재현(represent)해내는 방법에 관한 연구이다[1]. 텍스트 시각화에 대한 연구는 텍스트의 내용을 다양한 시각 요소로 표현하는 기본적인 접근부터, 텍스트의 내용 또는 그 안에 숨어있는 스토리텔링을 새로운 관점으로 재조명 하거나, 보이지 않는 관계적 측면을 시각적 재현을 통해 드러내는 등의 다양한 접근 방식을 갖는다.

텍스트 시각화의 재료 데이터가 되는 텍스트의 경우, 그 양이 방대해질 경우 전체적인 주제와 내용 및 그 데이터가 갖는 관계 등을 파악하기가 매우 어렵게 된다[2]. 이러한 맥락에서 텍스트 데이터를 분석하여, 시각적으로 표현하고자 하는 요소를 도출한 뒤 이를 효과적인 시각적 요소로 매핑하여 하나의 이미지의 형태로 표현하고자 하는 텍스트 시각화에 관한 연구는, 방대한 데이터에서 파악하기 불가능한 복잡, 다양한 정보를 직관적으로 나타낼 수 있다는 점에서 정보 전달의 독창성 및 효율성을 갖는다. 본 연구에서는 이러한 텍스트 시각화 연구의 일환으로, 다양한 서적이 갖는 장르적 특성을 텍스트 데이터에 대한 분석을 토대로 도출한 후 이를 직관적인 하나의 이미지 포맷으로 표현하기 위한 방법론을 제시하고자 하였다. 이를 위해 서적의 장르적 성격을 나타내는 요인 도출을 위한 디지털 서적 데이터의 분석 및 처리 절차에 관한 연구와, 도출된 요인을 시각적 요소로 치환하는 방법에 대한 미학적, 디자인적 접근에 따른 구체적 내용을 기술한다.

2. 연구 방법

책의 장르는 대부분의 경우 출판사나 저자에 의해 분류되는데, 이는 주관적인 것으로 실제 책의 텍스트가 갖는 성격과는 다소 차이가 있을 수 있다. 본 연구에서는 이러한 점에 착안하여 서적 텍스트에 사용된 단어 데이터를 분석하여 서적의 장르를 판별할 수 있는 방법론에 대하여 기술하고, 이러한 방법론의 타당성을 검증하며, 이를 통하여 도출된 각 서적 텍스트의 장르 정보를 시각적 요소로 매핑하여 한 장의 직관적인 이미지의 형태로 시각화하는 방법을 제안한다.

이를 위해 먼저 최대한 많은 디지털 서적 데이터를 수집하여 분석한 후, 이들 서적에 사용된 단어로 '보편빈도사전'을 제작한다. 다음으로 각 장르별로 대표 서적을 선정하여 이들을 장르 대표 서적으로 할당하고, 이들 대표 서적에서 사용된 단어들을 빈도수로 정리한 '장르빈도사전'을 제작한다. 그리고 이들 '장르빈도사전'에 등장하는 단어 중, 특정 장르의 성격을 갖기 보다는 보편적으로 많이 사용되는 단어를 제외하고, 실제로 특정 장르에서의 출현 빈도가 높은 단어를 판별하기 위해, '보편빈도사전'과 각 '장르빈도사전'의 빈도수를 비교, 분석하는 과정을 거쳐 '장르독자성사전'을 제작한다. 이렇게 제작된 '장르독자성사전'을 바탕으로 임의의 책을 알고리즘에 대입하여 각 장르독자성사전과 비교하는 과정을 거쳐 사용 단어의 장르 근접도를 도출한다. 이러한 일련의 과정을 통해 도출된 장르 근접도를 하나의 이미지로서 표현하기 위해 각 장르의 근접도에 해당하는 속성을 시각적 요소와 매핑하여 하나의 직관적 이미지로 시각화하는 방법을 제안하고, 결과 이미지 분석을 통해 제안된 시각화 방법론의 타당성을 검증한다.

II. 텍스트 시각화 사례 연구

1. Visualizing "His Dark Materials"

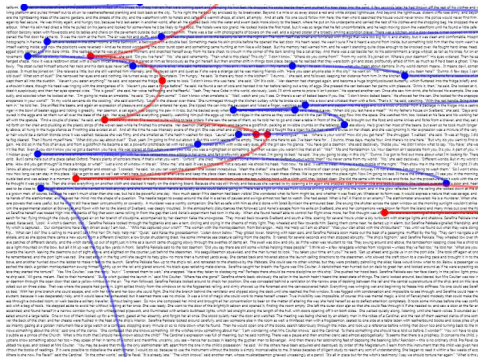


그림 1. "His Dark Materials"의 시각화 일부

T. Legan과 L. Becker(2010)[3]는 그들의 논문에서 Pullman의 3부작 소설인 "His Dark Material"을 시각화하는 방법론을 제안하였다.

[그림 1]과 같이, 세 권의 텍스트 전체를 펼쳐놓고 남녀 주인공 이름인 'Lyra'와 'Will'이라는 단어를 점으로 찍은 후 연결한 시각화 결과물을 통해 1권에서는 남자 주인공이 등장하지 않은 채 여자 주인공만으로 이야기가 전개된다는 내용의 흐름을 짐작할 수 있고, 두 주인공의 이름이 모두 존재하지 않는 공백 부분에서는 기타 등장인물(Mary Malone, Lord Asreil, Mrs. Coulter 등)에 관한 묘사가 주로 나타남을 알 수 있는 등 문학 작품의 시각화를 통해 그 속에 드러나는 이야기의 흐름 및 관계를 시각화한 연구이다.

2. BibleViz

BibleViz[4]는 Chris Harrison과 Christoph Romhild가 성경의 텍스트를 데이터로 여러 가지 정보를 시각화한 일련의 작업이다. 이 작업은 총 3종류의 시각화 작업으로 이루어져있으며, 이중 가장 대표적인 것이 성경 텍스트 내의 교차 언급을 시각화한 'Bible Cross-References'이다. 성경 내에 교차적으로 언급되는 63,000여 개의 교차 언급을 막대(bar)와 호(arc)를 이용하여 시각적으로 표현하였다[그림 2]. 방대한 데이터를 인터랙션이 없는 하나의 이미지 형태로 시각화하기 위해 정보 전달이라는 기능적 측면보다는 심미적 측면에 중점을 둔 작업으로 볼 수 있다.

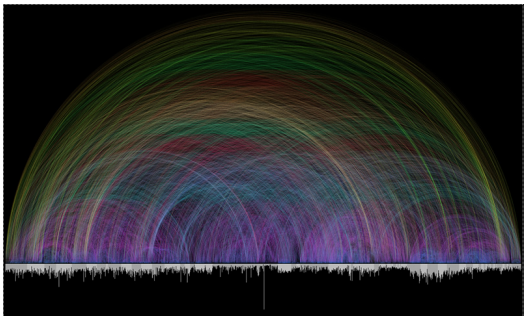


그림 2. Bible Cross-References

[그림 2]의 아래쪽에 나열된 막대그래프(bar graph)는 성경의 각 장(chapter)을 나타낸다. 성경의 각 권(book)은 흰색과 밝은 회색으로 번갈아가며 구분하여 표현하였고, 각 장에 있는 절(verse)의 수가 많을수록 막대의 길이가 길어진다. 총 63,779개의 교차 언급은 하나의 호(arc)의 형태로 나타내었고, 교차 언급된 두 개의 각 장(chapter)의 거리에 따라 색상을 다르게 입혀 마치 무지개와 같은 시각적 효과를 갖도록 하였다.

3. Writing without Words

Writing without Words[5]는 텍스트를 시각적으로 재현하는 방법에 대해 연구하고 다양한 작가들의 글쓰기 스타일 특성과 그 차이점을 시각화하고자 하는 프로젝트이다. 이 프로젝트에서는 소설이 갖는 텍스트 데이터를 '문학적 유기체(The Literary Organism)'로서 표현하여 개별 소설 고유의 문학적, 시각적 정체성을 나타내고자 하였다. 이와 함께 문장과 문장 길이의 시각화, 운율의 질감 시각화 등의 일련의 시각화 과정을 아우르는 매우 흥미로운 작업이다.

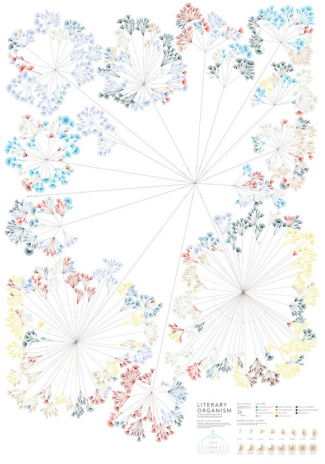


그림 3. Stephanie Posavec의 '문학적 유기체' (The Literary Organism) – Jack Kerouac의 소설 'On the Road'

[그림 3]은 Jack Kerouac의 소설 'On the Road'를 유기체적으로 시각화한 이미지이다. 단순한 트리구조로 표현하면서, 그 유기체적 느낌을 잘 살리기 위해 수작

업으로 제작하였다. 'On the Road'의 첫 번째 파트를 장(chapter)으로 분리하고, 각 장은 절(paragraph)로, 절은 문장으로, 문장은 단어들로 분리한 후, 각 문장에 따른 단어의 수에 따라 조직화하여 표현하였다. 여기에 11가지의 주제별 색상을 할당하여 적용함으로써 각 시각화 결과 이미지가 임의의 소설에 대한 주제와 내용을 시각적으로 나타내도록 하였다. 이에 따라 여러 문학 작품에 대한 주제와 성격 및, 작가에 따른 문장과 문체의 차이점 등을 느낄 수 있도록 했다는 점에서 그 의의를 갖는다.

4. TextArc

TextArc[6]는 책의 텍스트를 추상적으로 표현한 Brad Paley의 시각화 작품이다. 이 시각화 결과물은 컴퓨터 화면 상에서 두 번에 걸쳐 나타나게 되는데, 먼저 거대한 타원의 가장자리에 매우 작은 폰트로 이루어진 선들이 나타나고, 그 후에는 보이지 않는 스프링으로 연결된 단어들이 나열된다. 책에 많이 등장하는 단어는 크게 표현이 되고, 드물게 사용되는 단어들은 같은 공간을 공유하여 나열된다.

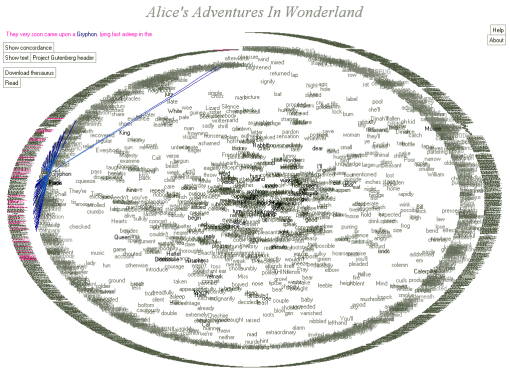


그림 4. Brad Paley의 TextArc – 'Alice's Adventure in Wonderland'

[그림 4]는 판타지의 고전인 '이상한 나라의 앨리스(Alice's Adventure in Wonderland)'를 TextArc로 시각화한 것이다. 이미지를 보면 'Alice'라는 단어가 화면 중앙에 큰 폰트로 자리잡고 있다. 이는 'Alice'는 주인공의 이름으로 책 전반에 걸쳐 계속하여 등장하기 때문이다. 또한 책 속에 등장하는 다양한 캐릭터들은 위치와 색깔을 달리하며 각각의 특성과 관계들을 시각적으로 표출해낸다. 예를 들면 해터(Hatter)와 도어마우스(Dormouse), 3월의 토끼(March Hare)는 비슷한 글자체로 가까이 위치하고 있어 주로 함께 등장함을 짐작할 수 있다. 본 시각화 작품을 통해 비록 소설의 내용을 데이터로 사용하였지만, 수학자이자 논리학자인 원작자 Lewis Carroll이 숨겨둔 치밀한 논리들이 시각화를 통해 재탄생될 수 있음을 알 수 있다.

Ⅲ. 사용 단어 기반 장르 판별 알고리즘 설계

1. 개념 및 절차

보통 책의 장르는 저자나 출판사의 임의대로 시장 상황에 따라 구분된다. 따라서 독자가 직접 책을 읽고 느끼게 되는 책의 장르와 상이한 경우가 발생하게 마련이다. 이를 위해 본 논문에서는 일반적인 장르 분류 기준으로 구분되는 특정 장르들의 고유의 단어 사전을 만들어 임의의 책을 이들 각각의 장르 사전들과 비교하는 과정을 거쳐 책의 단어 사용을 바탕으로 한 실질적 장르를 판별해 내는 것이다. 이는 구체적으로 다음과 같은 과정을 거친다.

1) 충분히 많은 책을 분석하여(본 연구에서는 총 4천권의 책을 분석하였다) 평균적인 단어의 빈도수를 측정하여 '보편빈도사전'을 만든다.
2) 각 장르에 해당하는 대표 서적을 골라 이 책들에 있는 단어의 빈도수를 바탕으로 각각의 '장르빈도사전'을 만든다.
3) 보편빈도사전과 장르빈도사전의 단어의 빈도를 비교하여 장르빈도사전에서 나타나는 특이 단어, 즉 특정 장리에 보다 빈번히 쓰이는 단어를 바탕으로 '장르독자성사전'을 만든다.
4) 이렇게 만들어진 장르독자성사전을 기준으로 판별을 위해 입력된 책에 사용된 단어들과의 유사도를 분석하여 각 장르와 입력 책과의 근접도(closeness)를 파악한다.

위의 장르 판별 과정 각각에 대한 세부 프로세스를 살펴보도록 하자.

2. 보편빈도사전 제작

대부분의 서적에서 보편적으로 많이 사용되는 단어를 빈도순으로 정리하기 위해 다양한 장르로 이루어진 4,000권의 디지털 서적을 데이터 파일 형태로 수집하였다. 본 연구에서는 데이터 수집의 용이성을 위해 영문 서적 데이터를 대상으로 하였다. 이렇게 수집된 책들에서 사용된 단어들을 추출해낸 데이터를 바탕으로 '보편빈도사전'을 제작하였다. 이를 위해 Java 기반 언어인 processing을 이용하여 프로그램을 개발하였고, 전형적인 string operation과 token 알고리즘, 그리고 binary search 알고리즘, quick sort 알고리즘 등 데이터베이스 구성을 위한 기능을 할당하였다. 이 사전은 단어 빈도(frequency)와 그에 따른 순위(rank)를 보관한다.

분석 결과 4,000권의 책 속에는 총 약 50만 개의 단어가 등장하는데, 이를 빈도수로 나열한 단어 순위의 상위 20%를 차지하는 약 10만개의 단어가 전체 사용 단어의 98% 이상의 비율을 차지하고 있었다. 특히 1-2회 정도만 등장하는 단어가 많은데 이들은 보통 의성어, 의태어, 고유명사 혹은 오타에 기인한 것이다. 데이터베이스의 총량에 따라 본 사전의 제작기간이 크게 늘어나고, 빈도 수 하위 2%의 단어는 큰 의미가 없으므로 이들을 제외하고, 상위 20%에 속하는 10만 단어를 바탕으로 보편빈도사전을 제작하였다. [표 1]는 보편빈도사전에 포함된 단어들을 빈도수 순으로 나열한 것이다.

표 1. 보편빈도사전 내부 구조

단어	빈도	순위
the	11583621	1
and	5586383	2
to	5320279	3
of	4923275	4
a	4669669	5
i	3328795	6
he	3046996	7
in	2942094	8
.....		
zhark	18	98765
zina	18	98766
zinnia	18	98767
zinzu	18	98768
zog	18	98769
zombielike	18	98770
zulfi	18	98771
zw	18	98772

3. 장르빈도사전 제작

앞서 제작한 보편빈도사전이 대부분의 서적에서 공통적으로 사용되는 단어를 정리한 것이라면, 다음으로는 각 장르별로 많이 쓰이는 단어를 정리한 '장르빈도사전'을 제작하는 단계가 필요하다. 그러나 서적의 장르는 매우 다양하여 모든 장르에 대한 정의 및 구분은 쉽지 않다. 본 연구는 단어로 장르의 특성을 파악하는 방법론적 기초 연구이므로 연구를 위한 실험 및 검증의 편의를 위해 보편적 서적 분류체계로 구분되고 있는 장르 중 4가지를 선정하여 실험을 진행하고자 하였다. 서적 장르 구분의 기준으로서, 전 세계적으로 널리 이용되고 있는 인터넷 서점인 아마존닷컴(amzon.com)[7]의 서적 카테고리의 대분류 - 중분류 카테고리를 참고로 하였고, 결과적으로 선정된 4가지 장르는 각각 '판타지(Fantasy)', '철학(Philosophy)', 'S. F(Science Fiction)', '여성소설(Women's Fiction)'이다[표 2].

표 2. 선정된 네 장르의 대분류-중분류 카테고리(출처-아마존닷컴)

대분류	중분류
Science Fiction & Fantasy	Fantasy
	Science Fiction
Politics & Social Sciences	Philosophy
Literature & Fiction	Women's Fiction

다음으로 위의 네 가지 장르 각각을 대표할 수 있는 대표 서적을 선별하여 각 장르별 대표 서적들에서 사용된 단어들의 출현 빈도수를 바탕으로 '장르빈도사전'을 제작한다. 네 가지 장르의 대표서적으로, 먼저 판타지 장르의 경우 과거 판타지 소설의 양대 산맥인 J. R. R.

Tolkien과 C. S. Lewis의 대표 작품인 '반지의 제왕'과 '나니아 연대기'와 함께, 현대 마법 판타지의 베스트셀러인 '해리포터와 마법사의 돌(1권)'을 추가로 선정하였고, 철학 장르의 경우 철학서의 고전인 플라톤의 저서와 칸트, 스피노자의 대표 저서를 선택하였다. 또한 여성소설 장르의 대표작으로 대표적 현대 여류 작가인 Jordi Picoult의 대표적 여성 소설로 알려진 'House Rules'와 함께 고전 여성소설의 대명사인 Jane Austin의 '오만과 편견(Pride and Prejudice)'과 'Sense and Sensibility)'를 선정하였고, S. F 장르의 대표 서적으로는 S. F 소설의 아버지로 불리우는 대표적 세 저자인 Arthur C. Clarke와 Robert A. Heinlein, 그리고 Isaac Asimov의 대표작을 각각 1-2권 선정하였다. 선정된 각 장르별 대표 서적은 [표 3]과 같다.

표 3. 각 장르별 대표 책 선정

장르	선정된 각 장르별 대표 서적
판타지 (Fantasy)	Lord of The Ring / by J.R.R. Tolkien
	Narnia – The Lion, the Witch and the Wardrobe / C. S. Lewis
	Harry Potter and the Sorcerer's Stone / J.K. Rowling
철학 (Philosophy)	Apology, Crito and Phaedo / Plato
	Critique of Pure Reason / Immanuel Kant
	The Ethics / Benedict de Spinoza
여성소설 (Women's Fiction)	Pride & Prejudice / Jane Austen
	Sense and Sensibility / Jane Austen
	House Rules / Jodi Picoult
S. F (Science Fiction)	2001 – A Space Odyssey / Arthur C. Clarke
	Childhood's End / Arthur C. Clarke
	Double Star / Robert A. Heinlein
	Starship Troopers / Robert A. Heinlein
	The Currents of Space / Isaac Asimov
	The Naked Sun / Isaac Asimov

4가지 장르빈도사전을 만든 결과, [표 4]와 같이 문장을 구성하는 필수적인 요소로서의 기본 단어(대명사, 전치사, 접속사 등)가 자연스럽게 높은 순위를 차지하고 있음을 알 수 있다. 따라서 이것만으로는 각 장르별 단어의 독자성을 판단하기에는 무리가 있기 때문에, 이러한 보편적 단어들을 제외한 장르 특유의 독자성을 강하게 띄는 단어들을 판별해내는 과정을 거쳐야 한다.

표 4. 장르별 단어 빈도 순위

순위	판타지		철학		여성소설		S. F	
1	the	17529	the	23407	the	14587	the	21024
2	and	11165	of	18009	to	10886	to	10193
3	of	7202	to	10544	i	10791	of	10032
4	to	6673	and	8859	and	10332	and	9822
5	a	6243	in	8842	of	8486	a	9472
6	he	5078	is	8097	a	7915	i	7102
7	in	4437	a	6965	her	5432	it	6997
8	it	4380	that	5919	in	5070	was	6784
9	was	4154	it	5133	you	4975	he	6632
10	i	3795	as	4953	it	4839	that	6108
...	.....							

4. 장르독자성사전 제작

앞서 제작한 장르빈도사전에서 보편 단어를 제외한 각 장르적 특성을 갖는 단어를 빈도순으로 간추려 '장르독자성사전'을 제작한다. 각 장르빈도사전과 보편빈도사전의 데이터베이스를 비교하여 특정 장르 사전에서만 등장하는 특이 단어를 골라내어 장르만의 독자성을 보여주는 단어를 추출한다. 이는 전체 사전에 비해 장르 사전에서 현격히 순위가 올라가는 단어, 즉 빈도순위 상승의 거리 값(distance)으로 순위를 정한 단어들의 집합이 된다. [그림 5]와 같이, 보편빈도사전에서 하위 빈도수로 랭크되었던 'wizard'나 'wand'같은 단어가 판타지 장르빈도사전에서는 높은 순위에 자리 잡고 있다. 이와 같은 경우, 해당 단어는 상승분(두 사전 비교 순위가 상승된 정도)에 따라 장르독자성사전에 순차적으로 기록된다.

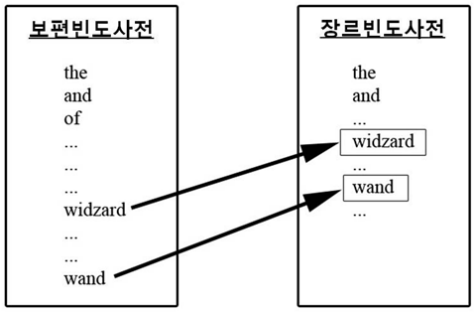


그림 5. 보편빈도사전과 장르빈도사전의 단어 순위 비교 과정

[표 5]는 각 장르별로 각각의 특이 단어를 모은 장르독자성사전의 일부이다. 각 장르독자성사전을 살펴보면 각 장르별로 각각의 특이 단어가 눈에 띈다. 초기 상위 빈도를 가진 단어는 선정된 장르별 대표 책들에서 사용되는 고유 명사가 대부분이지만(철학 장르는 예외), 이후로는 각 장르별로 특화된 단어들이 지속적으로 등장하는 것을 볼 수 있다.

표 5. 장르별 특이 단어를 모은 '장르독자성사전'의 일부

순위	판타지	철학	여성소설	S. F
1	harry	therefore	kate	baley
2	frodo	conception	elinor	daneel
3	gandalf	object	anna	rik
4	ron	experience	marianne	terens
5	sam	existence	edward	bonforte
6	hagrid	pure	rochester	karellen
...	...	...	...	...
41	wizard	necessarily	behaviour	trantor
42	journey	regard	obliged	bugs
43	legolas	proposition	engagement	discovery
...	...	...	...	...

5. 장르근접도 추출

이제 4개의 장르별로 각 장르의 특성을 강하게 갖는 단어들로 이루어진 장르독자성사전을 기준으로 임의의 책을 입력하여 그 책에 사용된 단어들과 각 장르독자성사전을 비교하는 과정을 통하여 그 책이 어떤 장르에 가까운지를 도출하고자 한다.

이를 위해 앞서 언급한 단어 빈도수 추출 프로그램을 통하여 입력 책의 단어를 빈도순으로 정리한 단어 사전을 만들고, 이를 각 장르독자성사전에 등재된 단어들과 비교하는 프로세스를 행하도록 알고리즘을 설계하였다. 임의의 입력 서적에 있는 단어가 각 장르독자성사전에서 발견될 경우, 이를 'Word-Hit'이라 칭하고, 각 동일 단어 사이의 순위의 차이를 구하여 그 값의 평균을 'Average Word Distance'라 명명하였다. 결국 특정 장르와 입력 책 간의 'Average Word Distance' 값이 작을수록, 그리고 Word-Hit이 높을수록 입력 서적이 해당 장르와 유사한 단어의 쓰임을 갖는다고 예측할 수 있다.

[표 6]은 임의의 7권의 서적들에 대해 장르 근접도를 계산한 결과이다. 실험을 위해 선정된 책은 각각 가족적이면서도 판타지 성격을 보이는 Lewis Carroll의 'Alice's Adventures in Wonderland(이상한 나라의 앨리스)'와 S. F 서적이면서도 철학적인 사고가 돋보이는 Stanislaw Lem의 'Solaris(솔라리스)', 그리고 판타지 소설의 대표작인 J. K. Rowling의 'Harry Potter and the Prisoner of Azkaban(해리포터와 아즈카반의 죄수)', 대표적 여류소설 작가인 Jodi Picoult의 소설 'Nineteen Minutes', Bertrand Russell의 철학책인 'The Analysis of Mind(정신분석)', 그리고 Adam Smith의 'An Inquiry into the Nature and Causes of the Wealth of Nations(국부론)', 가족 소설의 고전으로 자매들 간의 이야기를 다루고 있는 Louisa May Alcott의 'Little Women(작은 아씨들)'이다.

입력된 7권의 책들에 대한 장르 접근도 추출 결과 7권 모두 예외 없이 우리가 일반적으로 인식하고 있는 장르와 일치하는 결과를 보이고 있을 뿐 아니라, 'Solaris'와 같이 S. F이면서도 철학적인 내용을 동시에 갖는 책의 경우 철학 장르와 S. F 장르 모두 연관성이 있는 것으로 나타나, 한 책에 대한 복합 장르적 성격까지 유추할 수 있었다.

이러한 실험 결과를 바탕으로, 4장에서는 본 장에서 언급한 일련의 과정을 통해 결과적으로 도출된 각 장르별 근접도를 데이터로 하여 이를 시각적 요소로 치환시켜 책의 장르를 직관적으로 나타낼 수 있는 시각화 기법에 대해 제안한다.

표 6. 각 입력 책에 대한 Word-Hit과 Average Word Distance 추출 결과

장르	제목	Word - Hit				Word Average Distance			
		판타지	철학	여성소설	S. F	판타지	철학	여성소설	S. F
판타지	Alice's Adventures in Wonderland	274**	188	229*	108	116.03*	116.34*	117.45	118.13
	Harry Potter (book 3)	350**	128	208*	120	115.06**	121.31	120.74	119.28
S. F	Solaris	127	374**	243	307*	119.55	119.14*	119.51	118.05**
철학	The Analysis of Mind	82	880**	467	355	116.82	90.79**	113.36	114.30
	Inquiry into the Naturea	125	687**	600	288	116.41	103.88**	110.85	114.06
여성소설	Nineteen Minutes	86	145	270**	115	119.41	121.50	118.46*	120.17
	Little Women	190	267	525**	153	119.07	118.66*	118.54*	119.74

Ⅳ. 장르 시각화 및 분석

1. 장르 시각화 설계

본 연구에서는 임의의 서적 텍스트가 갖는 장르적 특성을 시각화하는 방법으로서 앞서 도출된 각 장르독자성 사전에 포함된 단어의 수와 함께 앞서 도출된 Average Word Distance를 통해 알 수 있는 4가지 장르에 대한 각 장르별 근접도를 요소로 하여 서적 텍스트 사용 단어 기반의 장르 정보를 직관적으로 표현하는 것을 최종 목표로 하였다. 따라서 이를 위해 시각적으로 표현되어야 할 필수 속성은 다음과 같다.

1) 장르독자성사전에 포함된 단어
2) Average Word Distance를 통해 알 수 있는 4가지 장르에 대한 각 장르별 근접도

본 시각화의 독창성 및 차별점은, 한 권의 책이 갖는 4가지 장르별 독자적 단어와 그 장르 근접도가 하나의 이미지로 표현함으로써, 더 높은 장르 근접도를 갖는 장르적 속성이 시각적으로 대비되어 드러나도록 설계하고자 하였다는 점이다. 따라서 각 장르별로 임의의 색상을 할당하여 각 장르의 장르독자성 사전에 포함된 단어들과 각 장르별 근접도를 시각적으로 표현하고자 하였다. 색상의 할당은 [표 7]과 같다.

각 장르별 색상을 [표 7]과 같이 할당한 후, 각 장르별 독자성사전에 있는 단어들과, 그 단어들의 장르 근접도를 시각적으로 표현하기 위해서 임의의 한 책에서 각 장르독자성사전에 포함된 단어가 나타날 경우 그 단어를 하나의 점(dot)으로 표현하되, 발견된 장르독자성사전에 해당하는 장르의 할당 색상을 갖도록 설계하였다. 그리고 각 단어에 대해 각 장르독자성사전에서 높은 순위에 위치할수록 점의 크기가 미세하게 커지면서 명도가 높아지도록 조정하였다. 이렇게 분산된 점으로만 표현할 경우 전체적으로 각 장르별 근접도를 명확하게 구분하기 어렵게 될 수 있으므로, 같은 색상을 갖는 두 점들을 하나의 선으로 연결하되 상대적으로 더 많은 Word-Hit을 갖는 장르일수록 시각적으로 더 두드러지게 표현될 수 있도록 [그림 6]과 같이 베지어 곡선(Bezier Curve)[1]으로 표현하였고, 그 곡률을 상대적으로 더 크게 할당하였다. 또한 도출된 Word-Hit과 Average Word Distance로 판별된 각 장르별 근접도가 낮을수록 더욱 흐릿하게 표현하여(blur effect) 그 시각적 감도가 낮아지도록 함으로써, 결국 최종 이미지에서 가장 높은 근접도를 갖는 장르의 점과 라인의 색상이 상대적으로 더욱 눈에 띌 수 있도록 설계하였다.

표 7. 각 장르별 색상 할당

장르	색상
판타지	주황
철학	보라
S. F	파랑
여성소설	노랑

1) n개의 점으로부터 얻어지는 n − 1차 곡선

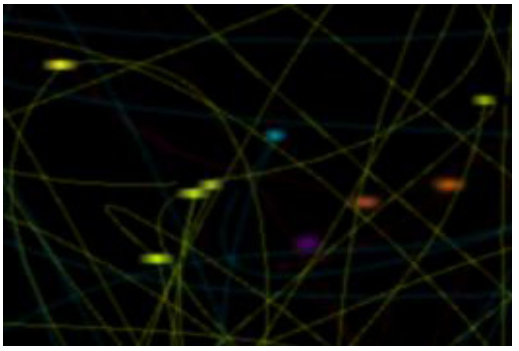


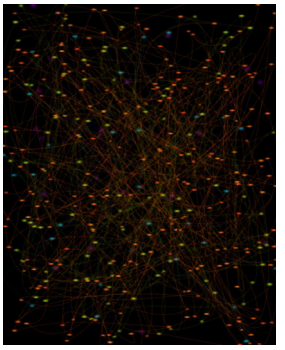
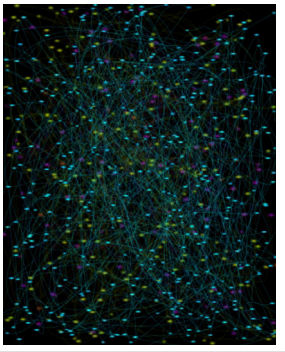
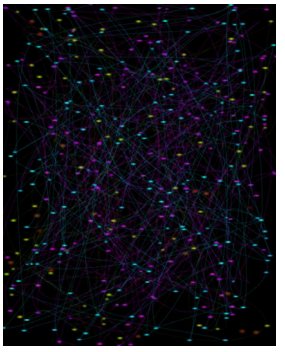
그림 6. 같은 색상의 점(dot)을 장르별 근접도에 따라 각기 다른 곡률의 베지어 곡선(Bezier Curve)으로 연결한 모습

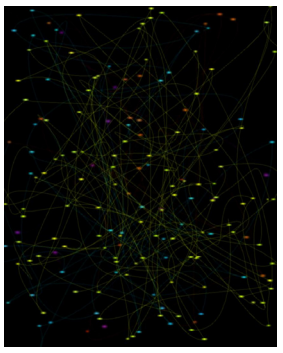
2. 시각화 결과 및 분석

앞서 설계한 장르 근접도 표현을 위한 시각화 방법론의 타당성 및 효율성 판단을 위해 위의 3장에서 장르 근접도 계산을 위해 임의로 선정된 각 장르 서적들 중 각각 한 권 씩을 선정한 후, 제안된 시각화 방법을 적용한 각각의 결과 이미지를 도출하였다. 실험을 위한 각 장르별 대표 서적으로는 판타지의 경우 '해리포터 3권', 철학서적으로는 B. Russeell의 '마음의 분석(The Analysis of Mind), S. F 장르로서 Stanislaw Lem의 '솔라리스(Solaris)', 여성소설로는 Luisa May Alcott의 '작은 아씨들(Little Women)'을 각각 선정하였다.

[표 8]은 선정된 각 장르 서적에 대한 시각화 결과 이미지를 책의 제목과 함께 정리한 것이다. 각각의 시각화 이미지를 살펴보면, 먼저 'Harry Potter 3권(해리포터와 아즈카반의 죄수)'의 경우 전반적으로 붉은 계열로 나타나 판타지 소설의 성격을 강하게 나타내고 있음을 알 수 있다. 그리고 Russell의 대표적인 철학 저서인 'Analysis of Mind(마음의 분석)'의 경우 전체적으로 푸른색을 갖는데, 다른 결과 이미지들과 비교해 볼 때 대표색(가장 두드러지는 색상)을 갖는 점의 개수가 상대적으로 많은 것을 볼 수 있다. 이는 철학 장르의 독자성 사전에 포함된 단어들을 다른 책들의 대표적 장르독자성사전의 일치 단어들보다 훨씬 그 빈도가 높다는 것을 의미하며, 이는 매우 일반적인 철학 저서의 성격을 갖고 있음을 내포한다. 'Solaris(솔라리스)'는 일반적 장르 분류상 S. F임에도 불구하고 철학적 내용을 함께 갖는 책으로, 그 시각화 결과물 또한 보라색과 파란색이 거의 비슷하게 나타나, 복합 장르적 성격을 알 수 있다. 마지막으로 'Little Women(작은 아씨들)'의 경우, 여성소설 장르의 대표색인 노란색이 다른 장르 색상에 비해 상대적으로 많이 나타나고 있지만, 그 절대량은 매우 적다는 것으로 미루어볼 때, 해당 장르와 가장 밀접하지만 그 특성은 강하게 두드러지지는 않는다는 정보까지도 함께 파악할 수 있다.

표 8. 임의의 책 4권에 대한 시각화 결과 이미지

책	시각화 이미지
Harry Potter (book 3) by J. K. Rowlling	
The Analysis of Mind by Bertrand Russell	
Solaris by Stanislaw Lem	

Little Women by Louisa May Alcott	

V. 결론

본 연구에서는 서적이 갖는 장르적 특성을 서적 본문에 직접 사용된 단어들을 바탕으로 파악해낼 수 있는 방법에 대해 고찰하고, 실험을 통한 검증을 토대로 서적 장르 시각화의 요소를 도출한 후, 이를 직관적이고 효율적으로 시각화하는 방법에 대해 서술하였다. 본 서적 장르 시각화 연구가 갖는 의의는 다음과 같다.

첫째, 일반적으로 저자나 출판사에 의해 주관적으로 분류되는 책의 장르를 책에 직접 사용된 단어 데이터의 분석을 바탕으로 해당 서적의 실질적 장르를 판단하는 방법론에 대하여 제안하고, 이를 검증하였다.

둘째, 도출된 장르 근접도를 바탕으로 제안된 시각화 결과 이미지를 통해 해당 서적이 갖는 장르 근접도 정보를 한 눈에 직관적으로 파악할 수 있다.

셋째, 시각화 결과 이미지를 통해 한 책이 갖는 복합 장르적 특성을 알 수 있음과 동시에, 이미지 내의 점(dot)의 개수와 곡선의 곡률, 밝기 등을 통해 대표 장르로 파악된 장르의 근접도(유사도)를 짐작할 수 있다.

나아가 본 연구에서 제안된 시각화 연구 프로세스는 개별 소비자 자신이 선호하는 서적들에 대한 개인별 선호 서적(또는 장르) 이미지를 제공하는 등 서적 추천 시스템 등의 북 커스터마이징(book customizing) 서비스 분야에도 다양하게 활용될 수 있다는 점에서 연구의 가치를 갖는다.

참 고 문 헌

[1] H. Kim and J. W. Park, "Textual Visualization based on Readability," Proceeding of ACM SIGGRAPH Asia 2011, 2011.

[2] 김효영, 박진완, "텍스트의 난이도 파악을 위한 가독성 정보의 시각화", 한국디지털디자인학회, Vol.12, No.2, 2012.

[3] T. Legan and L. Becker, "Visualizing the Text of Philip Pullman's Trilogy "His Dark Materials," Proceeding of NordiCHI 2010, 2010.

[4] http://chrisharrison.net/index.php/Visualizations/BibleViz

[5] http://itsbeenreal.co.uk/index.php?/wwwords/about-this-project/

[6] http://textarc.org/

[7] www.amazon.com

저 자 소 개

김 효 영(Hyoyoung Kim) 정회원

- 2006년 8월 : 성신여자대학교 미디어정보학부(공학사)
- 2010년 8월 : 중앙대학교 첨단영상대학원 영상학과(영상학석사)
- 2010년 9월 ~ 현재 : 중앙대학교 첨단영상대학원 영상학과 박사과정

<관심분야> : 데이터 시각화, 정보 시각화

박 진 완(Jin Wan Park) 정회원

- 1995년 2월 : 중앙대학교 컴퓨터공학과(공학사)
- 1998년 : Pratt CGIM Computer Media(MFA)
- 2003년 3월 ~ 현재 : 중앙대학교 첨단영상대학원 교수

<관심분야> : Art&Technology, Procedural Animation