JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 2, pp. 479-488, Feb. 2025

Check for updates

HTML 클러스터링과 심층 언어 모델을 활용한 전자책 템플릿 추천

장 동 호[1] · 서 정 헌[2] · 최 원 영[1] · 김 지 환[1] · 이 성 진[3] · 부 석 준[4] · 서 영 건[4*]
[1]경상국립대학교 컴퓨터공학과 학생 [2]경상국립대학교 AI융합공학과 학생
[3]경상국립대학교 소프트웨어공학과 교수 [4]경상국립대학교 컴퓨터공학과 교수

Recommendation of E-book Templates Using HTML Clustering and a Deep Language Model

Dong-Ho Jang[1] · Jeong-Heon Seo[2] · Ji-Hwan Kim[1] · Won-Young Choi[1] · Sungjin Lee[3] · Seok-Jun Buu[4] · Yeong Geon Seo[4*]

[1]Student, Department of Computer Engineering, Gyeongsang National University, Jinju 52828, Korea
[2]Student, Department of AI Convergence Engineering, Gyeongsang National University, Jinju 52828, Korea
[3]Professor, Department of Software Engineering, Gyeongsang National University, Jinju 52828, Korea
[4]Professor, Department of Computer Engineering, Gyeongsang National University, Jinju 52828, Korea

[요 약]

본 연구는 전자책 제작에 있어 템플릿 선택의 어려움을 해결하기 위해 HTML 기반의 전자책 템플릿 추천 시스템을 제안한다. 기존의 전자책 템플릿은 콘텐츠의 종류와 형식에 따라 레이아웃이 결정되지만 상대적으로 전자책 제작 경험이 부족한 편집자들이 콘텐츠 특성에 맞는 효과적인 템플릿을 선택하는 데 어려움이 있다. 본 연구에서는 유사도 기반 클러스터링 기법과 심층 언어 모델을 결합하여 원고의 특성에 맞는 전자책 템플릿을 추천하는 시스템을 개발하였다. 먼저, HTML 페이지의 구조적 유사성, 스타일 유사성, 원고 길이 유사성 등을 고려하여 템플릿을 클러스터링한 후 심층 언어모델을 사용해 원고의 콘텐츠 특성에 맞는 템플릿을 추천한다. 실제 전자책 출판 업계의 데이터셋을 활용한 실험을 통해, 제안된 추천 시스템이 실제 현장의 문제를 해결할 수 있는 가능성을 제시하였다.

[Abstract]

This study proposes an HTML-based e-book template recommendation system to address the challenges of selecting appropriate templates in e-book production. Traditional e-book templates determine layouts based on the type and format of the content; however, editors with limited e-book production experience often struggle to choose effective templates that align with the characteristics of content. In this study, we developed a system that integrates similarity-based clustering techniques with a deep language model to recommend e-book templates tailored to manuscript features. First, templates were clustered based on structural, stylistic, and manuscript-length similarities in HTML pages. Subsequently, a deep language model was employed to recommend templates best suited to the content characteristics of the manuscript. Experiments using datasets from the e-book publishing industry demonstrated the effectiveness of the proposed recommendation system in addressing real-world challenges in e-book production.

색인어 : 전자책, 템플릿 추천 시스템, HTML 클러스터링, 심층 언어모델, 유사도 기반 클러스터링
Keyword : E-book, Template Recommendation System, HTML Clustering, DLM, Similarity-based Clustering

http://dx.doi.org/10.9728/dcs.2025.26.2.479

Received 02 January 2025; **Revised** 31 January 2025
Accepted 11 February 2025

***Corresponding Author; Yeong Geon Seo**

Tel: +82-55-772-1392
E-mail: young@gnu.ac.kr

Ⅰ. 서 론

전통적인 종이책은 텍스트와 이미지를 중심으로 콘텐츠를 구성하며, 그 레이아웃은 이러한 요소들을 효과적으로 배치하는 데 초점을 맞추어 왔다[1]. 그러나 디지털 기술의 발전과 함께 전자책이 등장하면서 책의 형태와 콘텐츠의 표현 방식은 급격히 변화하였다. 종이책의 고정된 레이아웃이 갖는 한계를 넘어서 전자책은 텍스트, 이미지, 오디오, 동영상 등이 조화를 이루는 동적이고 상호작용적인 레이아웃을 구현할 수 있는 가능성을 열어주었다[2],[3].

이와 같은 변화 속에서 전자책 작가는 콘텐츠의 종류와 독자의 경험을 고려한 적절한 템플릿을 선택하는 것이 중요해졌다. 여기서 템플릿이란 콘텐츠의 특성에 맞게 디자인된 레이아웃과 스타일을 의미하며 전자책의 콘텐츠를 효과적으로 전달하고 독자에게 몰입감 있는 경험을 제공하는 데 중요한 역할을 한다. Han and Park[4]의 연구에 따르면 같은 내용의 전자책 콘텐츠라도 템플릿에 따라 독자의 경험이 크게 달라질 수 있다는 사실이 밝혀졌다. 이는 전자책 제작자들에게 적합한 템플릿 선택의 중요성을 잘 보여준다. 기존의 전자책 출판사들은 작가들이 전자책 제작 과정에서 겪는 어려움을 해결하기 위해 정기적으로 작가 교육을 실시하거나 카테고리별 예시 페이지를 제공하는 등의 방법을 사용해왔다. 그러나 이러한 방식은 전문가의 개입을 필요로 하므로 모든 작가에게 즉각적으로 지원을 하기에는 현실적인 제약이 있다.

본 연구에서는 이러한 문제를 해결하기 위해 클러스터링 기법과 심층 언어모델을 결합한 자동화된 템플릿 추천 시스템을 제안한다. 시스템은 전자책 데이터셋에서 페이지의 구조적 유사성, 스타일 유사성, 원고의 길이 등을 고려하여 템플릿을 클러스터링한 후 심층 언어모델을 사용해 원고의 특성에 맞는 최적의 템플릿을 추천한다. 이를 위해, 실제 전자책 출판업계의 데이터셋을 활용한 실험을 통해 제안된 시스템의 효과성을 테스트하고 평가했다. 실험 결과, 제안된 클러스터링 기법과 심층 언어모델을 결합한 시스템이 전자책 제작 과정에서 발생하는 템플릿 선택의 문제를 해결할 수 있다.

Ⅱ. 관련 연구

전자책 포맷은 크게 PDF, SVG, HTML 등으로 나눌 수 있으며 각 포맷은 그 특성에 따라 사용되는 목적과 방식이 다르다. 특히 HTML 전자책은 웹 기반의 기술을 활용하여 다양한 콘텐츠를 동적이고 상호작용적으로 표현할 수 있는 장점이 있다. 이러한 특성으로 인해 HTML 전자책은 텍스트, 이미지, 비디오, 오디오 등 다양한 미디어 형식을 손쉽게 결합할 수 있어 최근 전자책 출판에서 널리 사용되고 있다. 본 연구는 HTML 전자책에 초점을 맞추어 구조적 유사성과 스타일을 기반으로 한 자동화된 템플릿 추천 시스템을 제안한다.

2-1 HTML 콘텐츠 클러스터링

HTML 전자책에서는 콘텐츠가 HTML 마크업 언어로 구조화되므로 콘텐츠 간의 유사성을 분석하기 위해 HTML 문서의 구조적 특징과 스타일적 요소들을 고려하는 것이 중요하다. Gowda and Mattmann[5]은 웹 페이지 클러스터링을 활용하여 웹 콘텐츠의 유사성을 분석한 연구로 웹 페이지의 구조적 유사성과 스타일 유사성을 기반으로 각 페이지를 그룹화하고, 이를 통해 웹 콘텐츠를 사람의 개입 없이 효율적으로 분류할 수 있는 방법을 제시했다. 이 연구는 HTML 전자책 콘텐츠의 유사성을 분석하고 템플릿 추천 시스템에 적용할 수 있는 기초를 마련했다. Nair et al.[6]은 앞선 연구를 바탕으로 대규모 데이터셋에서도 효율적인 유사도 계산 방법을 제시했으며 HTML 문서 간 유사성을 계산 하는 데 LCS(Longest Common Subsequence) 방법을 사용하여 성능을 향상시켰다. LCS 방법은 두 HTML 문서 태그 간의 가장 긴 공통 부분 수열을 찾아내어 유사도를 계산하고, 이를 통해 콘텐츠 간의 유사성을 효율적으로 계산할 수 있었다.

이 두 연구는 웹 페이지 간의 유사성을 바탕으로 클러스터링 작업을 효율적으로 수행하는 방법들을 제시하고 있다. 하지만 이들 연구는 HTML 기반의 웹 페이지 콘텐츠를 대상으로 한 연구로 실제 HTML 전자책으로 확장 가능한지에 대한 검토가 필요하다. HTML 전자책은 일반 웹 페이지와 다르게 한 페이지에 담을 수 있는 콘텐츠 분량이 제한되어 있어 원고의 길이는 사용자 경험을 결정짓는 중요한 요소로 작용한다. 따라서 본 연구에서는 원고 길이와 같은 추가적인 요소를 클러스터링에 반영함으로써 전자책 콘텐츠의 그룹화가 보다 정교하게 이루어지도록 하였다.

2-2 심층 언어모델 기반 분류

최근 심층 언어모델(Deep Language Model, DLM)은 자연어 처리(NLP) 분야에서 큰 성과를 거두었으며 특히 문서 분류, 감정 분석 등 다양한 작업에서 뛰어난 성능을 보여주고 있다. 이러한 모델들은 대규모 텍스트 데이터를 학습하여 문맥을 이해하고 이를 바탕으로 주어진 텍스트의 의미를 정확하게 파악한다. Prasanthi et al.[7]은 트랜스포머 기반 모델을 사용하여 소셜 미디어 텍스트를 분류하는 연구를 수행했다. 이들은 BERT[8]와 RoBERTa를 활용하여 트위터 데이터를 분석했으며 해당 모델들이 비공식적인 문어체와 축약어, 철자 오류에 강인한 성능을 보임을 보고하였다. 특히, RoBERTa는 미세 조정 과정을 통해 감성 분류 작업에서 BERT보다 약 2~3% 높은 정확도를 기록하였다.

Rahman et al.[9]은 방글라어 텍스트 문서 분류를 위해 BERT와 ELECTRA[10] 모델을 활용한 성능 비교 실험을 진행했다. 세 가지 다른 방글라어 텍스트 데이터셋을 사용하여 모델을 실험했으며 두 모델 모두 사용된 세 가지 데이터셋 중 두 가지에서 우수한 성능을 보여주었다. 특히, ELECTRA

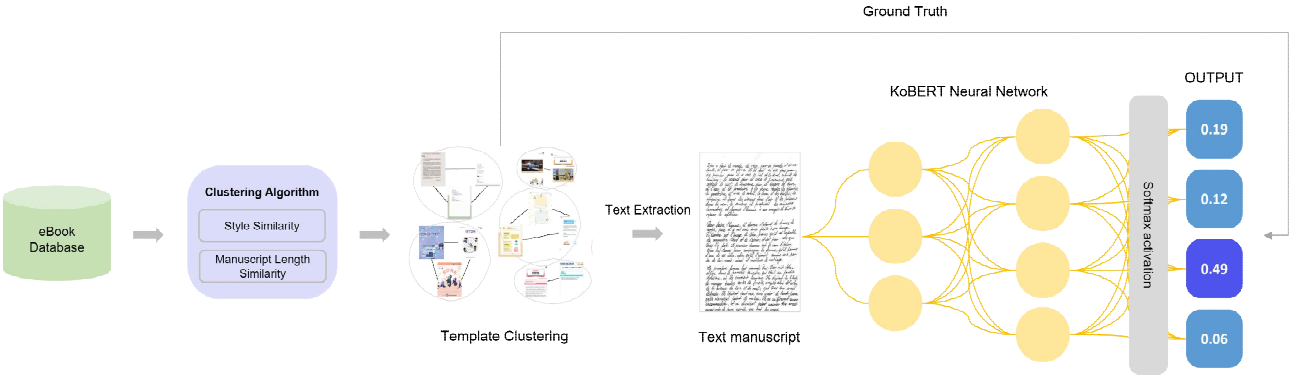


그림 1. 전자책 템플릿 추천을 위한 HTML 동적 클러스터링 및 심층 언어모델 결합 방법
Fig. 1. Proposed combining method of HTML dynamic clustering and DLMs for e-book template recommendation

모델은 데이터셋의 크기와 상관없이 전반적으로 더 높은 정확도와 F1 점수를 기록하며 우수한 성능을 발휘했다.

따라서, 본 연구에서도 BERT와 같은 트랜스포머 기반 모델을 활용하여 원고 텍스트의 의미와 문맥을 깊이있게 이해하고 이를 바탕으로 원고가 전달하고자 하는 정보를 효과적으로 표현할 수 있는 템플릿을 예측할 수 있도록 분류 모델로 채택했다.

Ⅲ. 제안된 전자책 추천 방법

본 연구에서는 전자책 템플릿 추천 시스템을 위해 유사도 기반 클러스터링과 심층 언어모델을 결합한 방법을 제안한다. 그림 1은 제안된 방법의 전체적인 흐름을 시각적으로 나타낸다. HTML 템플릿은 구조적 특성과 스타일을 기반으로 구조화되며 분류된 템플릿 데이터셋은 심층 언어모델이 학습하여 각 콘텐츠 유형에 최적화된 템플릿을 추천한다.

3-1 전자책 템플릿 구조화를 위한 HTML 클러스터 분류

제안된 방법은 전자책 템플릿 구조화를 위해 유사성 기반의 클러스터링 접근 방식을 채택하였다. 이 과정에서 HTML 페이지의 구조적, 스타일적 유사성은 물론 원고 길이 유사성까지 종합적으로 고려하여 템플릿을 분류한다. 유사성 계산은 전자책 페이지의 HTML 태그와 CSS 클래스를 기준으로 이루어지며, 이를 통해 각 템플릿을 동적으로 클러스터링한다. E. Marca의 github 저장소[11]에서는 이러한 유사도를 계산하기 위한 알고리즘을 제공하고 있다.

1) 구조적 유사성

구조적 유사성은 두 HTML 문서 태그 간의 Longest Common Subsequence 기반의 유사도 측정을 활용하여 HTML 문서의 계층적 특성을 반영하면서 대규모 데이터셋에서도 효율적으로 유사도를 계산할 수 있다. 구조적 유사성은 두 HTML 문서의 태그 집합을 A와 B로 정의하고 이를 바탕으로 (1)과 같이 계산된다.

$$S_{structure} = \frac{2 \times LCS(A,B)}{|A|+|B|} \quad (1)$$

2) 스타일 유사성

스타일 유사성은 두 전자책 페이지 간의 CSS 클래스 요소를 기반으로 계산된다. 전자책 템플릿에서 스타일은 페이지 레이아웃을 결정하는 중요한 요소로 동일한 템플릿을 사용하는 페이지들은 종종 동일한 스타일을 공유한다. 스타일 유사성은 두 HTML 문서의 클래스 집합을 A와 B로 정의하고 이를 바탕으로 (2)와 같이 계산된다. |A∩B|는 공통 클래스 요소의 개수를 의미한다.

$$S_{style} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \quad (2)$$

3) 원고 길이 유사성

원고 길이 유사성은 두 전자책 원고 간의 원고 텍스트 길이를 비교하여 콘텐츠의 유사성을 평가하는 지표이다. 같은 주제나 비슷한 내용의 전자책 페이지는 원고 길이가 비슷할 가능성이 크다. 원고 길이 유사성은 두 HTML 문서에서 텍스트의 길이를 추출하고 이를 바탕으로 (3)과 같이 계산된다.

$$S_{\leq ngth} = 1 - \frac{|L_1 - L_2|}{\max(L_1, L_2)} \quad (3)$$

4) 종합 유사성

종합 유사성은 구조적 유사성, 스타일 유사성, 원고 길이 유사성을 모두 고려하여 두 페이지 간의 전반적인 유사도를 평가한다. 각 유사성 항목은 가중치를 부여하여 선형 조합된 후 최종적인 페이지 유사도가 계산된다. 종합 유사성은 (4)와 같이 계산된다.

$$S_{total} = k_1 \cdot S_{structure} + k_2 \cdot S_{style} + k_3 \cdot S_{\leq ngth} \quad (4)$$

5) 클러스터링 방법

제안 방법에서는 전자책 템플릿을 효율적으로 분류하기 위해 종합 유사성 계산을 기반으로 한 클러스터링 접근 방식을 제안한다. 이를 구현하기 위해 다음과 같은 절차를 설계하였다. 먼저, 템플릿 데이터셋의 모든 HTML 페이지 쌍에 대해 종합 유사도를 계산하여 N×N 크기의 유사성 행렬로 표현하였다. 계산의 효율성을 위해 상삼각 영역에 유사도 값을 삽입한 후 하삼각 영역으로 복사하여 대칭 유사성 행렬을 완성하고 자기 자신과의 유사도를 나타내는 대각선 요소를 1로 할당하였다.

이후, 유사성 행렬에 대해 t-SNE를 사용하여 차원 변환을 수행한 다음, K-Means 알고리즘을 통해 클러스터링을 진행한다. t-SNE는 매니폴드 러닝 기법 중 하나로 고차원 데이터 내의 구조적 특징을 보존하면서 이를 인간이 이해할 수 있는 저차원 공간으로 변환하는 방법이다. 이를 통해 데이터의 군집 특성을 시각적으로 분석하고 K-Means를 적용하여 각 데이터가 속한 클러스터를 효과적으로 구분할 수 있다[12],[13].

3-2 심층 언어모델을 통한 템플릿 추천

심층 언어모델은 앞서 생성된 템플릿 클러스터를 바탕으로 템플릿 부류를 예측하는 방식으로 작동한다. 이 모델은 사용자로부터 입력받은 원고 텍스트를 임베딩하여 분류기에서 텍스트에 매핑되는 템플릿을 예측하는 역할을 한다.

심층 언어모델은 이를 위해 클러스터링을 통해 구조화된 전자책 데이터셋에서 각 페이지의 원고 텍스트와 해당 페이지가 가지는 클러스터의 번호를 입출력 쌍으로 받아 학습을 진행한다. 학습 과정에서 모델은 입력된 원고 텍스트의 문맥, 구조적 특징을 바탕으로 적합한 템플릿 부류를 예측할 수 있는 패턴을 학습한다. 훈련이 완료된 모델은 사용자로부터 입력받은 새로운 원고 텍스트를 분석하고, 이를 통해 가장 적합한 템플릿 부류를 출력한다.

Ⅳ. 실험 및 결과

4-1 데이터셋

본 연구에서는 전자책 템플릿 분류 및 군집화 성능을 평가하기 위해 실제 출판된 전자책 HTML 데이터를 사용하여 테스트 및 평가되었다. 이 전자책 데이터셋은 69권의 전자책으로 구성되어 있으며, 총 1,000개의 페이지를 포함하고 있다. 전자책의 주요 컨텐츠 특성은 표 1에 나타나 있다. 표에서 볼 수 있듯이 전자책의 템플릿은 텍스트 기반 콘텐츠부터 이미지, 오디오, 동영상, 상호작용 콘텐츠까지 다양한 유형을 수용할 수 있는 선택지를 제공한다.

4-2 실험 설정

전자책 템플릿의 추천 및 클러스터링 성능을 평가하기 위해 여러 평가 지표를 사용하였으며 구체적인 평가 지표로는 실루엣 점수(Silhouette Score), 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1-score) 등을 사용하였다.

$$Sillhouette-Score=\frac{(b_i+a_i)}{(\max(a_i,\ b_i))} \tag{5}$$

$$Accuracy=\frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{6}$$

$$\mathrm{Precision}=\frac{TP}{(TP+FP)} \tag{7}$$

$$Recall=\frac{TP}{(TP+FN)} \tag{8}$$

$$F1=\frac{(2\cdot \mathrm{Precision}\cdot Recall)}{(\mathrm{Precision}\cdot Recall)} \tag{9}$$

다음으로, 전자책 템플릿 군집화 및 추천 성능을 평가하기 위해 다양한 파라미터 설정을 실험하였다. 주요 파라미터는 종합 유사성 가중치 k1, k2, k3로 템플릿 군집을 조정하였다.

마지막으로, 전자책 추천 성능을 평가하기 위해 트랜스포머 기반의 심층 언어모델인 KoBERT를 사용하여 성능을 평가하였다. 이때, 모델이 입력값으로 사용할 원고 텍스트는 전처리 과정을 거쳐 HTML에서 추출된 텍스트로 단어 임베딩 기법을 통해 벡터화되었다.

4-3 실험 결과

1) 단일 유사성 클러스터링 성능 평가

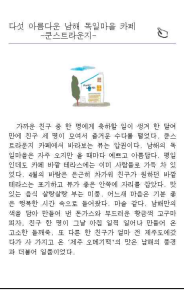
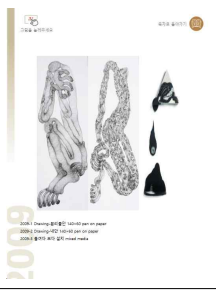
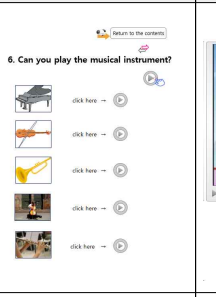
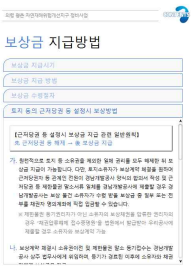
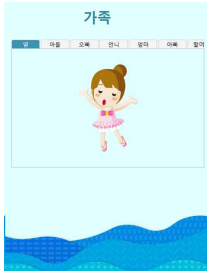
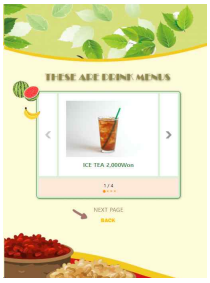
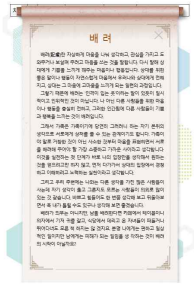
전자책 템플릿 클러스터링의 성능을 평가하기 위해 단일 유사성 평가를 실시하였으며, 이를 통해 각 유사성 기준이 전자책 템플릿 클러스터링에 미치는 영향을 파악하였다.

구조적 유사성은 페이지 간의 구조적 차이를 충분히 반영하지 못한 채 퍼져 있는 형태로 나타났다(그림 2-A 참고). 이는 전자책 템플릿에서 페이지 구조의 다양성으로 인해 유사성을 명확히 구분하기 어려운 경향을 보였으며, 기존 웹 페이지 클러스터링에 사용된 구조적 유사성 알고리즘이 전자책 도메인에는 적합하지 않음을 시사한다.

스타일 유사성은 구조적 유사성과 비교했을 때 상대적으로 군집화된 형태로 나타났으며, 이는 전자책 템플릿의 디자인 요소들이 일정한 패턴을 따르며 고르게 분포하고 있음을 시사한다(그림 2-B 참고). 이는 글상자나 이미지와 같은 각 페이지의 스타일 구성 요소들이 반복적으로 나타나거나 균등하게 배치되어 있어 클러스터링 과정에서 스타일 유사성에 기반한 페이지들이 일정한 그룹을 형성하는 경향을 보였음을 의미한다.

원고 길이 유사성은 M자 모양으로 나타났으며 원고 길이에 따라 템플릿들이 일정한 패턴을 따르기보다는 일정한 범

표 1. 전자책 주요 콘텐츠 특성 및 클러스터링 결과
Table 1. Key characteristics of e-book content and clustering results

Template Number	0	1	2	3	4
Cover Image					
Template Name	Plain text	Scrolling text	Image	Audio	Movie
Key Characteristics	A simple and intuitive layout for text-based content	A layout where long text scrolls naturally	Optimization of image content size and placement	A layout suitable for audio content	A layout suitable for video content
Included Content Types	General text, descriptions, table of contents	Novels, essays, explanations	Photos, illustrations, visual elements	Audiobooks, interviews, lectures	Videos, interactive media
Cluster Count	95	92	20	180	69
Template Number	5	6	7	8	9
Cover Image					
Template Name	Accordion	Tab	Slider	Quiz	Popup
Key Characteristics	A dynamic layout that allows sections to be collapsed or expanded	A layout that easily switches between multiple contents using tabs	A layout that allows content to be transitioned by sliding	A layout suitable for interactive quiz content	A layout with pop-up windows for interaction
Included Content Types	FAQs, tutorials, lists	Documents, lists, multi-category content	Images, galleries, presentations	Quizzes, surveys, tests	Notifications, additional information
Cluster Count	172	24	47	48	79

*We don't translate into English for the accurate conveyance of the word's meaning

위 내에서 분포하는 경향을 보임을 의미한다(그림 2-C 참고). 결과적으로 스타일 유사성이 가장 높은 클러스터링 성능을 보였으며, 구조적 유사성은 상대적으로 낮은 성능을 보였다. 이는 전자책 템플릿을 분류할 때 스타일 특성을 기준으로 한 분류가 효과적일 수 있음을 시사하며 향후 템플릿 추천 시스템에서 스타일 유사성을 중요한 기준으로 삼을 수 있다는 직관을 제공한다.

2) 복합 유사성 클러스터링 성능 평가

다음으로 유사성의 가중치를 복합적으로 조합하여 실험을 진행하였다. 복합 유사성 평가는 여러 유사성 기준을 결합하여 템플릿 클러스터링의 성능을 평가하는 방법이다. 복합 유사성 평가에서는 구조적 유사성, 스타일 유사성, 원고 길이 유사성을 다양한 조합으로 결합하여 실험을 진행했다. 실험 결과는 그림 3에 나타나 있으며, 각 조합이 클러스터링 성능에 미친 영향을 분석하였다.

먼저, 구조적 유사성과 스타일 유사성을 결합한 경우 스타일 유사성만을 사용할 때보다 클러스터링 성능이 향상되었다(그림 3-D 참고). 이는 구조적 유사성이 페이지 구조의 세부적인 차이를 보완하며 스타일 유사성과 함께 조화를 이루어 페이지들을 더 효과적으로 그룹화할 수 있음을 의미한다. 하지만, 이 경우 원고 길이 유사성이 포함되지 않아서 발생하는

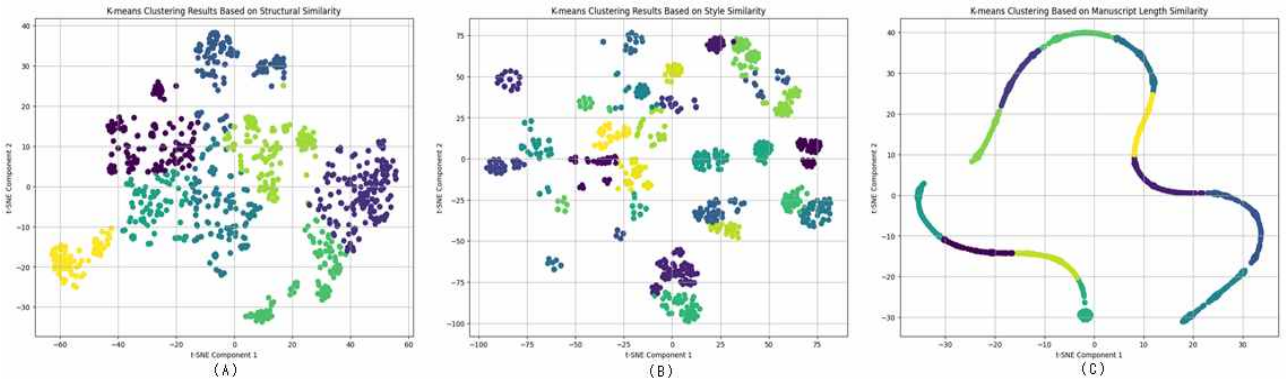


그림 2. 단일 가중치 기반의 유사성 평가(왼쪽부터 구조, 스타일, 원고 길이 유사성 순)
Fig. 2. Single weight-based similarity evaluation(from left to right: structural, style, and manuscript length similarity)

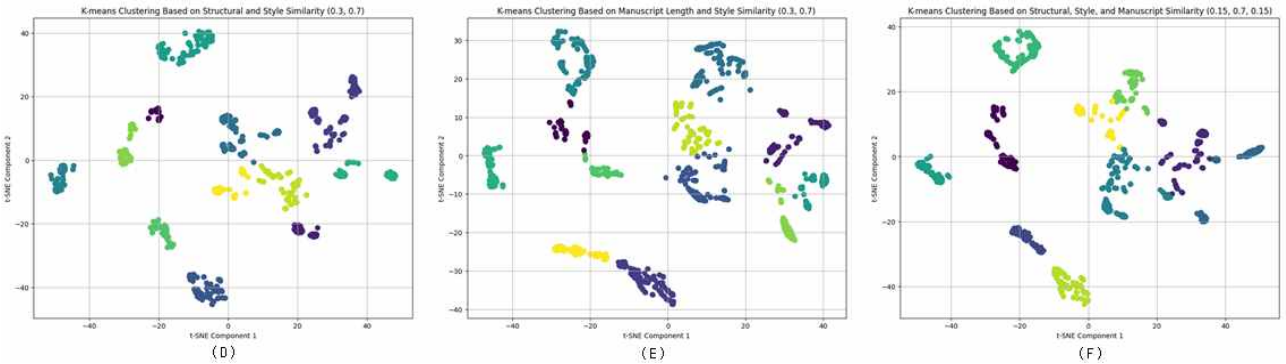


그림 3. 복합 가중치 기반의 유사성 평가(왼쪽부터 구조+스타일, 스타일+원고 길이, 구조+스타일+원고 길이 유사성 순)
Fig. 3. Composite weighted similarity evaluation (from left to right: structure + style, style + manuscript length, structure + style + manuscript length similarity)

문제점이 있다. 원고 길이 유사성은 템플릿 내에서 텍스트의 길이나 분량에 따라 차이를 반영하는 중요한 요소인데 이를 고려하지 않아 스타일이 같더라도 분량이 유사한 페이지들이 서로 다른 클러스터로 분류되었다. 따라서 원고 길이 유사성을 추가하여 템플릿의 분량에 따른 특성을 종합적으로 고려하는 것이 필요하다.

다음으로, 스타일 유사성과 원고 길이 유사성을 고려한 경우 구조적 유사성과 스타일 유사성을 결합한 경우 보다 상대적으로 퍼져 있는 경향을 보였다(그림 3-E 참고). 스타일 유사성은 디자인 요소에 따른 유사성을 반영하며 페이지들을 그룹화하지만, 원고 길이 유사성은 페이지의 텍스트 분량에 따라 페이지들을 더 넓은 범위로 분산시킨다. 이로 인해 페이지들이 더 세밀하게 분류되기보다는 넓은 범위에서 분포하는 경향이 나타났다.

마지막으로, 구조적 유사성, 스타일 유사성, 원고 길이 유사성 세 가지를 모두 결합한 경우 군집화가 상대적으로 더 집중되면서도 여전히 일부 페이지들이 넓은 범위에 분포하는 경향을 보였다(그림 3-F 참고). 이 경우 세 가지 유사성 기준이 상호 보완적으로 작용하면서 템플릿의 다양한 특성이 고려되었기 때문에 구조적 유사성과 스타일 유사성의 결합보다는 더 퍼져 있는 경향을 보였지만, 스타일 유사성과 원고 길이 유사성의 결합보다는 더 군집화된 결과가 나타났다.

3) 심층 언어모델 분류 성능 평가

복합 유사성 평가와 함께 심층 언어모델을 사용하여 템플릿 분류 성능을 평가하였다. 실험에서는 복합 유사성 평가에서 도출된 다양한 유사성 기준을 바탕으로 클러스터링된 전자책 페이지의 원고들과 클러스터 번호가 모델의 입력값과 출력값으로 사용되었다. 이 과정에서 모델의 목표는 주어진 원고를 입력받아 해당 원고가 속할 클러스터 번호를 정확하게 예측하는 것이다. 평가 기준으로는 micro F1 스코어를 사용하여 모델의 성능을 분석하였다. 실험 결과는 표 2에 나타나 있으며, 각 조합이 심층 언어모델 성능에 미친 영향을 분석하였다.

먼저, 스타일 유사성 가중치에 0.7, 원고 길이 유사성에 0.3을 부여한 경우 정확도는 75.94%, micro F1 스코어는 75.50%로 가장 우수한 성능을 보였다(표 2-G 참고). 이 조합은 스타일 유사성의 비중을 높게 설정하여 템플릿 디자인 요소의 유사성을 강하게 반영하면서도 원고 길이 유사성의 요소를 적절히 결합하여 템플릿의 텍스트 분량 차이를 일정 부분 고려한 결과로 해석된다. 다만, 스타일 유사성에 가중치를 0.8로 높게 부여한 실험에서는 정확도와 micro F1 스코어

표 2. 유사성 가중치 조합에 따른 클러스터링 성능 지표와 심층 언어모델 성능 지표
Table 2. Similarity weight combinations based clustering performance metrics and DLM performance metrics

Type	Index	Similarity Weight			Clustering Metrics	Deep Language Model Metrics			
		Structural	Style	Manuscript Length	Mean Silhouette Score	Accuracy	Precision	Recall	F1
Single Similarity	A	1	0	0	**0.18**	72.53%	73.61%	72.53%	71.26%
	B	0	1	0	0.39	71.00%	72.54%	71.00%	71.00%
	C	0	0	1	**0.51**	47.00%	47.72%	47.00%	**47.00%**
Composite Similarity	D	0	0.5	0.5	0.26	67.00%	70.29%	67.00%	67.00%
	E	0.3	0.7	0	0.38	70.50%	74.39%	70.50%	70.50%
	F	0	0.3	0.7	0.38	64.00%	66.15%	64.00%	64.00%
	G	0	0.7	0.3	0.34	75.94%	78.42%	75.50%	**75.50%**
	H	0	0.8	0.2	0.39	69.50%	72.54%	69.50%	69.50%
	I	0.1	0.8	0.1	0.41	71.00%	71.92%	71.00%	71.00%
	J	0.15	0.7	0.15	0.37	70.00%	73.63%	70.00%	70.50%

가 69.50%로 나타났다(표 2-H 참고). 이 경우 스타일 유사성의 가중치가 더 높아짐에 따라 분류 모델이 혼동을 겪는 것으로 분석된다.

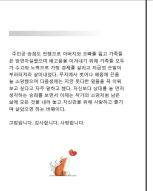
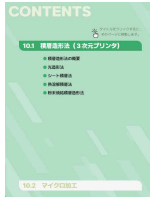
또한, 세 가지 유사성 기준을 모두 결합한 실험에서는 정확도와 F1 스코어가 71.00%를 기록하였다(표 2-I 참고). 이 결과는 각 유사성 요소들이 상호 보완적으로 작용하는 방식으로 클러스터링과 심층 언어모델의 성능을 높이려는 전략이었으나 스타일 유사성과 원고 길이 유사성만을 사용했을 때보다 상대적으로 성능이 낮은 경향을 보였다. 이는 모든 유사성 기준이 결합되면서 각 기준의 특성이 지나치게 분산되어 분류 성능에 혼란을 초래할 수 있음을 의미한다. 따라서, 가장 우수한 성능을 발휘한 스타일 유사성과 원고 길이 유사성의 조합은 템플릿 클러스터링에서 템플릿의 디자인 요소와 텍스트 분량의 균형을 잘 맞춘 것으로 향후 템플릿 분류 시스템에서 중요한 기준이 될 수 있음을 알 수 있다. 이와 관련된 클러스터링 결과는 표 1에 제시되어 있다.

4) 템플릿 추천 시스템의 실제 활용 예시

본 연구에서 제안한 템플릿 추천 시스템의 실제 동작 과정을 보여주기 위해, 전자책 제작 과정에서 사용되는 대표적인 원고를 입력으로 사용하여 시스템의 활용 사례를 실험적으로 검증하였다. 실험에서는 짧은 텍스트 중심 원고, 이미지 중심 원고, 정보가 많은 텍스트 중심 원고에 대해 템플릿 추천 시스템을 적용하였다.

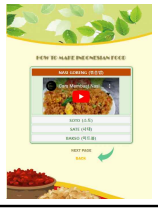
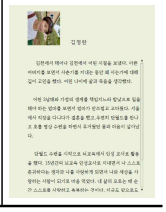
추천 시스템은 각 콘텐츠에 가장 적합한 레이아웃을 제공하는 템플릿을 추천하였다. 표 3의 짧은 텍스트 중심 원고에는 텍스트 기반 콘텐츠를 위한 간단하고 직관적인 레이아웃인 Plain Text, Scrolling Text, Accordion 순으로 템플릿을 추천하였고, 표 4의 이미지 중심 원고에는 이미지와 텍스트가 적절히 배치될 수 있는 Accordion, Plain Text, Scrolling Text 순으로 추천하였다. 또한, 표 5의 정보가 많은 텍스트 중심 원고에는 텍스트의 양을 고려하여 내용이 가독성 있게 배치되는 Scrolling Text, Tab, Accordion 순으

표 3. 전자책 콘텐츠 예시: 짧은 텍스트 중심 원고
Table 3. E-book content example: Short text-based manuscript

Input Data Example	The global economy has been rapidly changing since the early 20th century. The IT revolution from the late 1990s to the early 2000s caused…		
Template Name	Plain Text	Scrolling text	Accordion
Top-3 Probabilities	99.49%	0.18%	0.14%
Cover Image			

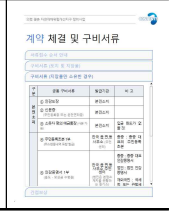
*We don't translate into English for the accurate conveyance of the word's meaning

표 4. 전자책 콘텐츠 예시: 이미지 중심 원고
Table 4. E-book content example: Image-based manuscript

Input Data Example	KOREAN FOOD MENUS Bibimbap 8,500Won Kimchi Stew 7,000Won Bulgogi 9,500Won		
Template Name	Accordion	Plain Text	Scrolling text
Top-3 Probabilities	95.51%	2.79%	0.48%
Cover Image			

*We don't translate into English for the accurate conveyance of the word's meaning

표 5. 전자책 콘텐츠 예시: 정보가 많은 텍스트 중심 원고
Table 5. E-book content example: Text-heavy manuscript

Input Data Example	A smart city refers to an urban model that uses advanced technologies to enhance the quality of life in cities. In recent years, smart cities have gained significant global attention, with various cities announcing plans for smart city development. For example, Singapore is aiming to achieve goals such as alleviating traffic congestion, enhancing public safety, and improving energy efficiency through its smart city project. Technologies such as IoT, AI, big data, and 5G play a crucial role in smart city development, as these technologies are integrated to solve various problems and improve the efficiency of urban management. Smart city projects involve not only technical aspects but also important policy and social challenges. Issues such as data security, privacy concerns, and the updating of public infrastructure are key challenges to be addressed during the smart city development process. The future of smart cities should focus on creating sustainable environments and improving the quality of life for citizens.		
Template Name	Scrolling text	Tab	Accordion
Top-3 Probabilities	92.88%	2.46%	1.95%
Cover Image			

*We don't translate into English for the accurate conveyance of the word's meaning

로 템플릿을 추천하였다.

제안 방법에서는 전자책 템플릿 분류 및 클러스터링 문제를 해결하기 위해 심층언어 모델과 유사성 기반의 클러스터링 방식을 결합한 새로운 접근법을 제안하였다. 이를 통해 전자책 템플릿을 분류하는데 있어 사람의 개입 없이 자동화된 방식으로 템플릿을 효과적으로 클러스터링할 수 있다는 것을 입증하였다. 그러나 본 연구에서 제안한 방법은 유사도 행렬을 생성하는 과정에서 계산 복잡도가 n 제곱에 도달하여 대규모 데이터셋에 적용하기 어려운 한계가 있다. 이는 클러스터링 알고리즘의 확장 가능성 및 효율성 측면에서 개선이 필요한 부분으로 향후 연구에서는 보다 효율적인 동적 클러스터링 기법을 도입하여 이러한 문제를 해결할 필요가 있다.

V. 결 론

본 연구에서는 전자책 제작 과정에서 겪는 템플릿 선택의 문제를 해결하기 위한 직관적인 템플릿 추천 시스템을 제안하였다. 기존의 전통적인 디자인 방식에 비해 동적이고 멀티미디어가 결합된 전자책 디자인의 특성을 고려하여 심층 언어모델과 유사성 기반 클러스터링 기법을 활용한 자동화된 템플릿 추천 시스템을 개발하였다. 이 시스템은 전자책 제작자들이 원고의 콘텐츠 특성과 구조에 맞는 적합한 템플릿을 직관적으로 선택할 수 있도록 도울 수 있다.

연구의 주요 결과는 클러스터링을 통해 전자책의 다양한 콘텐츠 유형을 사람의 개입 없이 효과적으로 그룹화하고 심층 언어모델을 활용하여 사용자의 요구에 맞는 추천할 수 있음을 보였다. 이를 통해 전자책 제작자들이 템플릿을 선택하는 데 있어 혼란을 겪거나 디자인상의 실수를 범하지 않도록 가이드라인을 제공할 수 있는 가능성을 보였다. 또한, 실제 전자책 출판 업계에서 사용되는 데이터셋을 활용한 실험을 통해 시스템의 실용성을 검증하고 현장에서 발생하는 문제들을 해결할 수 있는 방법을 제시하였다.

본 연구의 시사점은 다음과 같다. 첫째, 전자책 제작 과정에서 인공지능 기술을 활용한 자동화된 추천 시스템의 가능성을 입증하였다는 점에서 학문적 의의가 있다. 둘째, 전자책 출판 산업에서 디자인의 일관성을 향상시킬 수 있는 실질적인 도구를 제시함으로써 산업적 가치를 창출하였다. 셋째, 원고가 전달하고자 하는 정보를 효과적으로 표현할 수 있는 템플릿을 편집자에게 추천함으로써 전자책 제작의 진입 장벽을 낮추는 데 기여할 수 있다는 점에서 실무적 의의가 있다.

향후 연구에서는 템플릿 추천 시스템의 확장성과 효율성을 위해 동적 클러스터링 기법을 도입하고, HTML 외에도 PDF나 SVG와 같은 전자책 포맷을 지원할 수 있는 방법을 모색해야 한다. 또한, 심층 언어모델과 이미지, 비디오 처리 모델을 결합한 멀티모달 학습을 통해 텍스트, 이미지, 동영상을 동시에 분석하여 전자책의 동적 특성을 반영한 추천 시스템을 개발할 필요가 있다.

감사의 글

본 결과물은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구결과입니다. 또한 아라소프트의 지원에 의하여 이루어진 연구로서 연구 과정에서 아라소프트의 기술적 지원과 자원을 제공해주신 모든 관계자들께 감사 드리며, 전자책 제작에 귀중한 데이터와 피드백을 제공해주신 전자책 출판 업계의 전문가들과 참가자들께도 감사드립니다.

참고문헌

[1] M. Kim, A Study on the Role of the Layout on the Editorial Design -Focusing on Typography Magazine-, Master's Thesis, Chosun University, Gwangju, February 2014.
[2] H. Kim, "Electronic-Book Design Process through an

Investigation into History and Design Elements of the Book -On Form and Function with Consideration of Product-Environment Design-," *Journal of Korea Design Forum*, No. 31, pp. 77-86, May 2011. https://doi.org/10.21326/ksdt.2011..31.007

[3] H. J. Park and W. J. Choi, "A Study on the Systematic Implementation of e-Book Design," *A Journal of Brand Design Association of Korea*, Vol. 13, No. 1, pp. 131-146, March 2015. https://doi.org/10.18852/bdak.2015.13.1.131

[4] H.-W. Han and K.-E. Park, "Experientiality and Reading Experience in e-book," *Journal of the Korea Contents Association*, Vol. 11, No. 12, pp. 171-181, December 2011. https://doi.org/10.5392/JKCA.2011.11.12.171

[5] T. Gowda and C. A. Mattmann, "Clustering Web Pages Based on Structure and Style Similarity," in *Proceedings of 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, Pittsburgh: PA, pp. 175-180, July 2016. https://doi.org/10.1109/IRI.2016.30

[6] V. V. Nair, M. van Staalduinen, and D. T. Oosterman, "Template Clustering for the Foundational Analysis of the Dark Web," in *Proceedings of 2021 IEEE International Conference on Big Data*, Orlando: FL, pp. 2542-2549, December 2021. https://doi.org/10.1109/BigData52589.2021.9671936

[7] K. N. Prasanthi, R. E. Madhavi, D. N. S. Sabarinadh, and B. Sravani, "A Novel Approach for Sentiment Analysis on Social Media Using BERT & ROBERTA Transformer-Based Models," in *Proceedings of 2023 IEEE 8th International Conference for Convergence in Technology (I2CT),* Lonavla, India, pp. 1-6, April 2023. https://doi.org/10.1109/I2CT57861.2023.10126206

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805v1, October 2018. https://doi.org/10.48550/arXiv.1810.04805

[9] M. M. Rahman, M. A. Pramanik, R. Sadik, M. Roy, and P. Chakraborty, "Bangla Documents Classification Using Transformer Based Deep Learning Models," in *Proceedings of 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI),* Dhaka, Bangladesh, pp. 1-5, December 2020. https://doi.org/10.1109/STI50764.2020.9350394

[10] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-Training Text Encoders as Discriminators rather than Generators," arXiv:2003.10555, March 2020. https://doi.org/10.48550/arXiv.2003.10555

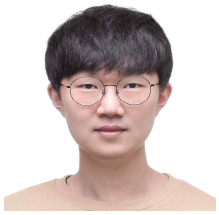
[11] GitHub. Html-Similarity [Internet]. Available: https://github.com/matiskay/html-similarity.

[12] L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, Vol. 9, No. 11, pp. 2579-2605, 2008.

[13] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, Vol. 9, No. 8, 1295, August 2020. https://doi.org/10.3390/electronics9081295

장동호(Dong-Ho Jang)

2019년~현 재: 경상국립대학교 컴퓨터공학과 재학
※관심분야 : AI, 머신러닝, LLM

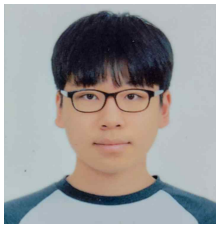


서정헌(Jeong-Heon Seo)

2013년 : 경상대학교 컴퓨터과학과(학사)
2021년 : 경상대학교 AI융합공학과(석사)
2023년 9월~현 재: 경상국립대학교 AI융합공학과 박사
※관심분야 : 멀티모달, AI, Agent, LLM

김지환(Ji-Hwan Kim)

2020년~현 재: 경상국립대학교 컴퓨터공학과 재학
※관심분야 : AI, 머신러닝, LLM



최원영(Won-Young Choi)

2020년~현 재: 경상국립대학교 컴퓨터공학과 재학
※관심분야 : AI, 머신러닝, LLM

이성진(Sungjin Lee)

2006년 : 한양대학교 컴퓨터공학과 졸업(공학사)
2015년 : 한양대학교 컴퓨터공학과 졸업(공학박사)
2017년 3월～현 재: 경상국립대학교 소프트웨어공학과 부교수
2024년～현 재: 경상국립대학교 정보전산처장
※관심분야 : AI, 머신러닝, LLM, 항공SW, 시스템SW

부석준(Seok-Jun Buu)

2023년 : 연세대학교 컴퓨터과학과 졸업(공학박사)
2023년 9월～현 재: 경상국립대학교 컴퓨터공학과 조교수
※관심분야 : AI, 머신러닝, LLM, 사이버보안, 뉴로-심볼릭 인공지능

서영건(Yeong Geon Seo)

1987년 : 경상대학교 전산과(이학사)
1997년 : 숭실대학교 전산과(공학박사)
1989년～1992년: 삼보컴퓨터
2022년～2024년: 경상국립대학교 정보전산처장
1997년～현 재: 경상국립대학교 컴퓨터공학과 교수
※관심분야 : 컴퓨팅 사고, SLAM, AI, 컴퓨터 네트워크