Journal of The Korea Society of Computer and Information
Vol. 20, No. 2, February 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.2.121

# TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구

유 은 순*, 최 건 희**, 김 승 훈***

# Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels

Eun-Soon You *, Gun-Hee, Choi **, Seung-Hoon Kim***

## 요 약

도서 상품에 대한 정보량이 폭증하면서 고객이 도서 선택에 어려움을 겪는 상황이 발생하고 있다. 이에 따라 고객에게 적합한 도서 정보를 제공하여 구매를 유도하는 도서 추천시스템의 중요성이 커지고 있다. 하지만 도서의 서지정보나 사용자 정보 등을 이용한 기존의 추천시스템은 추천 결과의 신뢰도에 문제를 드러내고 있기 때문에 도서 본문 텍스트의 의미적 정보를 추천시스템에 반영하는 것이 필요하다. 따라서 본 논문은 이에 대한 선행연구로 TF-IDF기법과 소설의 외형적 구조를 이용한 소설 텍스트의 주제어 추출 방법을 제안하였다. 이를 위해 100권의 소설텍스트를 수집하고 각각의 소설을 머리말, 대화문, 비대화문, 맺음말의 4개의 구조로 분리한 후 TF-IDF 가중치를 계산하였다. 실험결과 본문 텍스트만을 이용했을 때 보다 머리말과 맺음말을 포함하고 대화문에 가중치를 높게 부여하였을 때 주제어의 추출 정확도가 42.1%의 성능 향상을 보였다.

▸ Keywords : 주제어, TFIDF, 소설구조, 도서추천시스템, 대화 가중치

## Abstract

With the explosive growth of information about books, there is a growing number of customers who find it difficult to pick a book. Against the backdrop, the importance of a book recommendation system becomes greater, through which appropriate information about books could be offered then to encourage customers to buy a book in the end. However, existing recommendation systems based on the bibliographical information or user data reveal the reliability issue found in their recommendation results.

•제1저자 : 유은순 •교신저자 : 김승훈
•투고일 : 2015. 2. 5, 심사일 : 2015. 2. 9, 게재확정일 : 2015. 2. 20.
* 단국대학교 미디어콘텐츠연구원(Institute of Media Content)
** 단국대학교 소프트웨어학과 (Dept. of Software Science, Dankook University)
*** 단국대학교 응용컴퓨터공학과(Dept. of Applied Computer Engineering, Dankook University)
※이 논문은 2014 한국지능정보시스템학회 추계학술대회에서 발표한 논문("도서 텍스트 본문 주제어 추출 연구[14]")을 확장한 것임

This is why it is necessary to reflect semantic information extracted from the texts of a book's main body in a recommendation system. Accordingly, this paper suggests a method for extracting keywords from the main body of novels, as a preceding research, by using TF-IDF method as well as the text structure. To this end, the texts of 100 novels have been collected then to divide them into four structural elements of preface, dialogue, non-dialogue and closing. Then, the TF-IDF weight of each keyword has been calculated. The calculation results show that the extraction accuracy of keywords improves by 42.1% in performance when more weight is given to dialogue while including preface and closing instead of using just the main body.

▸ Keywords : Keyword, TFIDF, Novel Structrue, Book Recommendation System, Dialog Weight

## I. 서 론

인터넷에서 도서 상품에 대한 정보량이 폭증하면서 고객이 도서 선택에 어려움을 겪는 상황이 발생하고 있다. 이 문제를 해결하기 위해 아마존(Amazon)을 비롯한 많은 국내외 인터넷 서점들이 추천시스템을 통해 고객에게 적합한 도서를 제공하고 구매를 유도하고 있다. 가장 대표적인 사례로 아마존은 고객의 구매정보와 고객이 클릭한 상품 정보를 바탕으로 도서를 추천한다. 하지만 현재 제공되고 있는 대부분의 추천시스템은 고객의 선호도 정보와 도서 서지 정보, 고객들의 리뷰 정보들을 이용하는데 그치고 있으며 도서 본문 내용에 대한 정보를 고려하지 않기 때문에 추천 정확도에 대한 신뢰도가 떨어진다. 이에 본 연구는 도서 본문의 내용 정보를 추천시스템에 반영하여 추천 성능을 향상시키기 위한 선행 연구로 도서 본문에 대한 주제어 추출을 제안한다.

문서의 유형이 다양해지고 그 양도 폭발적으로 증가하는 환경에서 주제어 추출은 방대한 문서의 내용을 쉽게 파악할 수 있도록 해줄 뿐만 아니라 문서 검색 및 분류 등 다양한 분야에 활용될 수 있기 때문에 그 중요성이 점점 높아지고 있다.

주제어 추출을 위해 본 논문은 가장 대중적인 문학 장르인 소설텍스트를 수집하고 단어의 빈도 값을 표현하는 TF-IDF 가중치 모델과 소설텍스트의 외형적 구조 정보에 가중치를 부여하는 방식을 제안한다. TF-IDF는 어떤 단어가 특정 문서에 자주 출현하지만 전체 문서집합에서는 출현 빈도가 낮은 값을 나타내는 것[1,2]으로 해당 단어가 문서의 내용을 대표하는 중요한 주제어인가를 평가하는데 일반적으로 많이 사용되고 있는 방법이다. 하지만 빈도 정보만으로는 단어의 중요성을 판단하는데 한계가 있기 때문에 본 연구는 단어의 빈도 값뿐만 아니라 소설 텍스트가 갖는 외형적 구조 정보를 이용한다.

텍스트는 그 유형에 따라 정형화된 구조를 갖고 있다. 예를 들어 논문은 제목과 요약, 결론 부분에서, 신문은 기사 제목과 본문 앞 문장에서 주제어를 추출하고 있다[3,4]. 소설 내용 또한 일반적으로 머리말과 부(部)나 장(章), 절(節), 그리고 맺음말 혹은 작가 후기가 위계적 질서를 이루고 있다. 소설의 머리말과 맺음말은 소설 내용의 일부로 보는 견해가 강하기 때문에 본 연구는 주제어 추출을 위해 소설 본문뿐만 아니라 머리말과 맺음말도 포함하였다. 또한 소설 연구에서 등장인물들의 대사가 작가의 작품세계와 주제를 드러내는데 유용한 형식이라는 점이 강조되어 온 점을 고려하여 대화문에 가중치를 부여하였다. 따라서 본 논문의 목적은 소설의 주제어 추출을 위해 소설 구조가 반영된 개선된 TF-IDF를 제안하는 것이다.

본 논문의 구성은 다음과 같다. II장에서는 TF-IDF를 이용한 주제어 추출과 관련된 연구를 소개하고 III장에서는 주제어 추출 프로세스를 제안한다. 그리고 IV장과 V장에서는 각각 실험 및 성능 평가와 결론 및 향후 연구를 기술한다.

## II. 관련 연구

### 1. TF-IDF 가중치 모델

TF값은 한 문서에 사용된 모든 단어들의 출현 빈도를 나

타낸 값으로 출현 빈도가 높은 단어일수록 해당 문서에서 중요도가 높은 것으로 판단한다. 하지만 TF 값이 높다고 해서 그 단어를 해당 문서의 키워드로 볼 수는 없다. 왜냐하면 다른 주제의 문서에서도 많이 사용될 수 있기 때문이다. 따라서 다른 문서 집합에도 보편적으로 출현하는 단어들의 IDF 값을 구해 주제어 추출에서 제외시켜야 한다.

TF-IDF는 어떤 단어가 해당 문서에서 자주 사용되지만 다른 주제의 문서 집합에서는 출현 빈도가 낮은 값을 표현하는 것으로 문서에서 단어의 중요도를 평가하는데 일반적으로 사용되고 있는 방법이다[1,2].

## 2. TF-IDF 가중치 모델을 이용한 주제어 추출

TF-IDF는 그동안 정보 검색과 텍스트 마이닝에서 텍스트의 주제어를 추출하는데 사용되었다. 국외의 경우 이미 구축된 대규모 코퍼스로부터 주제어를 추출하거나[3,4] 웹 문서의 검색 결과 성능을 향상시키기 위해 주제어를 추출하여 유사도에 따라 문서를 분류하였다[5,6].

국내의 경우 이성직 등은 정치, 경제 등 각 분야별로 수집된 뉴스 기사로부터 분야별 키워드를 추출하기 위해 TF-IDF 가중치 모델을 6가지로 변형하여 분야 간 교차비교 분석을 진행하였다[1]. 고광수 등은 기존의 특허문서의 제목과 요약, 대표도면의 내용정보를 위주로 키워드를 추출한 것과는 달리 TF-IDF를 이용하여 문서 전체로부터 키워드를 추출하여 특허 검색에서 질의 문서와 유사한 문서들을 찾아내는 연구를 진행하였다[2].

정석팔 등은 주어진 질의어에 대한 검색 결과를 분류하기 위해 문서 군집의 제목과 요약문에 나타나는 명사의 TF와 IDF를 이용하여 단어의 가중치를 계산하였다[7].

최홍구 등은 음악 가사와 음악에 대한 사용자의 SNS 비정형 데이터에 대해 TF-IDF 모델을 적용하여 대표 감정 주제어를 추출하여 가사 무드 분류기를 개발하였다[8].

TF-IDF는 문서의 내용을 대표하는 주제어를 추출하는데 유용하게 사용되어 왔지만 빈도 정보에 의한 방법은 한계가 있다. 따라서 본 논문은 빈도 정보뿐만 아니라 소설텍스트의 구조 정보를 이용하여 가중치를 부여함으로써 주제어 추출의 성능과 정확도를 향상시키고자 한다.

## 3. 소설의 외형적 구조

주제어 추출을 위해 텍스트의 유형에 따른 구조를 이용하는 방법들이 제안되었다. 웹 문서는 태그 정보를 이용할 수 있으며, 학술 논문의 경우 핵심 내용을 담고 있는 제목, 요약, 결론 부분에 출현한 용어에 가중치를 부여할 수 있고, 신문의 경우 기사 제목 및 본문의 앞 문장에 출현한 단어들이 주제어일 가능성이 높은 것으로 판단한다[9,10].

문학에서는 소설의 구조를 형식과 내용을 모두 포함하는 넓은 시각에서 다루고 있으며, 그 정의를 소설을 이루고 있는 모든 요소와 그 요소들의 관계로 기술한다[11]

일반적인 소설의 외형적 구조는 첫 페이지에 장르 표시와 제목 그리고 작가의 이름이 나오고 내용이 시작된다. 내용은 다시 부(部)나 장(章)으로 나뉘지만 내용이 방대할 경우 권(卷), 부(部)나 장(章) 그리고 절(節)로 이루어진다. 여기에 '머리말'이나 '맺음말'이 있는 경우도 있다[11]. 소설에서 이러한 단위들이 상하의 위계질서를 이루고 있으나 작가에 따라 그 단위의 사용이 다양하다.

소설의 머리말은 소설의 내용과 독립적인 경우도 있지만 소설의 집필 동기나 소설의 모티프가 됐던 사건, 소설을 이해하는데 도움이 되는 정보 등을 제공하기 때문에 소설 내용의 일부로 보기도 한다. 맺음말 또는 작가 후기 역시 소설 내용을 포함하기도 하지만 그렇지 않은 경우도 있다[11]

소설 구성의 기본 단계인 대화는 등장인물들의 발언을 그대로 기술하는 형식적 어법으로 일반적으로 인용 부호가 앞뒤에 붙어 있다. 대화는 인물의 성격을 직, 간접적으로 묘사하고 인물 간 갈등을 형상화하고 작가의 세계관이나 주제를 드러내는데 효과적인 역할을 한다[11,12].

본 연구는 주제어 추출을 위해 소설의 본문뿐만 아니라 그동안 주제어 추출에서 제외 되었던 머리말과 맺음말을 포함하고, 대화가 소설에서 차지하는 중요성과 기능을 고려하여 대화에 출현한 단어들에 가중치를 부여하였다.

# III. 소설 구조를 반영한 개선된 TF-IDF 기반의 주제어 추출 방법

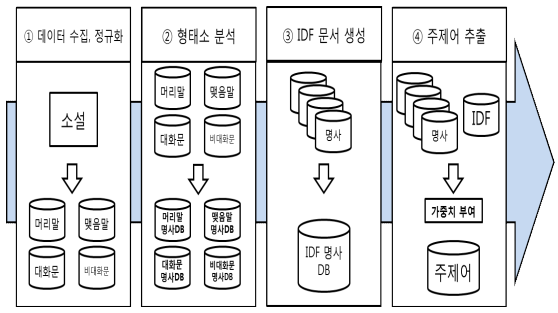


그림 1 . 주제어 추출 프로세스
Fig 1. Keyword Extraction Process

본 연구에서 주제어 추출은 그림 1과 같이 데이터 수집 및 정규화, 형태소 분석, IDF 문서 생성, 주제어 추출 순으로 총 4단계로 진행되었다.

## 1. 데이터 수집 및 정규화

주제어 추출을 위해 총 800권의 전자텍스트 소설을 수집한 후 위에서 기술한 소설의 형식적 구조인 머리말, 맺음말을 갖추고 있는 소설을 분류하여 총 100권의 소설텍스트를 최종 선별하였다. 선별된 텍스트는 추리, 로맨스, SF 등 특정 장르에 편중되지 않도록 다양하게 구성하였다.

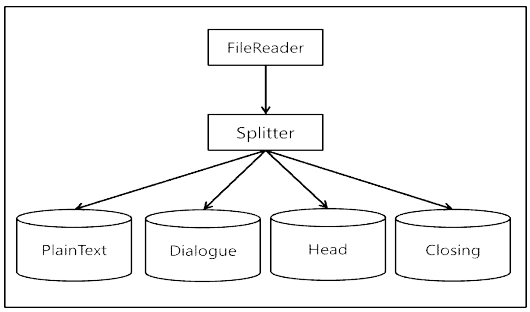


그림 2. 데이터 정규화 프로세스
Fig 2. Data Normalization Process

그림 2는 선정된 100권의 소설에 대한 정규화 과정을 나타낸 것이다.

각각의 소설을 EPUB 파일에서 TXT 파일로 변환한 후 이를 FileReader가 읽고 전체 내용을 저장한다. EPUB 파일의 경우 태그문자를 이용하여 소설의 구조화가 이미 이루어졌기 때문에 소설 구조를 태그문자로 파악할 수 있다. 하지만 각 출판사에 따라 그 구조가 다르기 때문에, 모든 출판사의 태그 구조를 파악하기에 많은 시간이 소요되는 문제점이 있다. 따라서 구조가 명확한 출판사의 EPUB 파일만을 이용하였다.

Splitter는 저장된 소설텍스트의 구조를 머리말(Head)과 맺음말(Closing), 대화문(Dialogue)과 비대화문(Plain Text)의 4가지 TXT파일로 분리한다. 다음 단계에서 진행될 형태소 분석에서는 문장단위로 분석이 진행되기 때문에 분리된 TXT파일들은 문장단위로 각 DB에 저장된다.

## 2. 형태소 분석

한나눔 형태소 분석기[13]를 이용하여 각 소설의 머리말과 맺음말, 대화문과 비대화문에 대해 형태소 분석을 진행하여 명사를 추출하였다. 그림 3은 형태소 분석 결과를 나타낸 것이다. 본 연구에서 명사만을 고려한 이유는 명사가 소설텍스트에서 차지하는 비율이 높고 중요한 정보를 전달하기 때문에 명사만으로도 내용의 핵심을 파악하는 것이 가능하기 때문이다.

한나눔 형태소 분석의 태그셋에서 명사는 고유명사(nq), 의존명사(nb), 대명사(np), 보통명사(nc)로 구분되는데, 의존명사와 대명사는 주제어를 표현하는 적합하지 않으므로 제외하였다. 이처럼 실제 분석에 필요 없거나 형태소 분석기의 오류로 나온 노이즈들을 제거하는 전처리 과정을 통해 명사를 필터링한 후 DB로 구축하였다. 형태소 분석으로 추출된 명사는 정규화된 문서 단위로 저장된다. 각 정규화된 문서에는 명사, 태그, 빈도수가 저장된다.

| | morpheme [PK] character varying(255) | tag character varying(255) | frequency integer |
|---|---|---|---|
| 43 | 시장 | ncn | 7 |
| 44 | 기준 | ncpa | 1 |
| 45 | 차이 | ncn | 8 |
| 46 | 내면 | ncn | 5 |
| 47 | 열간이 | ncn | 1 |
| 48 | 악보 | ncn | 4 |
| 49 | 강직 | ncps | 1 |
| 50 | 탐구욕 | ncn | 1 |
| 51 | 실용 | ncpa | 1 |
| 52 | 어딘가 | ncn | 16 |
| 53 | 재능 | ncn | 6 |
| 54 | 활동 | ncpa | 2 |
| 55 | 오늘 | ncn | 35 |
| 56 | 물체 | ncn | 2 |
| 57 | 웃음 | ncn | 2 |
| 58 | 수전 | ncpa | 1 |
| 59 | 주 | ncn | 1 |
| 60 | 영웅 | ncn | 1 |
| 61 | 여름 | ncn | 2 |
| 62 | 생활 | ncpa | 2 |
| 63 | 지시 | ncpa | 1 |
| 64 | 명심 | ncpa | 1 |
| 65 | 장갑 | ncn | 2 |
| 66 | 소름 | ncn | 1 |
| 67 | 저녁 | ncn | 1 |
| 68 | 멜로디 | ncn | 1 |
| 69 | 위로 | ncpa | 1 |
| 70 | 연마 | ncpa | 1 |
| 71 | 판단 | ncpa | 1 |
| 72 | 피아노 | ncn | 29 |
| 73 | 나이 | ncn | 1 |

그림 3. 형태소 분석
Fig 3. Morpheme Analysis

## 3. IDF 문서 생성

TF-IDF 가중치 이론에서 IDF 값은 전체 문서에서 보편적으로 등장하는 단어일수록 높게 부여된다. 이를 역수를 취하여 보편적으로 등장하는 단어일 경우 최종 가중치에서 낮게 계산된다. 만약 한 단어의 출현빈도수가 400이라면 해당 단어는 모든 문서에서 출현한 것을 의미하기 때문에 IDF 값은 0이 되고 주제어 선별에서 제외된다.

위에서 기술했듯이 100권의 소설은 머리말과 맺음말, 대화문과 비대화문의 4개의 텍스트로 분리되었다. 사용된 실험도서는 총 100권이므로 문서의 총 개수는 400개이다. 400개의 문서군을 이용하여 특정단어의 IDF값을 구하기 위해 필요한 문서를 생성한다. 각 문서에서 나온 단어들을 분석하여 해당 단어가 전체 문서에서 출현한 빈도수를 저장한다.

| | morpheme [PK] character varying(255) | tag character varying(255) | frequency integer |
|---|---|---|---|
| 1 | 종말사상 | ncn | 1 |
| 2 | 낚시바늘 | ncn | 1 |
| 3 | 종반 | ncn | 53 |
| 4 | 서당 | ncn | 3 |
| 5 | 낫날 | ncn | 3 |
| 6 | 인어공주 | ncn | 4 |
| 7 | 사우스 | ncn | 8 |
| 8 | 자궁경부 | ncn | 1 |
| 9 | 토담 | ncn | 5 |
| 10 | 생명 | ncn | 164 |
| 11 | 조소거리 | ncn | 1 |
| 12 | 용급차 | ncn | 1 |
| 13 | 준비 | ncn | 212 |
| 14 | 드리기 | ncn | 1 |
| 15 | 만화경 | ncn | 7 |
| 16 | 예언서 | ncn | 4 |
| 17 | 낙담 | ncpa | 38 |
| 18 | 철골구조 | ncn | 1 |
| 19 | 보전기구 | ncn | 1 |
| 20 | 특별수업 | ncpa | 2 |
| 21 | 곱팡이 | ncn | 29 |
| 22 | 갱년기 | ncn | 3 |
| 23 | 만석꾼 | ncn | 4 |
| 24 | 촌락 | ncn | 6 |
| 25 | 방심 | ncpa | 41 |
| 26 | 유유자적 | ncpa | 4 |
| 27 | 박애 | ncpa | 9 |
| 28 | 느럼보 | ncn | 5 |
| 29 | 해넘 | ncn | 2 |
| 30 | 미주 | ncn | 2 |

그림 4. IDF 문서
Fig 4. IDF Documnet

그림 4는 생성된 IDF 문서의 일부를 보여준 것이다. IDF 문서는 명사, 태그 그리고 명사가 전체 문서군에서 출현한 빈도수를 저장한다. IDF 문서는 각 명사 DB와 같은 구조로 저장된다. IDF 값은 전체 문서군에서 출현한 빈도수를 보기 때문에 특정 문서에서 특정 단어가 높은 빈도수를 보이더라도 전체 문서군에서 출현한 빈도수는 1이 증가한다. 본 연구에서는 전체 문서군의 수는 400개이므로 한 단어가 전체 문서군에서 출현할 수 있는 빈도수는 최대 400이다.

다음 식(1)은 IDF값을 계산하는 방법을 나타낸 것이다.

$$IDF = \log \frac{|P|}{|p_i | w_i \in p_i|} \quad \cdots\cdots\cdots\cdots\cdots (1)$$

식 (1)의 $|P|$는 문서군에 포함되어있는 문서의 수를, $|p_i | w_i \in p_i|$ 는 단어$w_i$가 등장한 문서의 수를 나타낸다. 단어 $w_i$가 등장한 문서 수가 많을수록 그 단어는 흔하게 사용되는 단어이다. 따라서, $|p_i | w_i \in p_i|$ 값이 높을수록 주제어일 확률이 떨어진다. $|p_i | w_i \in p_i|$ 값을 전체 문서의 수 $|P|$로 나눈 뒤 역수를 취한다. 값이 크게 커지는 것을 막기 위해 로그를 취하여 최종 IDF값을 얻는다.

## 3. 주제어 추출

앞 단계에서 생성된 각 문서와 IDF 문서를 토대로 추출된 명사에 가중치를 부여한다. 각 네 가지로 분류된 문서에서 추출된 명사의 빈도수를 이용하여 TF값을 구한다. IDF문서를 이용하여 해당 단어의 IDF값을 구한 뒤 TF-IDF 가중치 기법을 사용하여 각 명사에 가중치를 부여한다. 부여된 가중치를 오름차순으로 정렬하여 각 문서에 하나의 최종 주제어리스트들을 출력한다.

식(2)는 문서 $p$에서 특정단어 $i$의 $TF_{p,i}$값을 표현한 것이다.

$$TF_{p,i} = \frac{N_{p,i}}{\sum_k n_{k,i}} \quad \cdots\cdots\cdots\cdots\cdots (2)$$

식 (2)에서 $N_{p,i}$는 문서에서 등장한 특정 단어 빈도수를, $\sum_k n_{k,i}$는 문서에 등장한 모든 단어 빈도수를 나타낸다. 각 문서 DB에 저장된 단어의 빈도수 $N_{p,i}$를 해당 문서의 비율로 나타내기 위해 문서의 전체 단어 빈도수 $\sum_k n_{k,i}$로 나누어준다. 왜냐하면 단순히 빈도수를 이용하게 되면 절대적 크기로 비교되기 때문에 문서에서 비중이 크지 않지만 자주 등장하는 단어가 높은 TF값을 받을 수 있기 때문이다. 따라서, 전체 빈도수로 나누어 그 비율을 이용하여 해당 단어 $i$의 TF값을 구한다.

본 연구에서는 소설의 구조를 이용하기 때문에 각 구조에 가중치 값을 부여한다. 구조에 대한 가중치는 해당 구조가 얼마나 중요한 지표가 되는지를 설정한다. 본 논문에서는 소설의 대화가 주제를 전달하는데 있어 중요한 역할을 한다는 점을 고려하여 대화문에 대한 가중치를 다양하게 설정한다. 머리말과 맺음말, 대화문과 비대화문에 대해 부여하는 가중치의 합은 식(3)과 같다.

$$w_h + w_c + w_d + w_n = 1 \quad \cdots\cdots\cdots\cdots (3)$$

식 (3)에서 $w_h$는 머리말의 가중치, $w_c$는 맺음말의 가중치, $w_d$는 대화문의 가중치, $w_n$는 비대화문의 가중치를 각각 나타낸다. 각 문서의 특정 단어는 위와 같은 가중치를 문서마다 부여받는다. 모든 가중치의 합은 1로 설정하였다.

소설 텍스트는 모두 4가지의 문서로 구성 되어있으며 앞서 설명한 식(1), 식(2), 식(3)을 이용하여 소설 텍스트의 특정 단어 $i$의 최종 TF-IDF 가중치를 구한다.

식(4)는 최종 TF-IDF의 가중치를 나타낸 것이다.

$$TFIDF_i = (w_h \times TF_{h,i} \times IDF) + (w_c \times TF_{c,i} \times IDF) + (w_d \times TF_{d,i} \times IDF) + (w_n \times TF_{n,i} \times IDF) \quad \cdots\cdots (4)$$

식 (4)의 $TFIDF_i$ 는 소설 텍스트에서 나온 단어의 최종 가중치를 나타낸다. $TF_{h,i}$는 머리말에서 등장한 단어의 TF값을, $TF_{c,i}$는 맺음말에서 등장한 단어의 TF값을, $TF_{d,i}$는 대화문에서 등장한 단어의 TF값을, $TF_{n,i}$는 비대화문에서 등장한 단어의 TF값을 나타낸다. 각 문서들의 데이터를 계산하고 특정 단어 $i$의 가중치 $TFIDF_i$를 계산하여 저장한다. IDF는 모든 문서의 집합이므로 특정단어 $i$의 IDF값은 반드시 존재한다. 이와는 반대로 특정 문서 $p$의 특정단어 $i$의 $TF_{p,i}$는 존재하지 않을 수 있다. 소설의 구조를 이용하였기 때문에 해당 문서에 존재하지 않는다면 특정단어 $i$는 주제어와 멀어진다.

계산된 TFIDF 가중치를 이용하여 오름차순으로 정렬한 후 상위 주제어 10개를 선정하여 주제어 DB에 저장한다.

## IV. 실험 및 평가

### 1. 실험 방법

본 연구의 실험은 앞서 설계한 주제어 추출 프로세스를 구축하여 진행되었다. 첫째, 머리말과 맺음말, 대화문과 비대화문의 형식적인 구조가 명확한 전자책 소설 데이터 100권을 구조적으로 분리하였다. 형태소 분석을 위해 각 구조는 문장 단위로 DB화 하였다. 둘째, 형태소 분석기를 이용하여 분리된 소설 데이터에서 각 명사, 태그, 빈도수를 추출하여 DB화 하였다. 셋째, 추출된 명사 데이터를 이용하여 IDF 문서를 생성하였다. 넷째, 명사 데이터와 IDF 문서를 이용하여 각 명사에 TF-IDF 가중치를 부여하였다.

### 2. 실험 검증

분석된 실험 결과를 검증하기 위해 본 연구와는 무관한 학생 10명이 수작업으로 뽑은 소설 텍스트의 주제어와 본 실험을 통해 추출한 상위 주제어 10개를 비교하여 일치율을 측정하였다.

표 1은 100권의 소설에 대한 주제어 추출 일치율 결과의 일부 보여주고 있다.

표 1. 주제어 일치율
table 1. rate of concordance of keywords

| 소설 텍스트 제목 | 일치율 |
|---|---|
| 1리터의 눈물 | 70% |
| 걸리버 여행기 | 60% |
| 꿈꾸는 책들의 도시 | 70% |
| 나 황진이 | 70% |
| 사라진 비둘기 | 40% |
| 하얀손길 | 40% |
| 동물농장 | 80% |
| 연금술사 | 80% |
| 연어이야기 | 60% |
| 아름다운 상실 | 70% |
| 데미안 | 80% |
| 렉스 | 80% |
| 수레바퀴 아래서 | 50% |
| 비스틀리 | 50% |

표 1은 소설의 머리말, 비대화문, 대화문, 맺음말의 가중치 비율을 각각 10:35:45:10으로 설정하고 실험한 결과의 일부를 나타낸 것이다.

학생 10명이 소설 〈동물농장〉으로부터 추출한 10개의 주제어는 '동물', '농장', '동무', '풍차', '돼지', '봉기', '혁명', '동지', '첩자', '인간'이며, 본 연구에서 실험을 통해 나온 상위 10개의 주제어는 가중치 순서대로 '동무', '동물', '농장', '풍차', '돼지', '우유', '봉기', '동지', '혁명', '침대'이다. 일치하는 주제어는 '동무', '동물', '농장', '풍차', '돼지', '봉기', '동지', '혁명'으로 80%의 일치율을 보였다. 하지만 소설 〈사라진 비둘기〉의 주제어 경우 40%의 낮은 일치율을 보였다.

이러한 결과는 소설 〈동물농장〉의 경우 대화문이 전체 소설 텍스트에서 차지하는 비율은 23.5%인 반면에 소설 〈사라진 비둘기〉의 대화문이 전체 소설 텍스트에서 차지하는 비율은 14.5%로 대화문보다 비대화문에 주제어를 더 많이 포함하고 것으로 풀이된다.

그림 5는 머리말과 맺음말의 가중치를 각각 0.15씩 부여하고 대화문의 가중치를 다르게 부여했을 때 나타난 실험 결

과를 그래프로 나타낸 것이다. 대화문의 가중치 비율은 5% 단위로 올리고, 이에 비례하여 비대화문의 가중치를 5% 단위로 낮추어 실험을 진행하였다. 그림 5에서 보여준 그래프에서 대화문의 가중치의 비율을 높게 할수록 주제어의 일치율이 올라가는 것을 확인 할 수 있었다. 대화문의 가중치 비율이 50%일때 주제어 일치율이 53.6%으로 최고점을 보였다. 대화문의 가중치 비율이 50%를 넘어가면 주제어 일치율이 하락하는 그래프를 보였다.

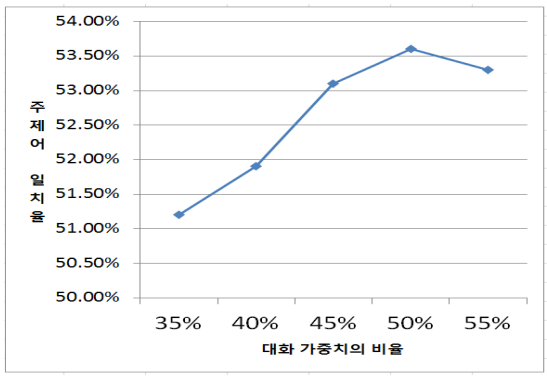


그림 5. $w_h + w_c = 0.3$ , $w_d + w_n = 0.7$
Fig 5. $w_h + w_c = 0.3$ , $w_d + w_n = 0.7$

그림 6은 머리말와 맺음말의 가중치를 각각 0.1씩 부여하고 대화문의 가중치를 5%씩 올리고, 이와 비례하여 비대화문의 가중치를 5%씩 감소시켰을 때 나타난 실험 결과이다.

마찬가지로 100권의 소설텍스트의 주제어 일치율을 비교하여 평균을 낸 결과, 대화문의 가중치의 비율을 45%로 부여하였을 때 주제어 일치율이 평균 62.1%로 최고점을 나타냈다. 그리고 대화문의 가중치를 0.5이상을 부여하였을 때 계속해서 감소하는 것을 보였다.

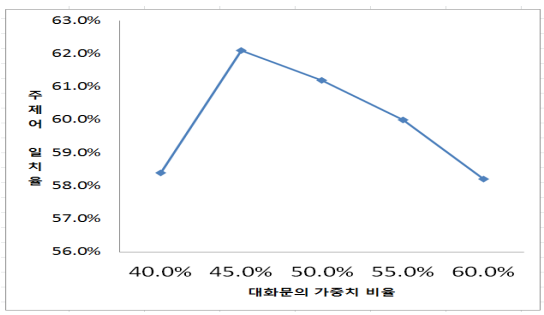


그림 6. $w_h + w_c = 0.2$ , $w_d + w_n = 0.8$
Fig 6. $w_h + w_c = 0.2$ , $w_d + w_n = 0.8$

## 3. 결과 분석

그림 7은 소설 구조를 고려한 주제어 추출 실험 2가지와 소설의 구조를 고려하지 않은 실험 결과를 그래프로 나타낸 것이다.

그림 7에서 보여준 실험 결과 "머리말 : 맺음말 : 대화문 : 비대화문"의 가중치 값을 "0.1 : 0.1 : 0.45 : 0.35"으로 부여했을 때 주제어 일치율이 62.1%로 가장 높았다. 이 결과는 그림 7에서처럼 소설의 구조를 고려하지 않고 본문 텍스트에서만 주제어를 추출했을 때 나온 20%의 일치율 보다 42.1% 더 높은 것으로 나타났다.

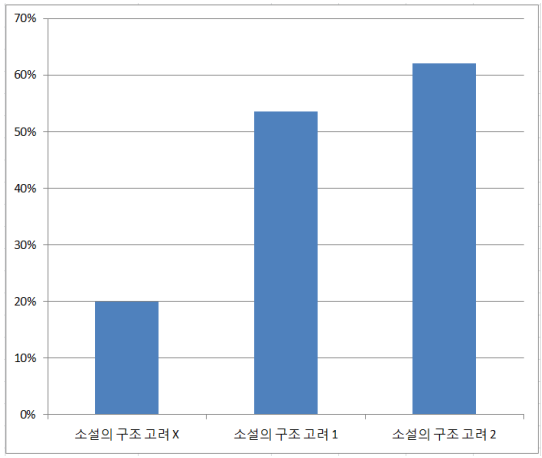


그림 7 소설 구조를 고려하지 않은 주제어 추출과 소설 구조를 고려한 주제어 추출 결과 비교
Fig 7. Do not take into account the main story structure extraction and Considering the main extraction novel structure

본 연구에서 제안한 주제어 추출 프로세스 실험결과를 통해 대화문의 가중치와 비대화문의 가중치를 똑같이 설정하였을 때보다 대화문의 가중치의 비율을 더 높게 하였을 때 주제어 추출 일치율의 정확도 더 높다는 것을 확인하였다.

하지만 대화문의 가중치를 높게 부여하면 일치율이 올라가지만, 그림 5와 그림 6에서 보듯이 대화문의 가중치가 일정 비율 이상 올라가게 되면 오히려 일치율이 더 낮아지는 결과를 확인하였다. 이는 소설의 구조에서 대화문이 주제를 드러내는데 유용한 역할을 하지만 머리말과 맺음말 그리고 비대화문도 주제어 추출에서 중요한 부분임을 의미한다.

## V. 결론 및 향후 연구

도서의 내용적 의미를 간과하고 도서의 서지 정보와 사용자 정보만을 이용하는 기존의 추천 시스템의 성능을 향상시키기 위한 선행연구로 본 연구는 TF-IDF와 소설 텍스트의 구조를 이용한 소설 텍스트의 주제어 추출 방법을 제안하였다.

우선 주제어 추출을 위해 소설 텍스트 100권을 수집하고 수집된 소설 텍스트들을 소설의 형식적 구조인 머리말, 맺음말, 대화문, 비대화문 4가지로 분리하였다. 그리고 각각의 구조에서 명사를 추출하고 명사, 태그, 빈도수를 DB형태로 저장한 후 저장된 명사 DB를 이용하여 IDF 문서를 생성하고 명사, 태그, 문서군에서의 출현빈도수를 DB형태로 저장하였다. 마지막으로 저장된 데이터를 이용하여 각 소설의 구조에 가중치 값을 다르게 부여하면서 상위 10개의 주제어를 선정한 후 학생 10명이 소설 텍스트에서 수작업으로 추출한 주제어 10개와 비교하여 일치율을 측정하였다.

실험 결과 소설 구조를 고려하지 않은 경우 주제어 일치율은 20%를 나타낸 반면 본 연구에서 제시한 소설 구조를 적용한 결과 일치율은 62.1%를 나타내어 42.1%의 성능향상을 보였다. 이러한 결과를 통해 소설에서 대화는 주제어를 추출하는데 있어 핵심적인 역할을 하고 있음을 확인할 수 있었다.

100권의 소설 중 주제어 일치율이 80%를 나타낸 소설도 있었지만 일치율이 40%를 나타내는 것도 있었다. 이는 대화에 대한 비율이 적거나 대화문보다는 비대화문에 주제를 내포한 것으로 해석된다.

본 연구에서는 주제어의 대상을 명사에 한정하였지만 소설의 장르적 특성 상 동사 역시 스토리를 전달하는데 중요한 역할을 하기 때문에 향후 연구에서는 주제어 추출에 동사도 함께 고려하는 연구를 진행 할 예정이다.

## Acknowledgement

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2014년도 문화기술연구개발지원사업의 연구결과로 수행되었음.

## REFERENCES

[1] S. G. Lee, H.-J. Kim, "Keyword Extraction from News Corpus using Modified TF-IDF", The Journal of Society for e-Business Studies, Vol.14, No.4, pp.59-73, 2009

[2] G.-S. Go, W.-K. Jung, Y.-G. Shin, S.-S. Park and D.-S. Jang, "A Study on Development of Patent Information Retrieval Using Textmining", Journal of the Korea Academia-Industrial cooperation Society, Vol.12, No.8, pp.3677-3688, 2011

[3] P. Soucy, G. W. Mineau, "Beyond TFIDF weighting for text categorization in the vector space model" In IJCAI, Vol. 5, pp. 1130-1135, 2005

[4] O. Zamir, O. Etzioni, O. "Grouper: a dynamic clustering interface to Web search results", Computer Networks, Vol.31, No.11, pp.1361-1374, 1999

[5] J. Martineau, T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", In Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media, 2009

[6] J. Ramos, "Using tf-idf to determine word relevance in document queries", In Proceedings of the First Instructional Conference on Machine Learning, 2003

[7] S.-P. Jung, S.-H. Lim, J.-H. Jeon, B. M. Kim and H. A. Lee, "Web Search Result Clustering using Snippets", Journal of KISS: Databases, pp.321-331, 2012

[8] H.-G. Choi, S. J. Jun, and E.-J. Hwang, "Multi-Modal Scheme for Music Mood Classification", Korea Information Science Society, pp.259-262, 2011

[9] H.I. Shin, U.I Yun, H.M. Ryang and G.B. Pyun, "An analytical Study for Extracting Topic Words on Text Documents", Korean Society For Internet Information, Vol.2011, No.6, pp.133-134, 2011

[10] S.-H. Jang, S.-S. Kang, "Keyword - based Document Clustering Algorithm", Korea Information Science Society. Vol.29, No.1B, pp.469-471, 2002

〔11〕 C.-H. Kim, Theory of the novel structure, Korean Studies Information, pp.16-17; 45-51; 203-204, 2010
〔12〕 H. S. Kim, "Types, Discourse Functions of Quotation and Speech Presentation in Novel", The Journal of Language and Literature, pp.113-142, 2000
〔13〕 www.kldp.net/projects/hannanum
〔14〕 GunHee. Choi, H-S. An, J-S. Park, "Main body of the text books extraction research", Proceedings of the Korea Inteligent Information System Society Conference pp.191-193, 2014

## 저 자 소 개

유 은 순
1995: 인하대학교 불어불문학과 문학사.
2000: Franche-Comté 대학교 언어학 석사.
2007: Franche-Comté 대학교 언어학 박사
현 재: 단국대학교 미디어콘텐츠연구원 리서치펠로우
관심분야: 스토리텔링, 추천시스템, 소셜미디어, 빅데이터, 감성
Email : tesniere@naver.com

최 건 희
현 재: 단국대학교 소프트웨어학과 학부생
단국-삼성 모바일 학부연구생
관심분야: 운영체제, IoT, 모바일 플랫폼
Email : rjsgmlgood@naver.com

김 승 훈
1985: 인하대학교 전자계산학과 공학사.
1989: 인하대학교 대학원 전자계산학과 공학석사.
1998: 포항공과대학교 대학원 컴퓨터공학과 공학박사
현 재: 단국대학교 응용컴퓨터공학과 교수
관심분야: 네트워크, IoT, 빅데이터, 추천시스템
Email : edina@dankook.ac.kr