한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

# 도서 데이터와 본문 텍스트 통합 마이닝을 기반으로 한 도서 콘텐츠 장르 분석 및 시각화 시스템 구현

홍민하°, 박경훈*, 이원진**, 김승훈***
°단국대학교 멀티미디어공학과
*(주)스타네이션
**단국대학교 소프트웨어학과
***단국대학교 응용컴퓨터공학과
e-mail:mhdjae66@hanmail.net°, jamespark@starnation.co.kr*
god7300@dankook.ac.kr**, edina@dankook.ac.kr***

# Implementation of Analysis of Book Contents Genre and Visualization System based on Integrated Mining of Book Details and Body Texts

Min-Ha Hong°, Kyoung-Hoon Park*, Won-Jin Lee**, Seung-Hoon Kim***
°Dept. of Multimedia Engineering, Dankook University
*Starnation Inc.
**Dept. of Software Science, Dankook University
***Dept. of Applied Computer Engineering, Dankook University

● 요 약 ●

최근 IT기술의 발달로 인하여 다양한 분야에서 IT기술을 활용한 융합기술의 시도가 많아지고 있다. 특히 인터넷의 발달과 전자책(e-Book) 시장규모가 커짐에 따라 도서에 대한 정보가 많아지고 있으며, 이러한 정보를 분석하여 활용하는 서비스 시스템에 대한 관심이 높아지고 있다. 하지만 현재 서비스되고 있는 대부분의 온라인 서점에서는 도서의 기본 서지정보와 같이 도서 본문 내용과는 무관한 출판사나 서점에서 도서를 관리하기 위한 정보만을 제공하고 있으며, 도서에 대한 다양한 정보를 활용한 키워드 추출 및 장르 분류를 통한 검색의 효율성 제공이 미흡한 현실이다.

본 논문에서는 도서의 본문 텍스트 정보를 마이닝 처리하여 도서 페이지의 흐름에 따라 포함되어있는 장르를 분류하고 이에 대한 결과를 사용자에게 친화적인 시각화 기법으로 제공되는 시스템을 설계하고 구축하였다. 제안한 서비스 시스템은 의미 분석을 기반으로 도서 정보의 구체적, 실제적, 직관적 정보를 제공하여 도서 추천 서비스에 활용될 것이다.

**키워드:** 도서(book), 텍스트 마이닝(text mining), 시각화(visualization)

## I. 서 론

오늘날 인터넷의 발달과 전자책(e-Book) 시장규모가 커짐에 따라 도서에 대한 대량의 데이터들이 쏟아져 나오고 있으며, 이에 따라 온라인 서점 웹사이트들은 여러 도서 정보를 제공하고 있다. 또한, 소비자 입장에서도 관심 있는 책을 찾기 위해 서적 판매 사이트에서 키워드를 입력하거나 주제 분류를 이용하면 되고, 책을 고르기 위해서는 웹사이트에서 제공하는 책에 대한 다양한 정보를 활용하면 서점에서 책을 일일이 뒤지는 것보다 효과적이다[1].

하지만, 저작도구(authoring tool)의 발달로 인해 다양한 분야에서 수많은 도서들과 문헌들이 출판되고 있고, 도서검색 가능을 제공하고 있음에도 불구하고 사용자들은 자신이 원하는 도서들을 찾기 위해 점점 더 많은 시간과 노력을 기울이고 있다[2]. 현재 국내외의 도서정보를 제공하는 대부분의 사이트에서는 도서 이미지, 서명, 저자 등 기본 서지정보만을 제공하고 있다. 일부 사이트에서 관련도서에 대한 추천이나 도서의 일부분을 발췌한 내용을 서비스해주고 있지만, 도서 전체 내용에 대한 정보 제공은 하지 않고 있다.

그러나 도서 콘텐츠의 다양한 정보 전달과 서적 구매 유도 및 도서 콘텐츠 이용 활성화를 위해서는 도서 본문에 대한 의미 분석을 기반으로 구체적, 실제적, 직관적 정보를 사용자들에게 시각화시켜주는 서비스가 필요하다.

따라서 본 논문에서는 도서의 본문 텍스트 정보를 마이닝 처리하여

도서의 흐름에 따른 장르를 분류하고, 이에 대한 결과를 사용자에게 친화적인 시각화 기법으로 제공되는 시스템을 설계하고 구축하였다.

## II. 관련 연구

효과적인 도서 추천 웹사이트를 구축하기 위해서는 우선적으로 도서에 대한 정보 제공과 직관적인 시각화 표현이 이루어져야한다. 이와 관련된 기존의 연구들 중 [1]에서는 현재 활발하게 운영 중인 인터넷 서점 사이트를 인터페이스 측면, 정보원 구성 측면, 기능적 측면 세 가지에 중점을 두고 조사, 분석하여 효과적인 인터넷 서점 웹사이트 구축 모형을 제시하였으며, 내용의 충실성과 차별성 그리고, 도서에 대한 풍부하고 다양한 정보 제공의 필요성을 설명하고 있다. [3]은 서적 텍스트에 사용된 단어 데이터를 분석하여 서적의 장르를 판별하고 판별된 장르 정보를 시각적 요소로 맵핑하여 이미지 형태로 시각화하는 방법을 제안하였다. 하지만 도서 전체의 흐름에 따른 상세 장르 분석 및 그에 대한 시각화는 찾아볼 수 없었다. 또한, 현재 제공되는 온/오프라인 서점이나 도서관에서는 도서의 장르를 하나로 분류하는데, 뉴스 등 일반 텍스트보다 텍스트 양이 많은 도서를 하나의 장르로만 판단하는 것은 정보를 온전히 활용하였다고 볼 수 없다.

정보의 빅뱅 시대에는 방대한 정보를 분류하고 유용한 정보를 골라내어 사용자에게 차별화된 정보를 제공해주는 서비스가 중요하므로, 본 논문에서 제안하는 도서 정보와 도서 본문 텍스트의 내용적 의미 분석을 통합한 도서 정보 제공 시스템의 설계 및 구축에 관한 연구가 필요하다.

## III. 도서 콘텐츠 장르 분석 및 시각화 시스템

본 논문에서는 도서 본문 텍스트를 활용하여 도서에 포함된 장르를 분석하고 이를 사용자 친화적인 시각화한 방법으로 정보를 제공하는 서비스 시스템을 설계, 구축한 모형을 제안한다. 제안한 모형은 도서 데이터와 본문 텍스트 통합 마이닝을 기반으로 한 내용 중심의 지식 마이닝 체계를 개발하고, 이를 기반으로 사용자 친화적 도서 정보 시각화 및 사용자 참여형 도서 추천 큐레이션 플랫폼 기술을 개발하기 위한 사전연구이다. 다음 그림 1은 제안한 서비스 시스템의 전체적인 구성도이다. 도서의 본문 텍스트를 분석하는 처리과정을 거쳐 나온 분석된 결과를 도서 장르 요약 분석, 도서 장르 상세 분석, 두 가지 구성으로 시각화하여 표현한다.

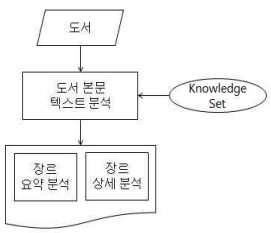


그림 1. 시스템 구조
Fig. 1. The proposed System Architecture

### 1. 도서 장르 요약 분석

도서 본문 전체 텍스트를 분석한 정보를 이용하여 장르의 정도를 간략하게 표현한 서비스이다. 먼저, 도서 본문을 이용, 형태소 분석하여 추출된 단어들을 가지고 도서에 포함되는 장르들을 파악한다. 장르 파악은 그림 3와 같이 장르별로 관련 단어 리스트를 가지고 있는 Knowledge Set을 미리 구축해놓고 이를 활용하였다. 두 번째는 텍스트 마이닝하여 분석된 모든 장르들 중 일정 기준치 이상에 부합하는 3~4개를 선택하여 도서 요약분석의 항목(장르)으로 적용하여 나타내었다. 즉, 그래프의 항목들이 해당 도서를 읽으면서 파악할 수 있는 대표적인 장르들인 것이다. 도서 본문 전체 텍스트가 없을 경우는 도서의 요약/서지 정보만 형태소 분석하여 장르의 정도를 표현한다. 마지막으로 장르의 정도는 항목(장르)의 많고 적음, 늘고 줄어드는 양을 쉽게 비교하여 볼 수 있는 가로 막대그래프를 형태로 시각화 처리하여 제공하였다.

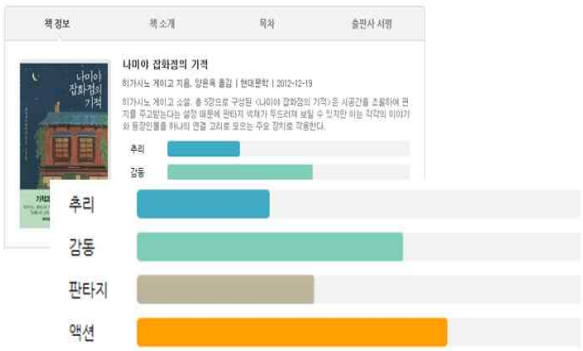


그림 2. 도서 장르 요약 분석 결과 및 시각화
Fig. 2. Book Analysis Result Summary and Visualization

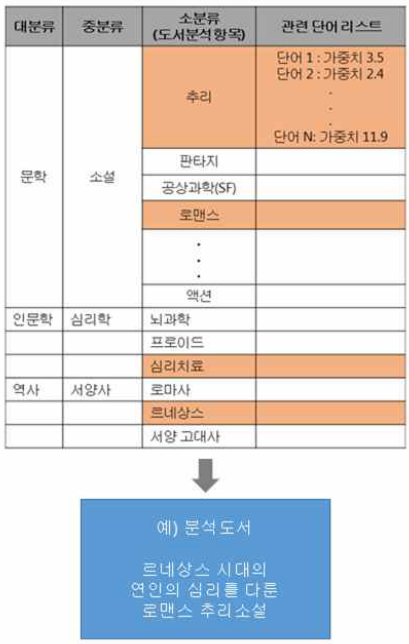


그림 3. Knowledge Set을 이용한 장르 분석
Fig. 3. Genre Analysis using Knowledge Set

## 2. 도서 장르 상세 분석

본 논문에서 제안하고 구축한 시스템에서는 도서 안에 포함되어있는 다양한 장르들을 파악할 수 있도록 도서의 구간별로 나타나는 장르의 정도를 Area 그래프로 시각화하였다. 이러한 방식은 도서의 구간을 구분하고, 장르의 해당 정도를 구분하여 장르의 포함률을 나타낸다. 따라서 사용자는 한 권의 도서 안에 어떤 장르들이 얼마나 포함되어 구성되어 있는지를 한 눈에 파악할 수 있어 간접적으로 도서를 읽은 효과를 얻을 수 있다.

도서 본문 full text를 형태소 분석하여 추출된 단어들은 도서의 장르를 파악하는 기준이 된다. 분야별로 많이 사용하는 단어들이 있고, 본문에 사용된 단어가 동음이의어로 사용되기도 하기 때문에 단어의 장르파악은 그림 4의 Knowlege Set을 이용해 해당하는 장르를 파악한다. 분석된 모든 장르들 중 일정 기준치 이상에 부합하는 n개의 장르를 그래프의 계열로 적용하여 시각화하였다.

또한, 그래프의 x축은 페이지 단위로 도서의 구간을 구분하고, y축은 장르의 해당 정도를 구분하여 표시한다. x축, y축은 도서에 따라 각각 n구간, m단계로 설정될 수 있다. y축이 커질수록 해당 장르의 성격이 짙다고 볼 수 있다. 따라서 사용자는 해당 그래프를 통해 책 한권의 장르변화가 어떻게 진행되는지를 직관적으로 확인할 수 있다. 다음의 그림 5에서는 300페이지 부근에는 액션과 추리 장르의 성격이 강한 반면, 700페이지 부근에서는 판타지 장르의 성격이 강함을 나타내고 있다.

| 소분류 (도서분석항목) | 관련 단어 리스트 |
|---|---|
| 추리 | 단어 1 : 가중치 3.5<br>단어 2 : 가중치 2.4<br>.<br>.<br>.<br>단어 N: 가중치 11.9 |

그림 4. Knowledge Set
Fig. 4. Knowledge Set

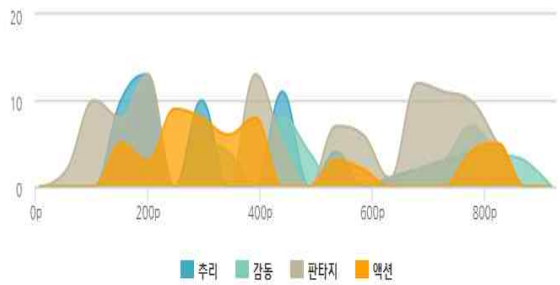


그림 5. 도서 장르 상세 분석 결과 및 시각화
Fig. 5. Book Detailed Analysis Result and Visualization

## IV. 결 론

본 논문에서는 도서 본문 텍스트를 활용하여 도서의 흐름에 따라 포함된 장르를 분석하고, 그 결과를 시각화한 서비스 시스템을 설계, 구축한 모형을 제시한다. 그 내용은 다음과 같다. 우선, 도서 본문 텍스트를 형태소 분석하여 추출한 단어들을 Knowledge Set을 이용하여 마이닝한다. 장르 요약 분석에서는 분석된 장르들 중 일정 기준치 이상에 부합하는 장르들을 각각 항목으로 두어 포함정도를 막대그래프로 표현한다. 장르 상세 분석에서는 x을 페이지 구간, y축을 포함 단계로 설정하여 도서의 페이지 흐름에 따라 포함되는 장르와 그 정도를 Area 그래프를 통해 나타낸다.

이는 도서 데이터와 본문 텍스트 통합 마이닝을 기반으로 한 내용 중심의 추천 지식 마이닝 체계를 개발하고, 이를 기반으로 사용자 친화적 도서 정보 시각화 및 사용자 참여형 도서 추천 큐레이션 플랫폼 기술을 개발하기 위한 사전연구로 활용될 것이다. 또한, 이러한 서비스는 의미 분석을 기반으로 도서 정보에 대해 정확한 추천을 가능하게 하고 양질의 도서 콘텐츠를 제공하여 고객의 만족도와 충성도를 높일 것이다. 그리고 도서 정보의 구체적, 실제적, 직관적 정보 전달을 통한 서적 구매 유도 및 도서 콘텐츠 이용 활성화를 시킬 것이다. 그리하여 기존의 베스트셀러 중심에서 양서 중심으로 도서 선택의 폭을 넓혀 도서 콘텐츠의 균형 잡힌 소비를 유도하고 소량 다품종 판매 형태로의 온라인 서점 패러다임을 변화시킬 것이다.

## Acknowledge

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2014년도 문화기술연구개발지원사업의 연구결과로 수행되었음

## 참고문헌

[1] Lan-Ju Lee, "A Study on a Design of Effective Websites for Internet Bookshops," Journal of the Korean society for information management, Vol. 18, No. 1 = No.39, pp. 43-64, March 2001.

[2] Sang-Gi Lee, "A Study of Intelligent Recommendation System based on Naive Bayes Text Classification and Collaborative Filtering," Journal of information management, Vol. 41, No. 4, pp. 227-249, Nov. 2010.

[3] Hyo-Young Kim and Jin-Wan Park, "Book Genre Visualization based on Genre Identification Algorithm," The Journal of the Korea Contents Association, Vol. 12, No. 5, pp.52-61, 2012.