Smart Media Journal / Vol.14, No.10 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.10.16

A!rrange: AI 기반 스마트 북마크 관리 브라우저 확장 프로그램

(A!rrange: AI-based Smart Bookmark Management Browser Extension)

김정우*, 서재오*, 임한빈*, 김미수**
(Jeong Wu Kim, Jae Oh Seo, Han Bin Im, Mi Soo Kim)

요 약

본 논문에서는 AI 기반의 스마트 북마크 관리 시스템인 “A!rrange”를 제안한다. 본 시스템은 웹페이지 내용을 자동으로 분석하여 적절한 제목을 생성하고, 의미 기반 클러스터링을 통해 유사한 북마크를 자동 분류함으로써 효율적인 정보 관리를 지원한다. 제목 생성 모듈은 계층적 구조를 채택하여 요약문을 생성한 뒤 이를 기반으로 최종 제목을 생성하며, 실험 결과 ROUGE-L 기준 약 85%의 성능 향상을 확인하였다. 클러스터링 모듈은 Sentence-BERT 임베딩, UMAP 차원 축소, HDBSCAN 알고리즘을 기반으로 구성되며, ‘제목 + 요약’ 조합이 가장 우수한 분류 성능을 나타냈다. 이러한 결과는 기존 북마크 시스템의 한계를 극복하고, 사용자 중심의 자동화된 북마크 관리 기술로 확장 가능함을 시사한다.

■ 중심어 : 제목 생성 ; 클러스터링 ; 자연어 처리 ; 브라우저 확장 프로그램 ; 스마트 북마크

Abstract

This paper proposes “A!rrange,” an AI-powered smart bookmark management system. The system automatically analyzes web page content to generate appropriate titles and performs semantic clustering to categorize similar bookmarks, thereby supporting efficient information management. The title generation module adopts a hierarchical structure, first summarizing the content and then generating the final title based on the summary. Experimental results show that this approach improves performance by approximately 85% in terms of ROUGE-L score. The clustering module utilizes Sentence-BERT embeddings, UMAP for dimensionality reduction, and the HDBSCAN algorithm, with the “title + summary” input combination yielding the best clustering performance. These results suggest that the proposed system effectively overcomes the limitations of conventional bookmark systems and can be extended as a user-centric automated bookmark management solution.

■ keywords : Title Generation ; Clustering ; Natural Language Processing ; Browser Extension ; Smart Bookmark

Ⅰ. 서 론

현대 사회는 디지털 정보의 폭발적인 증가로 인해 심각한 정보 과부하 (information overload) 현상을 경험하고 있다. 사용자는 일상에서 웹 콘텐츠를 소비하며 중요한 정보를 손쉽게 저장하고 관리할 수 있는 효율적인 방법을 필요로 하고 있다. 브라우저의 북마크 기능은 오래 전부터 웹페이지 저장을 위한 대표적인 도구로 사용되어 왔지만, 현재의 북마크 시스템은 사용자에게 많은 불편을 주고 있다. 구체적으로, 북마크의 수가 많아질수록 목록이 어지럽고 원하는 페이지를 찾기가 어려워지며, 폴더를 통한 관리 또한 시간이 지날수록 복잡하고 비효율적이 된다.

자체적으로 진행한 설문조사 결과에 따르면, 약 80%의 사용자가 브라우저 북마크 기능을 사용하고 있으나, 응답자의 60% 이상이 북마크 수가 증가할수록 관리에 어려움을 겪는다고 밝혔다. 또한, 사용자 중 절반가량이 별도의 폴더 구분 없이 무작정 저장하는 것으로 나타났다. 특히 응답자의 90% 이상은 웹페이지 내용을 자동으로 분석하여 적절한 제

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업과 (2021-0-01409) 인공지능융합혁신인재양성사업 연구 결과(IITP-2023-RS-2023-00256629)로 수행되었음

접수일자 : 2025년 07월 02일
수정일자 : 2025년 08월 01일
게재확정일 : 2025년 08월 25일
교신저자 : 김미수 e-mail : misoo.kim@jnu.ac.kr

목과 자동 분류 기능을 제공하는 서비스가 매우 유용할 것이라고 답변했다. 이러한 설문 결과는 기존 북마크 시스템의 한계를 극복하고, 보다 효율적이고 사용자 중심적인 관리 방식을 제공하는 기술 개발이 필요함을 시사한다.

최근 자연어 처리 및 인공지능 기술은 개인화된 정보 관리 서비스에 점차 널리 활용되고 있다. 특히, 대규모 사전학습 언어모델의 등장으로 인해 웹 콘텐츠 분석 및 요약, 제목 생성 기술의 성능이 크게 향상되었다. 예를 들어, 네이버의 "Clova for Writing" 시스템은 이미 자동 요약 및 제목 생성 기술을 상용화하여 사용자의 콘텐츠 작성과 관리 효율을 높이고 있다[1].

본 연구는 위와 같은 기술적 발전을 바탕으로 AI 기반의 효율적인 웹 콘텐츠 관리 시스템인 "A!rrange"라는 브라우저 확장 프로그램을 제안하고자 한다. 이를 통해 사용자가 북마크를 효율적으로 관리하고, 원하는 콘텐츠를 신속하고 정확하게 찾을 수 있도록 돕는 것을 목표로 한다. 본 연구의 주요 기여는 다음과 같다:

(1) 한국어에 특화된 계층적 제목 생성 방식을 도입하여 기존의 단순 생성 방식보다 높은 정확도를 달성한다.
(2) 밀도 기반 클러스터링 방법을 적용하여 사용자 친화적이고 정확한 자동 분류 성능을 제공한다.

이를 통해 기존의 브라우저 북마크 관리 방식의 한계를 극복하고, 사용자의 정보 관리 효율성과 생산성을 높이고자 한다.

Ⅱ. 제안 시스템 및 아키텍처

1. 시스템 개요 및 주요 기능

"A!rrange" 시스템은 AI 기반의 자동화된 북마크 관리 브라우저 확장 프로그램이다. 이 시스템은 사용자가 저장한 웹페이지 정보를 바탕으로 효율적인 북마크 관리를 지원한다. 주요 기능으로는 첫째, 웹페이지의 핵심 내용을 분석하여 직관적이고 명확한 제목을 자동 생성하는 기능이 있다. 둘째, 웹페이지 간 의미적 유사성을 분석하여 유사한 북마크를 자동으로 그룹화하고 카테고리화하는 스마트 분류 기능이 있다. 셋째, 북마크 통계와 시각화, 방문 기록 분석, 맞춤형 추천 컬렉션, 공유 기능 등의 부가 기능도 함께 제공된다. 그림 1은 "A!rrange" 시스템의 메인 인터페이스로, 자동 분류된 북마크 목록과 카테고리 기반 탐색 기능을 보여준다. 그림 2는 대시보드와 설정 화면으로, 사용자의 북마크 사용 패턴 분석과 데이터 관리 기능을 확인할 수 있다.

그림 1. A!rrange" 시스템의 북마크 목록 및 자동 분류 화면

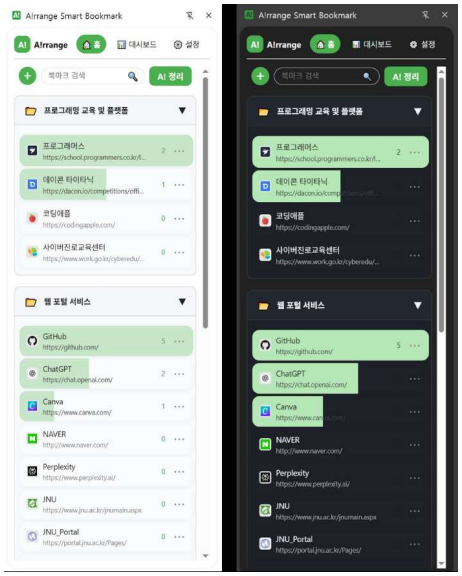
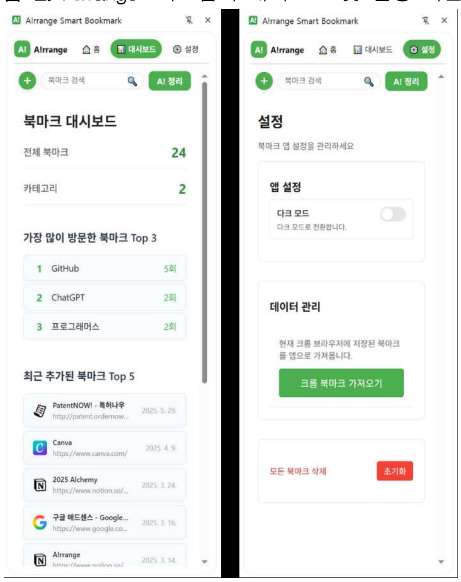


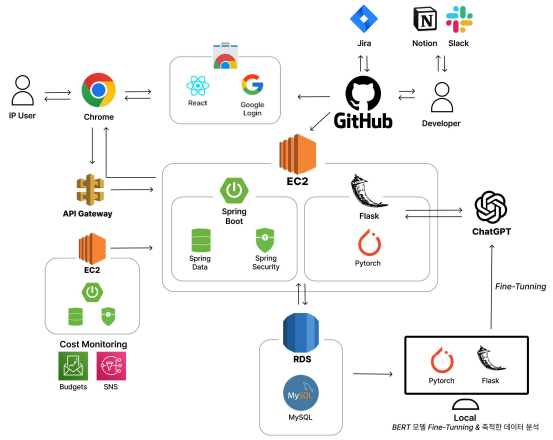
그림 2. A!rrange" 시스템의 대시보드 및 설정 화면



2. 시스템 아키텍처

그림 3은 시스템 아키텍처를 요약한다. 시스템의 아키텍처는 프론트엔드, 백엔드, AI 모델 및 분석 모듈, 데이터 수집 및 처리 모듈로 구성된다. 프론트엔드는 React와 TypeScript 기반의 브라우저 확장 프로그램으로, 사용자에게 직관적인 UI/UX를 제공한다. 백엔드는 FastAPI로 구축된 API 서버를 중심으로 구성되며, MySQL 기반의 RDS 데이터베이스와 함께 AWS Lambda를 활용한 클러스터링 API, AWS EC2에 배포된 제목 생성 API가 포함된다. AI 모델 및 분석 모듈은 PyTorch 및 OpenAI API를 활용하였으며, 의미 분석 및 생성 기능을 담당한다.

그림 3. 시스템 아키텍처



3. 제목 생성 모듈 설계

제목 생성 모듈은 계층적 구조를 갖추고 있다. 구체적으로, KoBART 모델을 통해 웹페이지의 핵심 내용을 요약한 뒤, 이를 KoGPT로 전달하여 최종 제목을 생성한다.

4. 클러스터링 모듈 설계

클러스터링 및 분류 모듈은 Sentence-BERT 기반 텍스트 임베딩을 통해 웹페이지 내용을 의미적으로 벡터화한 후, UMAP을 이용하여 고차원 벡터의 차원을 축소함으로써 계산 효율성을 높인다. 이후 HDBSCAN 알고리즘을 통해 밀도 기반 클러스터링을 수행한다[2].

Ⅲ. 제목 생성 모듈 개발

1. 개요 및 개발 필요성

자연어 처리 분야에서 텍스트 생성 기술이 발전함에 따라, 제목 생성은 정보 요약과 사용자 관심 유도에 유용한 기술로 주목받고 있다. 뉴스 기사, 블로그 포스트, 웹 문서 등 다양한 형태의 콘텐츠에서 생성된 제목은 콘텐츠의 핵심을 간결하게 전달할 뿐 아니라, 사용자의 클릭과 탐색에 직접적인 영향을 미친다. 정보의 양이 폭증하는 오늘날, 수동으로 제목을 작성하는 것은 비효율적이며, 자동화된 제목 생성 기술에 대한 수요가 산업계와 사용자 모두에게서 커지고 있다.

그러나 실사용 텍스트는 종종 장문, 복잡한 문장 구조, 부가적 설명 및 예시 등을 포함하고 있어, 단순한 end-to-end 제목 생성 모델이 핵심 정보를 제대로 파악하지 못하는 경우가 빈번하다. 특히, 모델이 중요 문장을 과소평가하거나, 맥락과 무관한 단어를 선택하는 등의 문제는 생성 제목의 품질 저하로 이어진다.

이를 해결하기 위한 방법 중 하나로 계층적 제목 생성 (hierarchical title generation) 기법이 주목받고 있다[3]. 이 방식은 두 단계로 구성된 계층적 생성 구조를 따른다. 먼저 원문 전체로부터 핵심 내용을 요약한 요약문을 생성하고, 이후 이 요약문을 입력으로 활용하여 최종 제목을 생성하는 방식이다. 이 중간 단계인 요약 생성은 모델이 원문에서 핵심 정보를 선별하는 데 도움을 주며, 결과적으로 제목 생성의 정확성을 높이는 데 기여할 수 있다.

영어를 중심으로 한 기존 연구에서는 이러한 계층적 생성 구조의 효과가 이미 입증되었으나, 한국어를 대상으로 한 실증적 분석은 매우 제한적이다. 특히, 구조가 다른 한국어 문장 특성과 문맥 의존성이 강한 언어 특성상, 계층적 생성 방식의 효과가 그대로 적용될 수 있을지에 대한 검증이 필요하다.

따라서 본 장에서는 한국어 웹 문서를 대상으로, 계층적 제목 생성 구조의 적용 가능성과 실질적인 성능 향상 효과를 실험적으로 분석하였다. 이를 통해 한국

어 환경에 적합한 효율적이고 직관적인 제목 생성 기법을 제안하고, AI 기반 북마크 자동화를 위한 핵심 기술로서의 활용 가능성을 확인하고자 한다.

2. 실험 설계

가. 데이터셋

본 연구에서는 AI Hub에서 제공하는 대규모 웹데이터 기반 한국어 말뭉치[4]를 활용하여 제목 생성 실험을 수행하였다. 해당 말뭉치는 뉴스, 블로그, 쇼핑 후기, 리뷰, 지식인 문서 등 다양한 도메인으로 구성되어 있으며, 실생활에서 자주 접하는 웹 콘텐츠 유형과 유사하다.

데이터셋은 총 8,500건의 문서로 구성되어 있으며, 이 중 6,800건은 학습용, 850건은 검증용, 나머지 850건은 테스트용으로 분할하였다. 각 샘플은 평균 300자에서 500자 분량의 원문 웹페이지 내용을 담은 content와, 사람이 직접 작성한 실제 제목을 나타내는 title로 이루어져 있다.

나. 실험 설정

제목 생성을 위한 실험은 총 4가지 조건으로 구성되며, 각 조건은 계층 구조와 파인튜닝 여부에 따라 다르다. 각 실험 구성은 표 1과 같다.

표 1. 실험 구성

번호	설명
1	원문 → KoGPT → 제목 생성
2	원문 → KoGPT Fine-tuned → 제목 생성
3	원문 → KoBART → 요약문 → KoGPT → 제목 생성
4	원문 → KoBART → 요약문 → KoGPT Fine-tuned → 제목 생성

KoBART[5]는 한국어 BART 기반의 요약 생성 모델이며, KoGPT[6]는 카카오브레인에서 개발한 한국어 GPT 모델이다. 본 실험에서의 파인튜닝은 총 6,800건의 학습 데이터셋을 기반으로 3 epoch 동안 수행되었으며, 과적합 방지를 위해 early stopping 기법을 적용하였다. 학습에는 학습률을 5e-5로 설정한 AdamW 옵티마이저를 사용하였고, weight decay는 0.01로 설정하였다. 배치 크기는 16으로 하였으며, 최대 입력 길이는 256, 출력 길이는 64로 설정하였다. 또한 학습의 안정성을 높이기 위해 gradient clipping을 적용하였으며, 이때 최대 노름 값은 1.0으로 설정하였다.

다. 평가 지표

제목 생성의 성능 평가는 정량적 평가와 정성적 분석의 두 측면에서 이루어졌다. 정량적 평가는 ROUGE[7]지표를 활용하였으며, 구체적으로 단일 단어 단위 중복을 측정하는 ROUGE-1, 연속된 두 단어의 중복을 측정하는 ROUGE-2, 그리고 원본과 생성 문장 간 최장 공통 부분 수열을 기반으로 한 ROUGE-L을 사용하였다.

이와 더불어, 각 실험 조건에서 생성된 제목 예시를 비교 분석함으로써 정성적 평가도 함께 수행하였다. 정성적 평가는 문법적 자연스러움, 핵심 정보의 반영 정도, 그리고 표현의 다양성과 같은 기준을 바탕으로 수행되었으며, 모델이 생성한 제목의 실용성과 품질을 보다 심층적으로 분석하는 데 목적이 있다.

3. 실험 결과 및 분석

가. 정량적 평가

제목 생성 실험의 정량적 평가는 ROUGE 스코어를 기준으로 수행하였다. 표 2는 실험에 사용된 네 가지 모델 설정에 따른 ROUGE-1, ROUGE-2, ROUGE-L 점수를 비교한 결과이다. 실험 4는 모든 지표에서 가장 높은 성능을 기록하였으며, 특히 실제 제목과의 구조적 유사성을 평가하는 ROUGE-L 기준에서 실험 1 대비 약 85.6% 향상된 결과를 나타냈다. 이는 요약 기반 계층적 생성 구조와 파인튜닝된 KoGPT 조합이 제목 생성에 있어 가장 효과적임을 시사한다.

또한 실험 2는 단일 모델 기반으로 파인튜닝을 적용한 경우로서, 실험 1 대비 모든 지표에서 향상된

성능을 보였다. 이는 동일한 모델 구조 내에서도 파인튜닝이 일정 수준의 성능 향상에 기여함을 보여준다. 반면, 실험 3은 요약 구조를 도입하였음에도 불구하고 가장 낮은 ROUGE 점수를 기록하였다. 이는 KoGPT에 파인튜닝이 적용되지 않은 상태에서, 요약문을 입력으로 사용했음에도 불구하고 불안정한 출력이 발생했기 때문으로 해석된다.

표 2. 제목 생성 실험 성능 비교 (ROUGE Score)

지표	실험 1	실험 2	실험 3	실험 4
ROUGE-1	0.0680	0.1004	0.0454	0.1251
ROUGE-2	0.0131	0.0146	0.0055	0.0200
ROUGE-L	0.0666	0.0999	0.0454	0.1236

나. 정성적 평가

정량적 평가와 함께, 생성된 제목의 질적 특성을 분석하기 위해 정성적 평가도 수행하였다. 표 3은 특정 인스턴스를 대상으로 각 실험에서 생성된 제목을 나열한 것으로, 문맥 적합성, 핵심 정보 반영 정도, 문장 구조의 자연스러움 등을 비교 분석하였다.

실험 1의 경우, 문맥과 무관하거나 무의미한 출력을 생성하는 경향이 강하게 나타났다. 예를 들어 인스턴스 1에서는 "저작권법 위반으로..."와 같은 법률 문구가 등장해, 원문과 전혀 관련 없는 결과를 보였다. 실험 2는 비교적 안정적인 문장 구조와 일부 핵심 키워드를 포함하였으나, 표현의 반복이나 문장 구성의 어색함 등으로 인해 완성도에서 한계가 있었다.

실험 3에서는 중간 요약문을 입력으로 사용하였음에도 불구하고 KoGPT에 대한 파인튜닝이 적용되지 않아, URL이 출력되거나 의미 없는 내용이 포함되는 등 비정상적인 결과가 다수 발생하였다. 반면, 실험 4는 대부분의 인스턴스에서 의미 있고 핵심 정보를 반영한 제목을 생성하였다. 예를 들어, 인스턴스 3에서 생성된 “익산시, 베트남 혁신상품 전용매장 입점지원사업 선정”은 실제 제목과 높은 유사도를 보이며, 정보 전달력과 자연스러움 측면 모두에서 우수한 성능을 보였다.

이러한 결과는 계층적 생성 방식이 제목 생성 성능 향상에 유의미한 기여를 할 수 있음을 보여준다. 특히 실험 2와 4의 비교를 통해, 동일한 파인튜닝 조건에서도 요약문을 포함한 계층적 구조가 더 정확하고 정보 밀도 높은 제목 생성을 유도함을 확인할 수 있었다. 향후 연구에서는 요약문과 생성 제목 간의 정보 일치도를 정량적으로 측정할 수 있는 평가 지표 도입을 고려할 필요가 있다.

표 3. 실험별 생성 결과 비교 (ROUGE Score)

ID	구분	내용
1	정답	코웨이 '고객 긴급이슈 24시간 내 대응한다'
	실험 1	저작권법 위반으로 저작권을 침해한 경우, 저작물을 복제하거나 판매·배포한 자는 5년 이하의 징역 또는 5천만
	실험 2	코웨이, '레드캡' 소비자중심경영 실천 선포
	실험 3	'고객중심 경영'을 슬로건으로 하는 레드캡은 ▲고객중심의 서비스 제공 ▲소비자중심 경영 실천 ▲고객
	실험 4	코웨이, 소비자중심경영 실천 '레드캡' 선포
2	정답	(이름) 팬클럽 '영웅시대', 수재민 위해 8억 9천만원 기부
	실험 1	http://www.facebook.com/blog.naver.c
	실험 2	'영웅시대' '영웅시대' '영웅시대' '영웅시대' '희망 팔찌 기부'
	실험 3	영웅시대(http://cafe.naver.com/yongga
	실험 4	'영웅시대' (이름) 팬클럽 '영웅시대' 기부금 8억 기부
3	정답	익산 식품기업, 베트남 식품시장 석권에 도전장
	실험 1	'익산=뉴시스】 (익산=뉴시스】 (익산=뉴시스통신사) (익
	실험 2	익산시, 우수혁신상품 베트남 입점지원사업 선정
	실험 3	베트남 혁신상품 전용매장 입점지원사업은 베트남 정부의 국책사업인 '국경없는 의사회' 사업
	실험 4	익산시, 베트남 혁신상품 전용매장 입점지원사업 선정

4. 결론

본 절에서는 KoGPT 기반의 한국어 제목 생성 모델을 대상으로, 계층적 생성 방식의 적용 여부와 파인튜닝의 효과를 실험적으로 분석하였다. 총 4가지 실험 조건을 비교한 결과, 요약문을 중간 입력으로 활용하고 KoGPT에 파인튜닝을 적용한 계층적 생성 방식(실험 4)이 정량적·정성적 평가 모두에서 가

장 우수한 성능을 보였다. 특히 ROUGE-L 기준으로 단순 제로샷 대비 약 85% 이상 향상된 점수는, 중간 요약 단계를 거친 계층적 구조가 제목 생성에 실질적인 기여를 한다는 점을 시사한다.

또한 정성적 분석에서도 실험 4는 문법적 안정성과 정보 전달력 측면에서 높은 완성도를 보였으며, 다른 실험 조건에서 나타난 맥락 무관 출력이나 불안정한 표현 문제가 크게 개선되었다. 이러한 결과는 단순한 end-to-end 생성 구조보다는, 요약이라는 정보 정제 단계를 거친 입력이 모델의 생성 품질을 향상시키는 데 효과적임을 입증한 것이다.

특히 본 결과는 AI 기반 스마트 북마크 관리 시스템 'A!rrange'에 적용될 제목 자동 생성 모듈의 핵심 기술로서, 사용자의 정보 탐색 효율성과 북마크 관리 편의성 향상에 실질적인 기여를 할 수 있음을 보여준다. 본 연구의 계층적 생성 전략은 사용자 맞춤형 북마크 제목 자동화를 위한 기반 기술로 활용될 수 있으며, 이후 클러스터링 및 추천 시스템과 연계하여 더욱 정교한 북마크 관리 솔루션으로 확장될 수 있다.

Ⅳ. 클러스터링 모듈 개발

1. 개요 및 개발 필요성

북마크의 양이 많아지고 구조가 복잡해짐에 따라, 원하는 정보를 다시 찾는 데 어려움을 겪는다. 특히 기존 브라우저의 북마크 시스템은 단순한 폴더 기반 분류에 의존하고 있어, 사용자가 직접 카테고리를 설정하고 수동으로 정리해야 하는 비효율성이 존재한다. 이로 인해 북마크가 누적될수록 정보 접근성과 관리 편의성이 급격히 저하된다.

이러한 문제를 해결하기 위해서는 저장된 웹페이지의 의미적 유사성을 자동으로 분석하고, 이를 기반으로 자동 분류 및 카테고리화할 수 있는 기술이 요구된다. 클러스터링은 이 목적에 적합한 접근 방식으로, 웹 콘텐츠 간의 주제적 연관성을 기반으로 비지도 학습을 통해 군집을 형성함으로써, 사용자 개입 없이도 체계적인 정보 구조화를 가능하게 한다.

본 장에서는 Sentence-BERT[8]기반의 임베딩 벡터를 활용하여 텍스트 간 의미적 유사도를 계산하고, UMAP[9]을 통한 차원 축소 및 HDBSCAN 알고리즘[10]을 통해 밀도 기반 자동 클러스터링을 수행하였다. 특히 HDBSCAN은 클러스터 수를 사전에 정의할 필요가 없고, 노이즈 데이터 처리가 가능하다는 점에서 다양한 주제와 길이의 웹 콘텐츠에 적합하다[11].

2. 실험 설계

가. 데이터셋

표 4는 클러스터링 실험에 사용한 데이터셋의 예시이다. 클러스터링 실험에는 실제 웹 콘텐츠 기반의 북마크 데이터를 활용하였다. 총 30개의 북마크 샘플을 수집하였으며, 각 샘플은 제목, 요약, 태그의 세 가지 텍스트 정보를 포함한다. 제목은 웹페이지의 원래 제목 또는 사용자가 직접 지정한 북마크 이름이며, 요약은 웹페이지 본문에서 TextRank 알고리즘을 이용해 생성한 핵심 문장으로 구성된다. 태그는 페이지의 주제를 나타내는 핵심 키워드로, 사용자 생성 또는 LLM을 활용한 보조 생성 방식으로 작성되었다. 이 데이터는 'UI/UX', '개발', '코딩테스트', '여행', '쇼핑', '영화', 'IT 트렌드' 등 다양한 주제의 콘텐츠로 구성되어 있으며, 요약문과 태그 생성을 보완하기 위해 ChatGPT-4o를 활용하였다.

한편, 본 실험에 사용된 북마크 샘플 수가 30개로 제한적이라는 점에서, 클러스터링 결과의 일반화에는 한계가 존재한다. 이는 연구의 한계로 작용할 수 있으며, 보다 신뢰도 높은 결과를 도출하기 위해서는 향후 수백에서 수천 개 이상의 대규모 북마크 데이터를 수집하고, 다양한 도메인에 대한 클러스터링 성능을 추가로 검증할 필요가 있다.

표 4. 클러스터링 실험 데이터셋

Label	Title	Summary	Tag
개발	보안 캠페인 출시	GitHub의 '보안 캠페인' 기능으로 개발자-보안 전문가 협업 강화 및 AI 기반 보안 부채 해결 지원	#GitHub #보안 #AI #코파일럿 #개발자협업
디자인	UX 포트폴리오 작성	디자인 철학과 데이터 기반 UX 포트폴리오 작성법	#UX #포트폴리오 #디자인철학 #UX분석 #디자인과정
여행	서울 테마별 여행 코스	서울의 다양한 테마별 명소 추천	#서울 #여행코스 #핫플레이스 #역사 #문화
쇼핑	COS 미니백&H&M 세일	COS 미니백 착용 후기 및 H&M 세일 구매기	#쇼핑 #COS #H&M #미니백 #세일
코딩테스트	DFS vs BFS 선택 기준 정리	그래프 탐색 시 DFS/BFS 선택 기준 설명	#DFS #BFS #그래프탐색 #백트래킹 #최단경로
영화리뷰	위플래쉬: 천재는 행복한가	위플래쉬'를 통해 천재성과 행복의 관계를 고찰	#위플래쉬 #영화리뷰 #천재 #현실성 #감독인터뷰

나. 실험 설정

클러스터링은 의미 기반 자동 분류를 목표로 하며, 다음과 같은 절차로 수행되었다:

(1) 텍스트 임베딩

각 샘플의 제목, 요약, 태그는 Sentence-BERT를 통해 768차원 임베딩 벡터로 변환하였다. 이 임베딩은 문장 간 의미적 유사성을 벡터 공간상에서 계산할 수 있게 해준다. 본 연구에서는 sentence-transformers/all-MiniLM-L6-v2를 사용하여 임베딩을 수행하였다. 해당 모델은 효율성과 속도 면에서 우수한 성능을 보이는 경량화된 BERT 계열 모델로, 다양한 일반 도메인 텍스트에 대해 문장 수준 의미를 효과적으로 포착할 수 있다.

(2) 차원 축소

고차원 임베딩 벡터는 시각화 및 클러스터링의 안정성을 높이기 위해 UMAP을 활용하여 2차원으로 축소하였다. 주요 파라미터로는 로컬 구조를 보존하기 위한 n_neighbors=15, 군집 간 밀집도를 극대화하기 위한 min_dist=0.0, 그리고 시각화를 위한 2차원 축소 설정인 n_components=2를 적용하였다.

(3) 클러스터링 알고리즘

차원 축소된 임베딩 벡터에 대해 HDBSCAN 알고리즘을 적용하였다. HDBSCAN은 밀도 기반의 비지도 클러스터링 기법으로, 클러스터 수를 사전에 지정할 필요가 없고, 잡음 데이터를 자동으로 구분할 수 있다는 장점이 있다. 본 실험에서는 클러스터의 최소 크기를 설정하는 min_cluster_size=3과, 이상치 감지를 위한 min_samples=1의 값을 주요 파라미터로 사용하였다.

(4) 입력 조합 실험 조건

다양한 조합의 입력 정보가 클러스터링 성능에 어떤 영향을 미치는지 확인하기 위해, 표 4의 7가지 조건으로 실험을 수행하였다.

표 5. 입력 조합 실험 조건

번호	입력 조합	구성 요소 설명
1	T	제목만 사용
2	S	요약만 사용
3	K	태그만 사용
4	T + S	제목 + 요약
5	T + K	제목 + 태그
6	S + K	요약 + 태그
7	T + S + K	제목 + 요약 + 태그 (전체 조합)

다. 평가 지표

클러스터링 성능 평가는 세 가지 주요 정량 지표와 정성적 분석을 통해 이루어졌다. 첫 번째로 Adjusted Rand Index (ARI)[12]는 예측된 클러스터와 사전에 정의된 실제 클래스 간의 일치도를 수치화한 지표로,

-1에서 1 사이의 값을 가지며 1에 가까울수록 이상적인 클러스터링을 의미한다. 두 번째로 Normalized Mutual Information(NMI)[13]는 클러스터링 결과와 실제 라벨 간의 정보량을 비교하는 지표로, 0에서 1 사이의 값을 가지며, 높은 값일수록 실제 구조를 잘 반영하고 있음을 나타낸다. 세 번째로 Silhouette Score[14]는 클러스터 간의 분리도와 클러스터 내부의 응집도를 함께 고려하는 지표로, 클러스터 품질을 종합적으로 평가하는 데 유용하다.

정량적 지표 외에도, 클러스터링 결과에 대한 정성적 분석을 통해 실질적인 군집 품질을 추가로 검토하였다. 이는 각 클러스터 내 샘플들의 주제 일관성, 핵심 키워드 유사성, 의미 혼동 여부 등을 중심으로 평가되었으며, 특히 '개발', 'UI/UX', '코딩 테스트'와 같이 의미적으로 인접한 주제 간의 분류 혼동 가능성도 함께 분석하였다. 이러한 다각적 접근을 통해, 클러스터링의 정확성뿐 아니라 실질적 활용 가능성까지 종합적으로 검토하였다.

3. 실험 결과 및 분석

표 5는 클러스터링 실험 결과를 요약한다. 클러스터링 실험 결과, 입력 조합에 따라 성능 차이가 뚜렷하게 나타났다. 정량적 평가에서는 '제목 + 요약(T+S)' 조합이 ARI, NMI, Silhouette Score 모두에서 가장 우수한 성능을 기록하였다. 이는 제목이 제공하는 간결한 주제 정보와 요약이 담고 있는 문서의 핵심 문맥 정보가 상호 보완적으로 작용하여, 클러스터 간 경계를 더욱 명확하게 형성한다는 점을 시사한다. 특히 단일 입력 조건이나 태그가 포함된 조합 대비, T+S 조합은 모든 지표에서 안정적으로 높은 점수를 보였다.

표 6. 클러스터링 실험 결과

입력 조합	ARI	NMI	Silhouette
T (title)	0.260	0.457	0.709
S (summary)	0.525	0.819	0.956
K (tags)	0.312	0.600	0.752
T + S	0.525	0.819	0.977
T + K	0.263	0.522	0.897
S + K	0.367	0.648	0.987
T + S + K	0.591	0.779	0.875

흥미롭게도, Silhouette Score를 기준으로 '제목 + 요약 + 태그(T+S+K)'와 같은 다중 조합보다도 T+S 조합의 클러스터링 성능이 더 높게 나타났다. 이는 태그 정보가 사용자 지정 또는 LLM 기반 생성이라는 점에서 불확실성과 일관성의 문제가 있을 수 있음을 시사하며, 오히려 불필요한 정보가 클러스터링 품질을 저해할 수 있다는 가능성을 보여준다.

정성적 분석에서도 이러한 경향은 동일하게 관찰되었다. T+S 조합은 클러스터 간 분리가 뚜렷하고, 각 클러스터 내부의 샘플들이 주제적으로 잘 응집되어 있었다. 예를 들어 '여행', '쇼핑', '영화'와 같은 일상 주제 클러스터는 명확한 경계를 형성하였고, 중심점 주변의 샘플들은 실제로 해당 주제를 대표할 수 있는 내용으로 구성되어 있었다.

반면, '개발', 'UI/UX', '코딩테스트' 등 의미적 인접성이 높은 주제들은 일부 혼합되거나 중복되는 클러스터로 분포하는 현상이 나타났다. 특히 '요약만'(S) 또는 '태그만(K)' 입력 조건에서는 이러한 경계 혼동이 더욱 뚜렷했으며, Silhouette Score가 낮게 측정되었다. 이는 클러스터링의 정밀도를 높이기 위해 정보의 품질뿐 아니라, 정보 간 조합의 균형이 중요함을 보여준다.

이러한 결과를 바탕으로, 클러스터링 성능 향상을 위한 전략으로 국소적 정밀 클러스터링 (Local Refinement Clustering) 접근이 제안될 수 있다. 예를 들어 1차 클러스터링 결과 내에서 경계가 불분명한 혼합 군집에 대해 2차 클러스터링을 적용함으로써, 세분화된 주제 구조를 형성할 수 있다.

요약하자면, 본 실험은 자동화된 의미 기반 클러

스터링 시스템에 있어 '제목 + 요약' 조합이 가장 핵심적인 입력 구조임을 실증적으로 입증하였으며, 향후 텍스트 기반 분류 시스템 설계 시 복합적이고 압축된 정보 조합의 중요성을 강조하는 근거가 될 수 있다.

4. 결론

본 장에서는 스마트 북마크 시스템의 핵심 기능 중 하나인 의미 기반 자동 클러스터링 모듈을 개발하고, 다양한 입력 조합에 따른 클러스터링 성능을 실험적으로 분석하였다. Sentence-BERT를 활용한 텍스트 임베딩, UMAP 기반 차원 축소, HDBSCAN 알고리즘을 통해 군집화를 수행하였으며, 그 결과 입력 정보의 종류와 조합이 클러스터 품질에 큰 영향을 미친다는 점을 확인할 수 있었다.

특히 '제목 + 요약 (T+S)' 조합이 정량적 평가 결과 높은 성능을 기록하였으며, 정성적 분석을 통해서도 가장 일관성 있는 클러스터 구조를 보여주었다. 이는 간결한 주제 정보인 제목과 핵심 문맥 정보인 요약이 결합되었을 때, 클러스터링 알고리즘이 보다 정밀하게 주제적 유사성을 파악할 수 있음을 의미한다. 반면 태그가 포함된 조합은 오히려 클러스터 품질에 혼선을 줄 수 있음을 시사하였다.

이러한 결과는 현재 개발 중인 AI 기반 스마트 북마크 시스템 A!rrange의 자동 분류 기능 설계에 있어, T+S 조합을 기본 입력으로 활용하는 전략이 가장 효과적임을 뒷받침한다. 더불어 향후 정밀한 주제 구분을 위해 국소적 정밀 클러스터링 (Local Refinement Clustering) 기법을 적용하거나, 사용자 피드백 기반 하이브리드 분류 시스템으로의 확장도 고려할 수 있다.

궁극적으로 본 클러스터링 모듈은 사용자의 북마크를 자동으로 의미 단위로 분류함으로써, 정보 탐색 시간 단축, 카테고리 혼잡 해소, 사용자 경험 향상이라는 실질적 효과를 기대할 수 있다.

V. 결 론

본 연구에서는 웹 콘텐츠 기반의 북마크 데이터를 효율적으로 관리하고 자동화하기 위한 AI 기반 스마트 북마크 관리 시스템 A!rrange를 제안하고, 이를 구성하는 제목 생성 모듈과 클러스터링 모듈의 설계 및 성능을 실험적으로 분석하였다.

제목 생성 실험에서는 KoGPT 기반 모델에 대해 계층적 생성 구조와 파인튜닝 적용 여부를 조합한 총 4가지 조건을 비교하였고, 그 결과 요약을 중간 입력으로 활용하고 KoGPT에 파인튜닝을 적용한 계층적 방식이 가장 우수한 성능을 보였다. 이는 요약을 통해 정보가 정제될 때 모델이 핵심 내용을 더 정확하게 파악하고 표현할 수 있음을 시사하며, 실제 북마크 제목 자동화에 효과적인 전략임을 입증하였다.

또한 클러스터링 실험에서는 Sentence-BERT 임베딩과 UMAP, HDBSCAN을 조합하여 웹 콘텐츠 간 의미 기반 자동 분류를 수행하였으며, 다양한 입력 정보 조합을 비교한 결과 '제목 + 요약' 조합이 가장 높은 군집 품질을 보였다. 이 조합은 정량적 지표에서 안정적인 성능을 나타냈으며, 태그 정보가 포함될 경우 오히려 군집 품질이 저하될 수 있다는 점도 함께 확인되었다. 이러한 결과는 클러스터링 입력 구성의 중요성과, 정제된 문맥 정보를 활용한 군집화 전략의 필요성을 보여준다.

두 모듈의 실험 결과를 종합하면, 본 연구에서 제안한 계층적 제목 생성과 의미 기반 클러스터링 전략은 실제 사용자 환경에서 북마크 자동화, 카테고리화, 탐색 효율성 향상을 가능하게 하는 실용적 성과를 지닌다. 특히 A!rrange 시스템은 기존 폴더 기반 정리 방식의 한계를 극복하고, AI 기반으로 의미적 맥락을 이해하여 사용자 맞춤형 정보 구조를 제공할 수 있다는 점에서 실질적 차별성을 갖는다.

향후에는 사용자 피드백 기반의 파인튜닝, 클러스터링의 계층적 구조 정교화, 다국어 웹페이지 대응, 추천 기능 고도화 등 다양한 확장 방향이 존재한다. 본 연구는 이와 같은 스마트 정보 큐레이션 기술의 기초적 기반을 마련함으로써, 웹 콘텐츠 환경에서의

AI 활용 가능성을 실증적으로 보여주었다는 점에서 의의가 있다.

REFERENCES

[1] 연합뉴스(2023), “네이버, ‘CLOVA for Writing’으로 자동 요약·제목 생성 서비스 제공,” https://www.yna.co.kr/view/AKR20231013049100017, (accessed Jul., 1, 2025).

[2] Eklund, A., Forsman, M., & Drewes, F. (2023). “An empirical configuration study of a common document clustering pipeline,” *Northern European Journal of Language Technology (NEJLT),* vol. 9, no. 1, 2023.

[3] Sun, X., Sun, Z., Meng, Y., Li, J., & Fan, C., “Summarize, Outline, and Elaborate: Long-text Generation via Hierarchical Supervision from Extractive Summaries,” arXiv preprint arXiv:2010.07074, 2020.

[4] AI Hub(2023), “웹사이트 기반 말뭉치,” https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=624, (accessed Jul., 1, 2025).

[5] SKT-AI, “KoBART: Korean BART-based Summarization Model,” GitHub, https://github.com/SKT-AI/KoBART, (accessed Jul., 1, 2025).

[6] Kakao Brain, “KoGPT: Kakao Korean Generative Pre-trained Transformer,” GitHub, https://github.com/kakaobrain/kogpt, (accessed Jul., 1, 2025).

[7] Lin, C. Y., “ROUGE: A Package for Automatic Evaluation of Summaries,” Text Summarization Branches Out, pp. 74 - 81, July 2004.

[8] Reimers, N., & Gurevych, I., “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” arXiv preprint arXiv:1908.10084, 2019.

[9] McInnes, L., Healy, J., & Melville, J., “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv preprint arXiv:1802.03426, 2018.

[10] McInnes, L., Healy, J., & Astels, S., “HDBSCAN: Hierarchical Density Based Clustering,” *Journal of Open Source Software,* vol. 2, no. 11, p. 205, 2017.

[11] Saha, R. (2023). Influence of various text embeddings on clustering performance in NLP. arXiv preprint arXiv:2305.03144.

[12] Steinley, D., “Properties of the Hubert-Arabie Adjusted Rand Index,” *Psychological Methods,* vol. 9, no. 3, pp. 386 - 396, 2004.

[13] Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M., “Normalized Mutual Information Feature Selection,” *IEEE Transactions on Neural Networks,* vol. 20, no. 2, pp. 189 - 201, 2009.

[14] Shahapure, K. R., & Nicholas, C., “Cluster Quality Analysis Using Silhouette Score,” *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),* pp. 747 - 748, Sydney, Australia, Oct. 2020.

저 자 소 개

김정우(준회원)

현재 전남대학교 에너지자원공학과 수료

<주관심분야 : 자연어 처리, LLM, 딥러닝>

서재오(준회원)

2023년 명지전문대학교 소프트웨어콘텐츠과 전문학사 졸업.

현재 전남대학교 인공지능학부 인공지능전공 재학.

<주관심분야 : 자연어 처리, LLM, Devops>

임한빈(준회원)

2019년 조선대학교 생명과학과 2년 수료

현재 전남대학교 인공지능학부 인공지능전공 재학.

<주관심분야 : 풀스택 개발, MLOps, LLM 프롬프트 엔지니어링>

김미수(정회원)

2013년 고려대학교 경영학부 학사 졸업.

2021년 성균관대학교 전자전기컴퓨터공학과 박사 졸업.

현재 전남대학교 인공지능학부 조교수

<주관심분야 : 자연어 처리, 소프트웨어 유지보수>