JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 5, pp. 39-49, May 2025
https://doi.org/10.9708/jksci.2025.30.05.039

Improving Topic-Specific Document Selection through BERTopic and Re-Clustering

Woon-Kyo Lee*, Ja-Hee Kim**

*Student, Graduate School of IT Policy, Seoul National University of Science and Technology, Seoul, Korea
**Professor, Graduate School of IT Policy, Seoul National University of Science and Technology, Seoul, Korea

[Abstract]

This study analyzes the performance of BERTopic-based clustering across various data distributions. A re-clustering method is proposed to improve the selection of documents on a specific target topic. Existing clustering techniques often face challenges in accurately selecting documents when the proportion of documents related to the target topic is very low or very high. To address this issue, sampling is performed on retrieved documents to include target topic documents at varying ratios. Clustering is performed using SBERT-based BERTopic with K-means and HDBSCAN algorithms, and the results are compared. In the re-clustering step, documents initially classified as outliers are re-clustered and merged with the original results. The results before and after re-clustering were compared using four evaluation metrics. Accuracy improved from 0.7251 to 0.9421, and the F1 Score increased from 0.8449 to 0.9423. ARI increased by 0.3626 and NMI by 0.2805. This indicates that the proposed method enhances clustering quality and improves the accuracy of document selection.

▸ **Key words**: BERTopic, Clustering, Re-Clustering, K-means, HDBSCAN

[요 약]

본 논문에서는 다양한 데이터 분포에 따른 BERTopic 기반 군집화 결과를 분석한다. 또한, 연구자가 원하는 목표 주제 문서를 선별하기 위한 재군집화 방법을 제안한다. 기존 군집화 기법은 목표 주제의 문서 비율이 매우 낮거나 높으면 정확한 문서 선별에 어려움을 겪는다. 이 문제 해결을 위해 검색으로 수집된 문서에서 목표 주제 문서가 다양한 비율로 포함되도록 샘플링을 수행한다. SBERT를 사용한 BERTopic 모델에 K-평균과 HDBSCAN 알고리즘을 적용하여 군집화를 수행하고 결과를 비교한다. 재군집화 단계에서는 이상치로 분류된 문서를 다시 군집화하여 원래 결과와 병합한다. 재군집화 전후 결과는 네 가지 평가지표를 통해 비교했다. 정확도는 0.7251에서 0.9421로, F1 Score는 0.8449에서 0.9423으로 향상되었다. ARI와 NMI는 각각 0.3626과 0.2805만큼 증가했다. 이는 제안된 방법이 군집 품질을 향상하고 문서 선별의 정확도를 높인다는 것을 보여준다.

▸ **주제어**: 버토픽, 군집화, 재군집화, K-평균, HDBSCAN

• First Author: Woon-Kyo Lee, Corresponding Author: Ja-Hee Kim
*Woon-Kyo Lee (johntato@seoultech.ac.kr), Graduate School of IT Policy, Seoul National University of Science and Technology
**Ja-Hee Kim (jahee@seoultech.ac.kr), Graduate School of IT Policy, Seoul National University of Science and Technology
• Received: 2025. 04. 01, Revised: 2025. 04. 18, Accepted: 2025. 05. 12.
• This paper is an extension of the work presented at the 71st Winter Conference of the Korea Society of Computer Information in 2025, titled "Evaluating Document Screening Performance with BERTopic and Re-Clustering."

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

텍스트 데이터의 급증과 함께 특정 주제의 문서를 분류하고 선별하는 군집화(clustering) 기법의 중요성이 커지고 있다[1]. 비정형 텍스트 데이터에서 특정 주제의 문서를 찾는 작업은 정보 검색, 추천 시스템, 문서 분류 등의 다양한 응용 분야에서 핵심적인 역할을 한다. 군집화 기법을 적용할 때, 특정 주제의 문서가 전체 데이터에서 차지하는 비율이 다른 주제에 비해 낮거나 높은 경우, 군집화 결과의 품질이 저하될 수 있다[9]. 지도학습에서 연구된 데이터 불균형(imbalanced data)문제와 유사하게 군집화에서도 데이터 분포가 편향될 때 군집화 결과가 왜곡될 수 있다.

전통적인 군집화 성능을 개선하기 위해 최근에는 버토픽(BERTopic)과 같은 토픽 모델링 기반의 군집화 기법이 주목받고 있다[3]. 버토픽은 사전 학습된 언어 모델과 토픽 모델링을 결합하여 의미적으로 유사한 문서들을 효과적으로 군집화하는 기법이다. 기존의 토픽 모델링 기법과 달리, 버토픽은 문서 임베딩을 활용하여 고차원 공간에서 문서 간의 의미적 유사성을 더 정교하게 반영할 수 있다[3][4][5]. 하지만 버토픽 기반 군집화에서도 연구자가 원하는 목표 주제를 가진 문서의 비율이 극단적으로 낮거나 높은 경우, 일부 문서가 이상치(outlier)로 분류되거나 적절한 군집을 형성하지 못하는 문제가 발생할 수 있다. 특히, 목표 문서가 다수의 군집으로 분산되거나 주요 군집에서 벗어나 단독으로 존재하는 경우, 기존 군집화 기법만으로는 효과적인 문서 선별이 어렵다[9].

본 연구는 버토픽 기반의 군집화 과정에서 이상치로 분류된 문서들을 재군집화(Re-Clustering)하는 방법을 통해 목표 주제 문서의 선별 성능을 개선하는 방법을 탐구한다. 기존 연구에서 텍스트 데이터의 군집화 성능을 향상시키기 위해 다양한 접근법이 제안되었으며, 특히 밀도 기반의 HDBSCAN 군집 기법을 버토픽과 함께 사용하면 불규칙한 모양의 군집이나 노이즈 처리에 강점이 있어 군집 품질을 높일 수 있다[2][3][5]. 하지만 이러한 연구들은 주로 군집의 노이즈 영향을 줄이고 일관된 문서 군집에 초점을 맞추었으며, 특정 주제의 목표 문서를 선별하는 데 있어 재군집화 기법의 효과를 분석한 사례는 부족하다.

이에 본 연구는 다양한 데이터 분포에 따른 버토픽 기반 군집화 결과를 분석하고, 목표 주제 문서를 더 정확하게 선별할 수 있도록 버토픽 기반의 재군집화 방법을 제안한다. 특히, 기존 군집화 과정에서 이상치로 분류된 문서들을 재배치함으로써 군집 품질을 개선하고, 목표 주제 문서의 선별 성능을 개선하는 것이 본 연구의 주요 목표이다. 이를 통해 버토픽 기반의 군집화 성능을 확인하고, 보다 신뢰성 있는 문서 선별 방법을 제시하고자 한다.

2장에서는 문서 군집화와 문서 선별 연구에 대하여 알아보고 3장에서는 연구 절차에 대하여 상세하게 설명한다. 4장과 5장에서는 연구 결과와 연구를 요약하고 향후 연구 과제를 논의한다.

II. Review of related works

1. Document clustering and Topic modeling

문서 군집화는 유사한 주체를 가진 문서들을 자동으로 그룹화하는 기술이다. 정보 검색, 추천 시스템, 주제 분석 등 다양한 영역에서 사용된다. 전통 기법으로 사전에 정의된 k개의 군집으로 나누는 K-평균 군집화(K-means clustering), 계층적으로 분류하는 계층적 군집화(Hierarchical clustering), 밀도 기반의 DBSCAN, 그래프 이론을 활용한 스펙트럴 군집화(Spectral clustering) 등이 있다. 전통적인 군집은 대부분은 거리 또는 유사도에 기반한다. 하지만 고차원, 희소성과 같은 데이터 불균형문제로 군집 성능에 한계가 있다[1].

최근에는 문서 간 거리 기반 아닌, 밀도 기반 접근 방식이 대안으로 주목받고 있으며, HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)이 대표적인 기법이다. HDBSCAN은 DBSCAN의 한계를 보완하여 군집의 밀도 변화에 유연하게 대응하며, 군집 되지 않는 데이터를 구분하는 특징이 있다. 문서 분석에서 중요하지 않거나 제거할 문서를 찾는 데 유리한 특성을 제공한다[2].

주제 모델링 기법은 단어 문서 간 확률분포로 주제를 도출하는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기반이 사용된다. 최근에는 문장의 의미를 반영할 수 있는 트랜스포머(Transfomer) 기반의 버토픽이 주목받고 있다[3]. 버토픽은 사전 학습된 트랜스포머(Transformer) 기반 언어 모델을 활용하여 문서 임베딩을 생성하고, 이를 군집화하여 주제를 도출한다. 클래스 기반 TF-IDF(Term Frequency-Inverse Document Frequency) 절차를 통해 각 클러스터의 핵심 단어를 추출함으로써 주제의 해석 가능성을 높인다. 이러한 접근법은 전통적인 LDA 모델과 비교하여 향상된 주제 일관성과 다양성을 제공한다[3][4]. 버토픽은 텍스트 임베딩을 생성하는데 SBERT(Sentence-BERT)를 활용하며, 기존 BERT 기반 모델보다 더 빠르고 효율적으로 문서 간 유사성을 계

산할 수 있다[5].

문서 군집화의 성능 평가는 일반적으로 정확도(Accuracy), F1 Score, ARI(Adjusted Rans Index), NMI(Normalized Mutual Information)등의 지표를 이용한다. 정확도와 F1 Score는 정답 레벨과 예측 라벨 간의 일치도를 평가하며, ARI와 NMI는 군집간의 유사성과 무작위성(Randomness)을 측정하는 지표로 사용된다[6].

2. Document Identification and Reclustering

문서 선별은 트렌드 분석 연구[7]나 체계적 문헌 고찰 연구[8]에서 매우 중요한 단계로, 연구의 목적에 부합하는 문서 선별 과정은 필수적이다. 선행 연구에서 문서 선별을 위해 K-평균 군집화를 활용해 문서 선별 성능을 개선한 사례[9]나 BERT와 오토인코더를 사용한 사례[14]가 있으나 여전히 문서 선별 과정에 개선의 필요성이 제기됐다.

일반적인 군집화 기법은 주어진 데이터 집합을 한 번의 프로세스로 군집화하는 방식이다. 하지만, 실제 데이터에서는 군집화 과정에서 일부 데이터가 적절하게 배치되지 않거나, 특정 군집이 과도하게 분산되는 문제가 발생할 수 있다. 군집화 성능을 개선하기 위해 반복적 군집화(Iterative clustering) 방법이 도입되었다[10][11]. 이는 초기 군집 결과를 바탕으로 추가적인 피드백을 제공하여 점진적으로 군집 품질을 향상하는 방식이다. 이러한 접근은 군집의 성능을 향상하고 복잡한 데이터에서도 효과적인 군집화를 가능하게 한다[10][11].

군집화의 주요 과제 중 하나는 군집 결과를 해석하고 의미를 부여하는 것이다. 비지도 학습 기반의 군집화 기법에서는 군집의 의미를 명확하게 정의하는 것이 어려우며, 이를 해결하기 위한 연구가 지속되고 있다[12]. 주제 표현(topic representation) 기반의 예측 라벨링(predictive labeling) 기법이 활용되며, 이는 군집 내 대표적인 키워드와 주제어를 분석하여 각 군집의 의미를 부여하는 방식이다[12].

전통적인 LDA 기반 방법은 단어-문서 행렬 분석을 통해 군집별 주요 키워드를 추출하는 방식이지만, 최근에는 BERT 기반의 버토픽 기법이 등장하여 문맥적 의미까지 반영할 수 있는 기법이 주목받고 있다[3]. 본 연구에서는 군집화 과정에 버토픽을 적용하여 주요 주제를 파악하고 군집화 결과를 정량적으로 분석한다.

군집별 주요 토픽 키워드와 평가 지표를 함께 비교함으로써, 군집화의 품질을 주제적 관점에서도 검토할 수 있도록 한다. 이러한 접근 방식은 기존 연구에서도 군집 해석의 중요한 도구로 사용되었다, 군집화된 데이터의 품질을 평가하는 일반적인 방법으로는 정확도, ARI, NMI와 같은 정량적 평가와 정성적 평가가 병행된다[13]. 다양한 군집화 알고리즘의 특징을 분석하여 사용자가 자신의 데이터와 문제에 가장 적합한 알고리즘을 선택하는 것은 중요하며, 군집화 평가 방법에 대한 논의는 필요하다[13].

본 연구는 버토픽 기반 재군집화 기법을 적용하여 특정 문서를 효과적으로 선별하는 방법을 제안한다. 군집 기법에 따라 어떤 차이가 있는지와 원하는 문서의 분포에 따른 군집 기법의 성능 차이도 알아본다.

III. Proposed approach

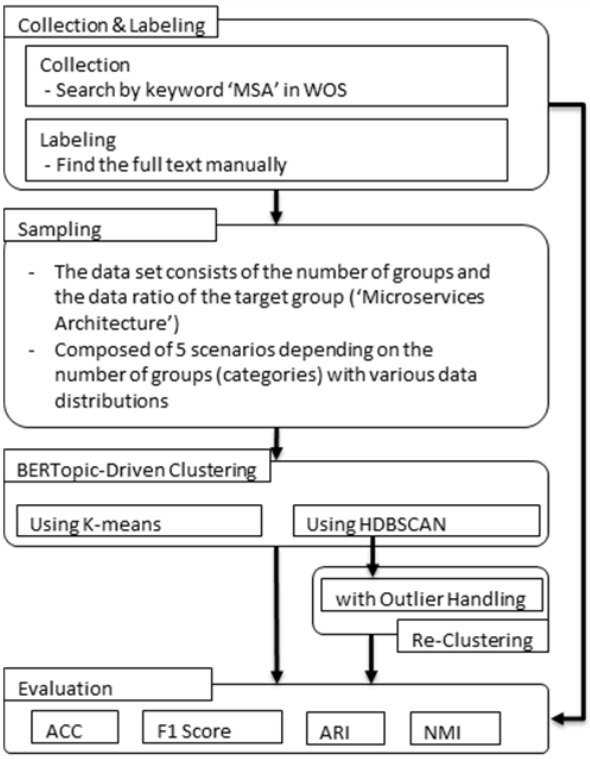

Fig. 1. Research Process

본 연구는 버토픽 기반으로 문서 집합에서 연구자가 원하는 문서를 선별하기 위해 Fig. 1과 같이 수집 및 라벨링, 샘플링, 군집화, 평가의 4가지 주요 과정으로 진행되었다. 주요 과정에 대하여 다음 각 절에서 상세하게 설명한다.

1. Collection and Labeling

수집 과정은 선별하고자 하는 목표 문서를 모으기 위하여 키워드를 통하여 검색하여 초기 문서 집합을 구성하는 과정이다. 본 연구에서는 초기 문서 집합을 구성하기 위하여 목표 문서의 주제를 'Microservice Architecture'로

설정하고, 검색 키워드는 'MSA'로 정했다. WOS(Web of Science)에서 검색 기간을 2000년부터 2024년 8월까지, 카테고리를 'computer' 관련으로 한정하여 2,003개의 초기 문서 집합을 수집했다. 연도별 수집된 문서의 수는 Fig. 2와 같으며, 전반적으로 매년 증가하였음을 알 수 있다.

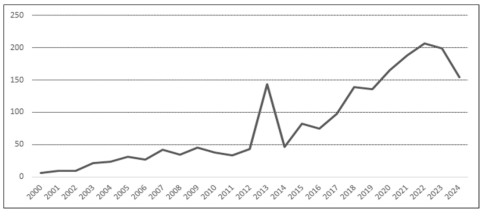

Fig. 2. Count of paper by publication year

수집된 문서 집합의 'MSA'에 대한 원형 표현(full form)을 알기 위해서 수집된 문서의 제목, 요약문, 본문에서 원형 표현을 수작업으로 찾아 라벨링을 진행했다. 라벨링 작업은 목표 문서 집합을 선별한 결과와 비교하기 위하여 사용된다. 위키백과에서 'MSA'로 검색하면 다중 서열 정렬(Multiple sequence alignment), 일본 해상 보안청(Maritime Safety Agency), 마이크로서비스 아키텍처(Microservice Architecture), 현대 표준 아랍어(Modern Standard Arabic) 등의 동음이의어가 있다[15]. 라벨링된 문서도 다양하게 분포되어 있으며, 여러 문서가 혼합되어 수집되었다. 원형 표현별 분포 현황은 그림 Fig. 3과 같다.

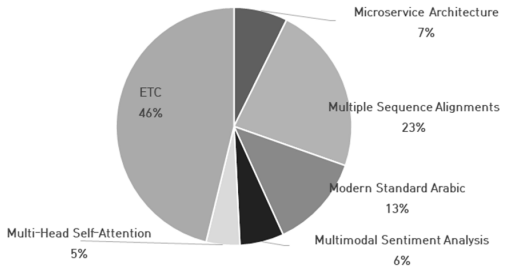

Fig. 3. Rate of data by label

원형 표현이 소량인 경우는 'ETC'로 분류했다. ETC에는 'Microstrip Antena', 'Mult-Scale Attention', 'Measurement System Analysis'등이 있다. 원형 표현을 찾을 수 없거나, 저자명, 참고문헌에 있는 경우는 대상에서 제외했다. 수집되어 라벨링된 대상 문서의 수는 1427개이다.

2. Sampling

수집된 문서에서 목표 문서 집합의 분포에 따른 군집의 결과를 확인하기 위하여 목표 문서 집합의 비율에 따라 샘플링 단계를 진행했다. 샘플링 단계에서는 라벨링된 문서 그룹의 수에 따라 5개의 시나리오를 구성했다. 시나리오별 문서 그룹의 수와 비율은 Table 1과 같다.

Table 1. Data rate of Group by Scenario

Scenario	Group Count	Full Text	Document Count	Rate
Scenario 1	2	Microservice Architecture	105	24.2%
		Multiple Sequence Alignments	328	75.8%
Scenario 2	3	Microservice Architecture	105	17.0%
		Multiple Sequence Alignments	328	53.2%
		Modern Standard Arabic	183	29.7%
Scenario 3	4	Microservice Architecture	105	15.0%
		Multiple Sequence Alignments	328	46.7%
		Modern Standard Arabic	183	26.1%
		Multimodal Sentiment Analysis	86	12.3%
Scenario 4	5	Microservice Architecture	105	13.7%
		Multiple Sequence Alignments	328	42.7%
		Modern Standard Arabic	183	23.8%
		Multimodal Sentiment Analysis	86	11.2%
		Multi-Head Self-Attention	66	8.6%
Scenario 5	6	Microservice Architecture	105	7.4%
		Multiple Sequence Alignments	328	23.0%
		Modern Standard Arabic	183	12.8%
		Multimodal Sentiment Analysis	86	6.0%
		Multi-HeadSelf-Attention	66	4.6%
		ETC	659	46.2%

각 시나리오는 목표 문서 그룹의 비율을 5%에서 95%까지 5% 단위로 설정하여 총 19개로 샘플링한 문서 집합으로 구성했다. 샘플 문서 집합은 시나리오와 비율에 따라 총 95개의 문서 집합을 생성했다. 시나리오와 문서 그룹의 비율에 따라 무작위로 추출하여 샘플 문서 집합을 만들었다. 목표 문서 집합의 비율과 목표 문서가 아닌 그룹의 비율 관계는 Fig. 4와 같다. 목표 그룹의 문서 수와 목표 그룹이 아닌 문서 수는 비율에 따라 다르게 조절했다. 목표 문서의 비율이 낮은 구간에서는 목표 문서의 수를 조정하고, 목표 문서의 비율이 높은 구간은 목표 문서가 아닌 문서의 수를 조절하여 샘플링했다.

Table 2. Target vs Total Documents by Scenario

Scenario		Rate of Target Document																		
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Scenario 1	Total	420	430	426	430	420	350	300	262	233	210	190	175	161	150	140	131	123	116	110
	Target	21	43	64	86	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105
Scenario 2	Total	600	610	613	525	420	350	300	262	233	210	190	175	161	150	140	131	123	116	110
	Target	30	61	92	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105
Scenario 3	Total	700	700	700	525	420	350	300	262	233	210	190	175	161	150	140	131	123	116	110
	Target	35	70	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105
Scenario 4	Total	760	760	700	525	420	350	300	262	233	210	190	175	161	150	140	131	123	116	110
	Target	38	76	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105
Scenario 5	Total	1420	1050	700	525	420	350	300	262	233	210	190	175	161	150	140	131	123	116	110
	Target	71	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105	105

Fig. 4. Target vs Non-Target Document Ratio

시나리오별 샘플 문서 집합은 Table 2와 같이 목표 그룹 문서와 전체 문서의 비율을 조절하여 구성했다. 예를 들어 시나리오 5의 목표 그룹 문서 비율 5% 구간의 경우는 ETC를 포함하여 대상 문서 1,427개 중 목표 그룹의 문서 수가 5%가 되도록 구성하여 전체 문서 수는 1,420개로 구성했다. 95% 구간의 경우는 목표 그룹의 105개 문서를 모두 포함하고 목표 그룹이 아닌 문서를 무작위로 추출하여 비율을 맞추어 전체 문서 수를 110개로 구성했다.

3. BERTopic-Driven Clustering

군집화 단계에서는 준비된 샘플 문서 집합에서 목표 문서 그룹를 선별하기 위해 버토픽 기반의 군집화를 수행했다. Fig. 5는 버토픽 기반의 군집화 단계를 보여준다. 문서 임베딩은 Sentence-Transformer 모델('all-MiniLM-L6-v2')를 사용했으며, 차원 축소는 UMAP(Unified Manifold Approximation and Projection)을 활용했다. 군집화 기법으로 HDBSCAN(Hierarchical DBSCAN)과 K-평균을 적용했다. 토큰화 도구와 주제 표현은 각각 TfidfVertorizer, C-TF-IDF를 적용했다. 버토픽 과정에서 군집화 기법에 따라 목표 문서 집합 선별의 차이를 확인하기 위해 버토픽 과정 중 군집화 기법 외에 다른 과정은 같은 방법으로 진행했다.

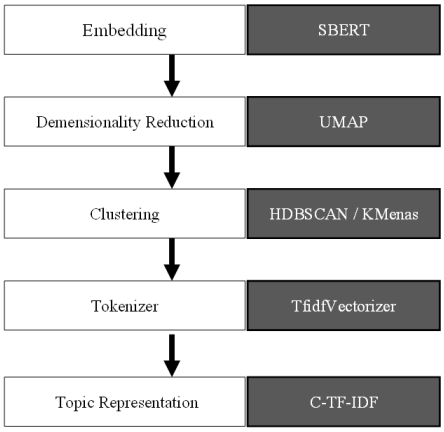

Fig. 5. BERTopic Driven Clustering Flow

재군집화는 HDBSCAN의 이상치를 재 군집화하여 진행했다. Fig. 6은 재군집화 과정을 사례로 보여준다. HDBSCAN 결과에서 이상치로 분류된 데이터를 대상으로 추가적인 재군집화를 수행했다. 재군집화는 이상치의 문서 수가 일정 기준(threshold) 이상일 경우에 진행했다. 재군집화의 이상치 기준값은 10으로 설정했다. 재군집화 과정은 초기 버토픽 기반의 군집화 과정과 같은 방법으로 수행했으며, 군집화 기법은 HDBSCAN을 적용했다.

4. Evaluation

평가에 앞서, 라벨링 과정에서 각 문서에 찾은 약어의 원형 표현을 정답 값(ground truth)으로 부여했다. 군집화된 문서 그룹이 목표 문서 그룹에 해당하는지 판단하기 위한 과정을 수행했다. 버토픽 기반의 문서 군집화를 통해

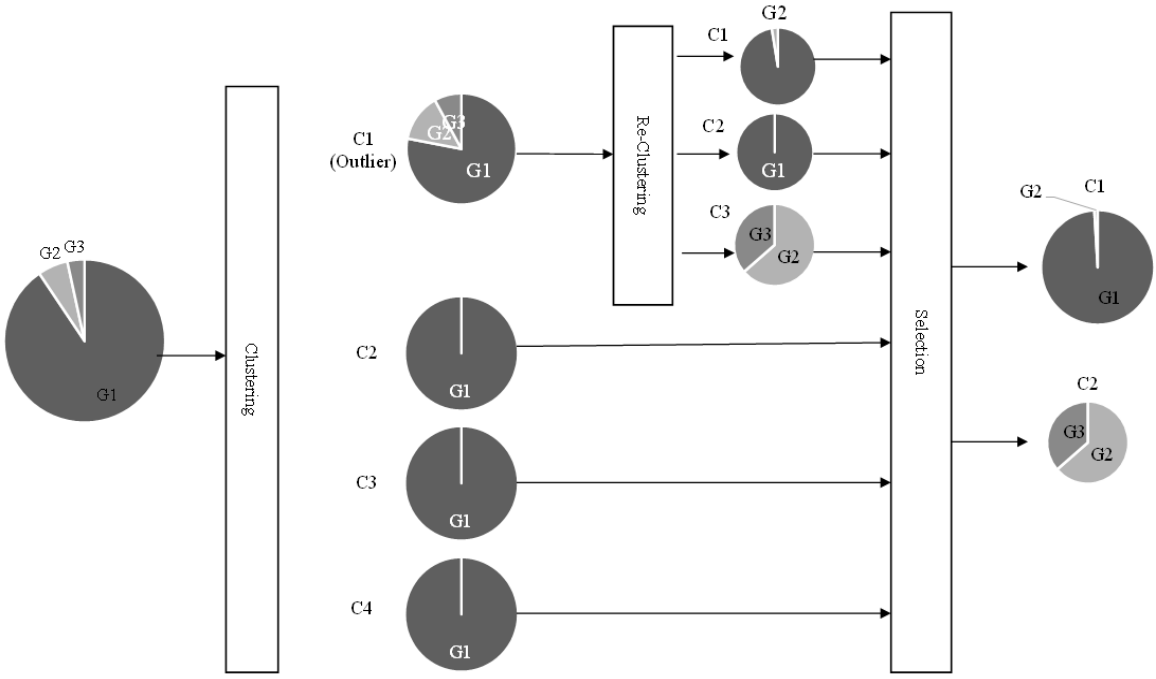

Fig. 6. Re-Clustering Flow

Topic	Count	Name	Representation	Representative_Docs
0	106	0_architecture_msa_service_systems	['architecture', 'msa', 'service', 'systems', 'development', 'services', 'applications', 'architectures', 'approach', 'architectural', 'based', 'testing', 'paper', 'data', 'quality', 'challenges', 'study', 'research', 'computing', 'model']	['An Extensible Data-Driven Approach for Evaluating the Quality of Microservice Architectures Microservice architecture (MSA) is defined as an architectural style where the software system is developed as a suite of small services, each running in its own process and communicating with lightweight mechanisms. The benefits of MSA are many, ranging from an increase in development productivity, to better business-IT alignment, agility, scalability, and technology flexibility. The high degree of microservices distribution and decoupling is, however, imposing a number of relevant challenges from an architectural perspective. ... ']
1	13	1_alignment_sequence_multiple_sequences	['alignment', 'sequence', 'multiple', 'sequences', 'algorithm', 'msa', 'using', 'methods', 'prediction', 'results', 'accuracy', 'high', 'optimization', 'problem', 'msas', 'model', 'analysis', 'quality', 'outlier', 'biological']	['A New Genetic Algorithm Using Gap Matrixes for Multiple Sequence Alignment As is well known, multiple sequence alignment (MSA) is a widely used technique in molecular sequence analysis, and the object of MSA is to find optimal alignment for several sequences. By using gap matrixes, this paper devised a new genetic algorithm to solve MSA problem, and the approach is examined by using a set of standard instances taken from the benchmark alignment database BAliBASE. Numerical simulations are performed to verify the significance of the proposed algorithm.', '...']
2	21	2_fusion_msa_analysis_attention	['fusion', 'msa', 'analysis', 'attention', 'datasets', 'language', 'information', 'network', 'features', 'model', 'standard', 'modern', 'feature', 'learning', 'dataset', 'data', 'used', 'paper', 'approach', 'performance']	['BCD-MM: Multimodal Sentiment Analysis Model With Dual-Bias-Aware Feature Learning and Attention Mechanisms Multimodal Sentiment Analysis (MSA) is gaining attention, but faces two main challenges: efficient extraction of cross-modal features without redundancy and removing spurious correlations between sentiment labels and multimodal features. In this paper, we propose a novel multimodal learning debiasing model, named Bilateral Cross-modal Debias Multimodal sentiment analysis Model (BCD-MM), to address these issues. Specifically, BCD-MM ultimately enhances the generalisation of the model to out-of-distribution (OOD) situations by improving the ability of cross-modal low-redundancy feature extraction and reducing the reliance on non-causal correlations. ...']

Fig. 7. Overview of Clustered Documents in BERTopic Results

각 군집에서 주제어(topic word)를 추출하고, 주제어 점수(topic word scores)를 활용하여 대표 주제어를 선정했다. 토픽의 대표 문서(Representation Docs)에서 나타난 문서 내용과 약어의 원형 표현을 함께 고려하여 해당 군집의 예측값(predicted label)을 결정했다. 예측값과 정답값을 비교하여 군집화 성능을 평가했다.

Fig. 7과 Fig. 8은 목표 문서 그룹을 판별하기 위한 주제어 점수와 토픽의 대표 문서의 예시를 보여준다. 예를 들어, 'Microservice Architecture' 관련 문서는 Topic 0에서 주제어와 및 주제 표현이 반영하고 있어 Topic 0을 목표 문서 그룹으로 식별할 수 있다. 군집화 성능을 평가하기 위해 정확도, F1 Score, ARI, NMI 지표를 활용했다.

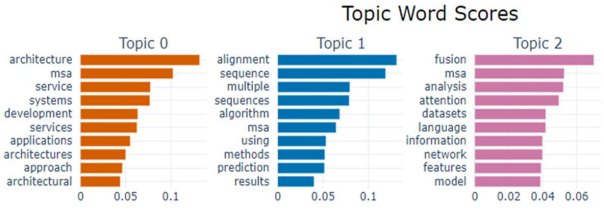

Fig. 8. Topic Word Scores

$$Accuracy = \frac{N}{T} \quad (1)$$
where N is number of correctly clustered documents, and T is Total number of documents.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

$$ARI = \frac{RI - Expected\ RI}{Max RI - Expected\ RI} \quad (3)$$
where RI is the Rand Index.

$$NMI(u,v) = \frac{2 \cdot I(u;v)}{H(u) + H(v)} \quad (4)$$
where u and v are the sets of actual and clustered labels, $I(u;v)$ is Mutual information between u and v, $H(u), H(v)$ are the entropies of each set.

정확도는 전체 문서 중에 예측값과 정답값이 일치하는 비율로 전반적인 군집 성능을 확인할 수 있는 지표이다. F1 Score는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로 목표 그룹 문서 수의 데이터 비율에 따라 군집 성능의 군집 결과를 평가할 수 있다. 군집화 결과와 구조적 일치도를 평가하기 위해 ARI와 NMI를 함께 사용했다. ARI는 라벨링에서 도출한 정답 값과 군집의 결과인 예측값 간 일치도를 고려하며, 무작위 군집화의 기대치를 보정한 지표로 군집 수가 상이한 경우에도 신뢰도 높은 비교가 가능하다. NMI는 예측값과 정답값 간의 정보 공유 정도를 측정하는 지표로, 군집 간 상호 정보를 정규화 값을 사용하여 군집 품질을 비교할 수 있다. 수식(1), (2), (3), (4)는 지표의 수식으로 군집의 결과를 평가했다.

IV. Study result

1. Results by Dataset Composition

목표 문서 집합을 찾기 위한 군집 기법 간의 평가지표 결과를 분석했다. Fig. 9는 각 시나리오별로 군집화 성능 변화를 나타내는 그래프이다. 그래프의 x축은 목표 문서의 데이터 비율을 나타낸다. y축은 평가지표의 값으로 0~1 사이의 값으로 표시한다. 시나리오는 문서 그룹의 수에 따라 나누어진다. 시나리오 1에는 2개의 문서 그룹이 포함되고, 시나리오 2, 3 및 4에는 각각 3, 4, 및 5개의 문서 그룹이 포함된다. 시나리오 5에서는 모든 문서 그룹이 포함되어 여러 주제의 문서와 노이즈도 포함되었다.

Fig. 9는 목표 문서의 비율 변화에 따른 네 가지 성능 지표의 추이를 시나리오별로 시각화한 결과를 보여준다.주제가 명확히 구분되는 시나리오 1, 2, 3, 4 에서는 목표 문서 비율이 40~90% 구간일 때 정확도, F1 Score, ARI, NMI 등 네 가지 성능 평가지표가 모두 높은 값을 나타냈다. 특히 시나리오 3의 경우 25~85% 구간의 네 가지 성능 평가지표가 모두 1.0의 값을 보여 해당 구간에서 목표 문서를 잘 군집하고 있다. 반면 5%, 10%, 95%와 같은 목표 문서의 비율이 극단적 구간에서는 네 가지 지표 모두 성능이 급격히 저하되었다. 이는 목표 주제의 문서 비율의 분포가 군집화 결과에 영향을 미치며 목표 주제의 문서 분포가 과도하게 적거나 많지 않은 경우라면 군집화 기법이 좋은 결과를 나타낼 수 있음을 시사한다.

시나리오 5는 혼합된 주제와 다량의 노이즈를 포함한 복잡한 데이터 환경을 구성했다. 이로 인해 대부분의 군집화 기법이 전반적으로 낮은 성능을 보였으며, 각 기법 간 성능 차이도 불안정하게 나타났다. 특히 동일한 문서 비율 구간에서도 평가지표가 큰 폭으로 변동하며, 일관된 군집 구조를 형성하는 데 어려움을 겪는 모습을 보였다. 이러한 결과는 단일 주제가 명확하게 구분되는 다른 시나리오와 달리, 주제 간 경계가 모호하고 노이즈가 많은 환경에서는 기존 군집화 방식으로는 의미 있는 결과를 도출하기 어려울 수 있음을 나타낸다.

제안된 재군집화 기법은 이러한 복잡한 환경에서도 상대적으로 안정적인 성능을 유지하며, 일관된 경향을 보여주었다. 예를 들어, Fig. 9의 (r)에서 K-Means의 경우 목표 문서 비율이 5%, 10%인 구간에서 F1 Score가 각각 0.30, 0.60으로 낮은 값을 보인 반면, 제안된 재군집화 기법은 같은 구간에서 각각 0.89, 0.79로 높은 성능을 나타냈다. 이는 기존 기법이 혼합된 환경에서 재군집화를 통해 주제 간 경계를 상대적으로 잘 구분했음을 보여준다. 반면, Fig. 9의 (r), (s), (t) 에서 목표 문서 비율이 35~ 45%인 구간에서는 재군집화 기법이 기존 K-Means보다 오히려 낮은 값을 기록했다. 해당 구간에서는 군집 내부에 다수의 다른 주제 문서가 혼합되어 재군집화 과정에서 의도치 않은 군집 분리가 발생하였고, 이로 인해 주요 주제에 대한 대표성이 오히려 약화하는 현상이 나타났다. 하지만, 목표 문서 비율이 50~80% 구간에서 정확도와 F1 Score가 0.9 이상이고 ARI와 NMI는 0.6 이상의 결과를 나타냈다. 낮은 성능을 보였던 다른 구간에 비해 상대적으로 양호한 결과를 확인했다. 이처럼 재군집화 기법은 복잡한 데이터 환경에서도 일정 수준 이상의 성능을 유지하므로 문서 수집 시에 검색 조건을 추가하거나 사전 필터링 과정을 통하여 목표 문서 비율을 조절할 경우 실용적으로 적용할 수 있음을 보여준다.

ARI와 NMI는 정확도나 F1 Score 보다 목표 문서 비율 변화에 더 민감하게 반응했고, 극단 비율 구간에서는 무작

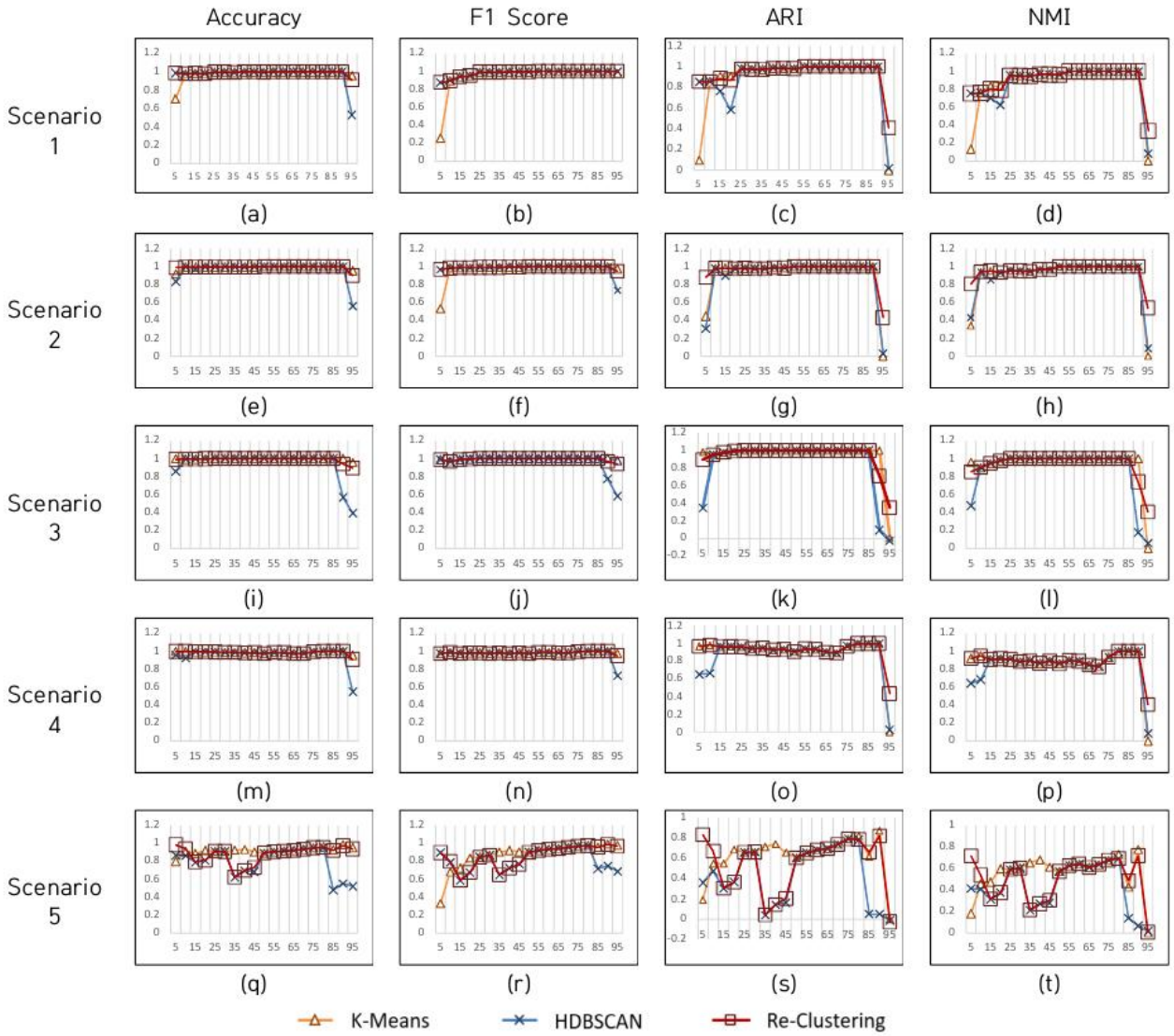

Fig. 9. Comparative analysis by scenario

위 분류에 가까운 낮은 값을 기록했다. 예를 들어, Fig. 9의 (b), (f)에서는 목표 문서 비율이 5%일 때 K-Means 기법 F1 Score가 각각 0.25, 0.53으로 낮은 값을 나타냈고 95%일 때는 0.98, 0.97로 비교적 높은 값을 보였지만, (c), (d), (g), (h)에 해당하는 ARI와 NMI 지표는 5% 구간에서 각각 0.09, 0.13, 0.44, 0.33으로 매우 낮은 값을 보였으며, 95% 구간에서는 모든 값이 0.00으로 하락하였다. 이는 ARI와 NMI가 군집 간의 구조적 분리에 민감하게 반응함을 보여준다. 특히 95% 구간에서는 대부분 문서가 동일한 주제로 구성되어 단일 군집으로 집중되었으며, 이 결과는 의미 있는 주제 구분이 이루어지지 않은 무작위에 가까운 군집화로 해석된다.

Table 4. Two-Way ANOVA Results

Bechmark		SS	df	MS	F	P-value
Accuracy	Ratio	1.4551	94	0.0154	2.6618	6.33E-09
	Method	0.1287	2	0.0643	11.0726	2.84E-05
F1 Score	Ratio	1.8808	94	0.0200	3.6403	2.58E-14
	Method	0.0163	2	0.0081	1.4864	0.228816
ARI	Ratio	18.0332	94	0.1918	10.9216	7.34E-43
	Method	0.2674	2	0.1337	7.6131	0.000662
NMI	Ratio	17.5355	94	0.1865	16.1903	8.39E-56
	Method	0.1435	2	0.0717	6.2311	0.002398

문서 비율과 군집화 기법이 군집화 성능에 미치는 영향을 분석하기 위해 이원 분산분석(ANOVA)를 수행했다. Table 4는 네 가지 평가지표에 대한 분석 결과를 보여준다.

Table 5. Number of Outlier Documents by Scenario

Scenario	Rate of Target Document																		
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Scenario 1	4	6	17	56															52
Scenario 2	99	2	13																49
Scenario 3	100	9																50	67
Scenario 4	32	55																5	50
Scenario 5	173	96				6		9	11							6	63	50	49

이원 분산분석 결과, 네 가지 평가지표 모두 문서 비율에 따라 통계적으로 유의미한 성능 차이를 보였다($p < 0.01$). 이 중 정확도, ARI, NMI는 군집화 기법에 의한 차이도 유의미한 것으로 나타났으며, 특히 ARI와 NMI는 두 요인 모두에 대해 민감하게 반응하였다. 반면, F1 Score는 군집화 기법에 따른 차이는 유의하지 않았다($p > 0.05$). 이러한 결과는 평가지표의 특성과 군집화 기법의 선택이 성능 평가에 중요한 변수임을 나타낸다.

군집화 성능 평가에서 세 가지의 주요 관찰 내용을 요약할 수 있다. 첫째는 목표 문서의 데이터 비율에 따라 성능 영향을 미친다. 둘째는 평가지표별 민감도에 차이가 있으며 ARI, NMI이 더 민감하다. 셋째는 시나리오 5와 같이 여러 주제가 혼합되거나 노이즈가 많은 환경에서는 복잡하고 민감한 결과를 보였다. 이는 데이터의 분포 특성과 군집화 기법을 고려하여 접근할 필요가 있음을 보여준다.

2. Results of re-clustering

재군집화 과정은 버토픽 기반 군집화 결과에서 이상치(outlier)로 군집된 문서를 별도로 추출하고, 동일한 임베딩 모델과 HDBSCAN 알고리즘을 적용하여 다시 군집화를 수행하는 방식으로 구성되었다. 재군집화 대상은 이상치 문서의 수가 10개 이상인 경우를 임계값으로 하여 문서 집합을 선정했다. 각 시나리오별 이상치 문서 수와 재군집화 적용 여부는 Table 5에 정리되어 있다, 이중 음영 표시가 없는 항목이 실제 재군집화가 적용된 그룹이다. 전반적으로 목표 문서의 비율이 극단적으로 낮거나 높은 경우에 이상치 문서가 많이 발생했다. 특히 목표 문서 비율이 5% 수준으로 낮은 경우에 재군집화 대상 문서 수가 집중되는 경향을 보였다.

Fig. 10은 재군집화 전후의 성능 평가지표 변화를 시각화한 결과이다. x축은 재군집화 전의 성능 값을 의미하고, y축은 재군집화 후의 성능 값을 의미한다. 그래프의 각 점은 색깔에 따라 평가지표를 구분한다. 대각선은 성능 변화의 기준선을 나타낸다. 대각선의 위쪽 파란색 영역은 재군집화 이후 성능이 향상된 경우이고 대각선의 아래쪽 빨간색 영역은 성능이 하락한 경우를 나타낸다. 전반적으로 대부분 점이 기준선 위에 분포하고 있어 재군집화가 목표 문서 그룹 식별에 긍정적인 영향을 주었음을 알 수 있다.

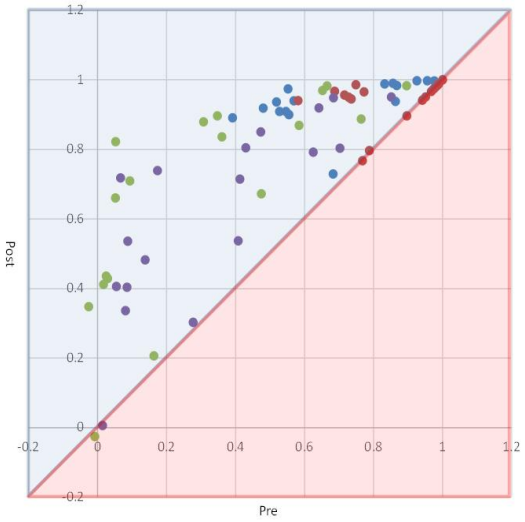

Fig. 10. Comparative analysis of pre and post re-clustering

정확도(파란색)와 F1 Score(빨간색)는 기준선 위의 0.8 이상 구간에 밀집되어 있으며, 재군집화 기법이 목표 문서 선별에 효과적임을 보여준다. 예를 들어, 시나리오 3의 95%구간에서 정확도는 0.39에서 0.89로, F1 Score는 0.58에서 0.94로 상승했다. 시나리오 5와 같이 노이즈가 많은 환경에서도 F1 Score가 평균 0.128 증가하는 등 전반적인 성능 향상이 확인되었다.

ARI(녹색) 및 NMI(보라색)는 상대적으로 넓은 분포를 보이지만, 대부분 점은 기준선 위에 위치하며 유의미한 개선이 이루어졌음을 보여준다. ARI와 NMI값이 0.2 이하에서 0.4 이상으로 향상되는 경우도 다수 있으며 재군집화가 기존 군집 구조의 불완전성을 보완하고 있음을 보여준다. 결과적으로 제안된 기법은 목표 문서 선별의 정밀도를 높일 가능성을 보여준다.

다만, 일부 ARI와 NMI 평가지표 값은 대각선 아래에 위치하며 군집화 성능이 소폭 감소한 것으로 나타났다. 이는

주로 지표의 절대값이 낮은 구간에서의 발생한 미세한 차이에 기인한 것으로 해석된다. ARI와 NMI는 무작위 군집화 결과를 기준으로 성능을 측정하는 지표로 0에 근접한다는 것은 군집화 결과가 무작위에 가까운 수준임을 의미하므로 전후 비교에서 개선 또는 감소가 있더라도 절대 성능이 무작위 수준에 머무른다면, 해당 비교의 유의미성은 제한적일 수 있다.

성능 개선 효과는 목표 문서 비율과 밀접한 연관을 보였다. 목표 문서 비율이 극단적이면 이상치로 군집된 문서가 많았으며, 이들에 대해 재군집화를 수행한 결과, 문서 비율이 낮은 경우보다 높은 경우에서 더 큰 성능 개선 효과가 나타났다. 특히 ARI 지표에서의 개선 폭이 두드러졌는데, 시나리오 1, 2, 3, 4의 95% 구간에서 각각 0.395, 0.400, 0.374, 0.223의 절대값 향상이 있었다. 반면, 시나리오 5의 95% 구간에서는 성능 향상이 소폭에 그쳤다. 재군집화 기법이 목표 문서가 일정량 이상 확보된 조건에서 더욱 효과적임을 시사한다. 초기 문서 수집 시 검색 조건 및 필터링 과정을 정교하게 설정한다면, 실질적인 목표 문서 확보와 군집화 성능 향상을 기대할 수 있다.

Table 6. Paired t-test Results

Metric	Mean (Pre)	Mean (Post)	Diff (Post-Pre)	t-value	p-value
Accuracy	0.7251	0.9421	0.2170	5.1002	4.45E-05
F1 Score	0.8449	0.9423	0.0974	3.2076	0.002581
ARI	0.3024	0.6651	0.3626	7.0829	9.19E-07
NMI	0.3444	0.6250	0.2805	6.7688	1.64E-06

재군집화 기법의 효과성을 검증하기 위하여, 재군집화 전후의 성능 지표 간 평균 차이에 대해 대응표본 t-검정을 시행했다. 정확도, F1 Score, ARI, NMI의 네 가지 평가지표별로 재군집화 전후의 측정값을 대응시켜 검증했다. Table 6은 검정 결과를 보여준다. 검정 결과, 네 가지 평가지표 모두 통계적으로 유의미한 성능 향상이 확인되었다. 구체적으로, 정확도는 평균 0.7251에서 0.9421로 증가하였고($t = 5.10$, $p < 0.001$), F1 Score는 0.8449에서 0.9423으로 향상되었으며($t = 3.21$, $p < 0.01$), ARI와 NMI 또한 각각 0.3626, 0.2805만큼 증가하여, 모두 $p < 0.01$ 수준에서 통계적으로 유의미한 결과를 나타냈다. 이러한 결과를 통해 재군집화 기법이 군집 품질 향상과 문서 선별의 정밀도 개선에 기여할 수 있음을 통계적으로 확인할 수 있었다.

종합하면, 군집화 성능 평가는 단순한 수치 비교를 넘어서 데이터 분포 특성과 평가지표의 민감도에 대한 해석이 함께 이루어져야 한다, 제안된 재군집화 기법은 목표 주제의 문서가 일정 비율 이상 확보된 조건에서 효과적인 성능 향상을 보였고, 복잡한 환경에서도 기존 방법 대비 유의미한 개선 가능성을 입증했다.

V. Conclusions

본 연구에서는 목표 주제의 문서가 다양한 비율로 분포하는 환경에서 버토픽 기반 군집화 성능을 분석하고, 이상치로 분류된 문서들을 재군집화하여 문서 선별 성능을 개선하는 방법을 탐구했다. 문서 분포가 극단적으로 적거나 많은 편향된 경우에도 재군집화 기법이 효과적으로 작용하는지를 평가했다.

실험 결과, 재군집화 방법을 적용함으로써 목표 주제 문서가 보다 의미적으로 일관된 군집을 형성할 수 있음을 확인했다. 다만, 일부 실험에서 ARI 및 NMI 지표가 감소하는 현상이 관찰되었으며, 이는 해당 지표가 극단적인 데이터 분포에서 군집 품질을 적절히 반영하지 못할 가능성이 있음을 시사한다. 따라서 군집화 성능 평가에서는 절대적인 수치 비교보다는 군집의 의미적 일관성과 실질적인 정보 이득을 종합적으로 고려할 필요가 있다.

본 연구는 기존 버토픽 기반 군집화의 한계를 보완하기 위해 재군집화 기법을 적용하여 문서 선별 성능을 평가했다. 기존 연구에서는 목표 주제의 문서가 희소하거나 과다한 경우 군집화 성능 저하를 해결하는 방법에 대한 논의가 부족했다. 본 연구는 이러한 문제를 해결하기 위해 이상치로 분류된 문서를 효과적으로 재배치하는 방법을 제안함으로써, 의미적으로 일관된 군집을 형성하고 문서 선별의 정확도를 향상시킬 수 있음을 보였다.

본 연구에서는 제한된 데이터셋을 사용하여 실험을 진행하였으며, 특정 도메인에 최적화된 결과를 도출했다. 따라서 다른 유형의 문서에서도 동일한 효과가 나타나는지는 추가 검증이 필요하다. 또한, 재군집화 과정에서 사용된 HDBSCAN의 최소 군집 크기 등의 기준값이 연구자가 설정한 값에 따라 결과가 달라질 수 있어 파라미터의 최적화 방안이 향후 연구에서 고려되어야 한다.

향후 연구에서는 제안된 재군집화 기법의 자동화 및 최적화 방안을 모색하고, 버토픽 기반 군집화와 다른 토픽 모델링 기법과의 비교 분석을 통해 재군집화 방법의 일반화 가능성을 검증할 필요가 있다. 다양한 도메인에서 본 연구의 기법을 적용하여 실험을 진행함으로써, 실용적인 활용 가능성을 높이기 위한 추가 연구가 필요하다.

REFERENCES

[1] Jain, Anil K., "Data clustering: 50 years beyond K-means." Pattern recognition letters, Vol. 31, No. 8, pp. 651-666, 2010.

[2] Campello, Ricardo JGB, Davoud Moulavi, and Jörg Sander, "Density-based clustering based on hierarchical density estimates." Pacific-Asia conference on knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160-172, 2013.

[3] Grootendorst, Maarten, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794, 2022.

[4] BERTopic Project, https://github.com/MaartenGr/BERTopic

[5] Reimers, N., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." arXiv preprint arXiv:1908.10084, 2019.

[6] Vinh, Nguyen Xuan, Julien Epps, and James Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?." Proceedings of the 26th annual international conference on machine learning, pp. 1073-1080, 2009.

[7] Na, Sang-Tae, et al., "Trend analysis using topic modeling for simulation studies." Journal of the Korea society for simulation, Vol. 25, No. 3, pp. 107-116, 2016.

[8] O'Mara-Eves, Alison, et al., "Using text mining for study identification in systematic reviews: a systematic review of current approaches." Systematic reviews, Vol. 4, pp. 1-22, 2015.

[9] Lee, Woon-Kyo, and Ja-Hee Kim, "Re-Clustering Documents to Enhance Search Accuracy with Imbalanced Abbreviation Data." Tehnički vjesnik, Vol. 31, No. 6, pp. 1845-1858, 2024.

[10] Ding, Hongxu, Wanxin Wang, and Andrea Califano, "iterClust: a statistical framework for iterative clustering analysis." Bioinformatics, Vol. 34, No. 16, pp. 2865-2866, 2018.

[11] Palanivinayagam, A., and Nagarajan, S., "An optimized iterative clustering framework for recognizing speech." International Journal of Speech Technology, Vol. 23, pp. 767-777, 2020.

[12] Röder, M., Both, A., and Hinneburg, A., "Exploring the space of topic coherence measures." In Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399-408, February 2015.

[13] Xu, R., and Wunsch, D., "Survey of clustering algorithms." IEEE Transactions on neural networks, Vol. 16, No. 3, pp. 645-678, 2005.

[14] Seobin Yoon, and Namgyu Kim, "Document Classification Methodology Using Autoencoder-based Keywords Embedding." Journal of the Korea Society of Computer and Information , Vol. 28, No. 9, pp35-46, 2023.

[15] https://ko.wikipedia.org/wiki/MSA

Authors

Woon-Kyo Lee received the B.S. degree in Computer Science and Engineering from Catholic Kwandong University, Korea, in 1997 and the M.S. degree from the Graduate School of IT Policy, Seoul National University of Science and Technology, Korea, in 2018. He is currently a Ph.D. candidate at the Graduate School of IT Policy, Seoul National University of Science and Technology Seoul, Korea. Since 2007, he has been serving as the Head of Division at Korea Financial IT Inc. He is interested in requirement engineering and text analytics.

Ja-Hee Kim received the B.S. and M.S. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea, in 1995 and 1997, respectively, and the Ph.D. degree in industrial engineering from KAIST, in 2003. She is a Professor with the Graduate School of Public Policy and Information Technology, Seoul National University of Science and Technology, Seoul, Korea. Her current research interests include data visualization, requirement engineering, text analytics and AI standards.