문서 클러스터링을 위한 학술지 논문의 구조적 초록 활용성 연구

Usability Analysis of Structured Abstracts in Journal Articles for Document Clustering

최상희(Sanghee Choi)*
이재윤(Jae Yun Lee)**

초 록

구조적 초록은 학술 논문의 주제를 표현하는 역할을 하여 학술 논문을 처리하는데 중요한 요소로 인식되어왔다. 이 연구에서는 구조적 초록을 구성하는 세부 필드의 속성을 4개로 분석하고 초록의 구조를 활용하여 문서 클러스터링에 적용할 수 있는 가능성을 고찰고자 하였다. 구조적 초록의 필드 속성을 문서 클러스터링에 적용한 결과 클러스터링 기법간의 편차가 있었으나 연구 목적이 제공하는 정보량에 비해 주제성이 커서 클러스터링 성능에 가장 큰 영향을 미치고 있는 것으로 나타났다. 또한 분석 결과 특정 필드에 특화되어 출현하는 필드 종속적인 단어가 발생하는 것으로 나타나 필드 종속적인 단어를 배제하고 집단내 평균연결 기법을 적용하였을 때는 클러스터링의 성능이 개선되는 것으로 분석되었다.

ABSTRACT

Structured abstracts have been regarded as an essential information factor to represent topics of journal articles. This study aims to provide an unconventional view to utilize structured abstracts with the analysis on sub fields of a structured abstract in depth. In this study, a structured abstract was segmented into four fields, namely, purpose, design, findings, and values/implications. Each field was compared in the performance analysis of document clustering. In result, the purpose statement of an abstract affected on the performance of journal article clustering more than any other fields. Furthermore, certain types of keywords were identified to be excluded in the document clustering to improve clustering performance, especially by Within group average clustering method. These keywords had stronger relationship to a specific abstract field such as research design than the topic of an article.

키워드: 구조적 초록, 학술지 초록, 문서 클러스터링, 클러스터링 자질, 클러스터링 기법
structured abstract, journal abstract, document clustering, clustering feature, clustering method

* 대구가톨릭대학교 도서관학과 교수(shchoi@cu.ac.kr) (제1저자)
** 경기대학교 문헌정보학과 교수(memexlee@kgu.ac.kr) (공동저자)
■ 논문접수일자 : 2012년 3월 10일 ■ 최초심사일자 : 2012년 3월 11일 ■ 게재확정일자 : 2012년 3월 21일
■ 정보관리학회지, 29(1), 331-349, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.1.331〕

1. 서 론

학술논문에서 초록은 독자가 가장 먼저 읽는 정보이자 동시에 유일하게 읽는 정보가 되는 중요한 요소이다. 초록은 학술 논문의 내용과 주제를 표현하는 가장 적절한 정보로 인식되기 때문이다. 초록은 독자들이 적합한 학술논문을 선정하는데 활용되기도 하고 학술논문을 주제 색인하는 데 활용되기도 하다. 그러나 초록은 저자들이 직접적으로 작성하는 정보이기 때문에 학술논문의 주제가 가장 잘 표현되기도 하지만 저자에 따라 제공되는 정보량이 달라지며 포함되는 내용에서도 편차가 나타나고 있다.

이와 같은 문제점을 개선하기 위하여 초록의 내용과 질적 일관성을 유지하기 위하여 학술지 초록에 세부 구조를 도입하는 시도가 80년대부터 시작되어 90년대에 이르러 활성화되기 시작했다. 구조적 초록은 다양한 연구를 통해서 비구조적인 초록에 비해 제공하는 정보량이 증가된 것으로 평가되었으며 검색 시 적합한 논문을 선택하는데 도움이 되었고 가독성을 향상시킨 것으로 나타났다. 그러나 구조적 초록을 분석한 대부분의 연구는 구조적 초록 전체의 효과성이나 가치를 평가한 경우가 대부분이고 구조적 초록의 세부 구조에 따라 제공하는 정보의 특성을 고찰하거나 구조적 초록을 활용하여 정보처리 성능이나 검색 효율을 향상시키는 방안을 제시한 경우는 거의 없었다. 이에 이 연구에서는 구조적 초록을 구성하는 세부 요소의 특성을 분석한 후 구조적 초록을 활용하여 문서 클러스터링의 성능을 개선시킬 수 있는 가능성을 타진하고자 한다.

구조적 초록의 세부 구성 요소를 분석하기 위하여 영국 Emerald 데이터베이스에서 제공하는 학술지 Library High Tech와 Library Review의 최근 5년간 게재된 학술논문을 수집하여 15개 주제 집단으로 재구성하여 실험집단을 구축하였다. 구축된 실험집단의 구조적 초록을 대상으로 초록을 구성하는 세부요소의 특성을 분석하기 위하여 필드 간 상관관계를 분석하였다. 또한 구조적 초록의 세부 필드가 문서 클러스터링에 미치는 영향력을 평가하기 위하여 각 필드를 선택적으로 배제하면서 계층적 클러스터링 기법으로 문서 클러스터링을 수행하였고, 필드를 배제한 각 경우별로 문서 클러스터링 개선여부를 평가하였다. 이 과정에서 분석된 결과를 토대로 문서 주제 클러스터링의 성능을 개선하는 목적으로 구조적 초록의 세부 필드 특성 중 특정 필드에 종속된 용어를 불용어로 처리하는 방안이 도출되었다.

2. 선행연구

학술논문의 초록은 일반적으로 연구배경, 연구 목적, 연구 결과, 방법 등 구체적인 요소들을 포함하고 있다. 따라서 초록의 구조도 이 요소들을 기반으로 형성된다. 영국의 Emerald 학술 데이터베이스가 제공하는 주요 학술지에서는 학술논문 초록의 구조를 목적(purpose), 연구 설계 및 방법(design/methodology), 결과(findings), 연구한계 및 시사점(research limitations/implications), 실용적 시사점(practical implications), 독창성 및 가치(originality/value)로 설정하여 필드 구분을 하여 초록을 처리하고 있다.

세부학문 분야별로 구조적 초록이 학술논문

에 포함되는 동향을 살펴보자면 의학 분야에서는 구조적 초록이 80년대 중반에 소개된 이후 90년대부터 활성화되고 있는 것으로 인식되고 있다. 구조적 초록이 의학 분야 학술지에 적용된 현황을 분석한 2005년 논문에서는 의학 분야에서 impact factor가 상위 30위 이내인 304개 학술지의 61.8%가 구조적 초록을 제공하고 있는 것으로 나타나 의학 분야에서 구조적 초록이 일반화되고 있는 현상을 말해주었다(Nakayama et al., 2005).

Medical Library Association에서는 발행하는 의학 학술지인 Journal of the Medical Library Association에 실리는 학술논문 초록의 기본 구조를 목적(objective), 연구방법(methods), 결과(results), 결론(conclusions)으로 제시하고 있으며 추가적으로 배경(background), 해석(interpretation) 등을 기술할 수 있도록 하였다. 또한 사례연구일 경우 연구문제(question), 사례환경(setting) 등으로 대치하여 구조적 초록을 생성하도록 안내하고 있다.

사회과학분야 사례로 교육학 분야 학술지 Educational Research는 실증연구논문의 초록 구조를 배경(background), 목적(purpose), 프로그램개요(program description), 사례(sample), 연구 설계 및 방법(design/methods), 결과(results), 결론(conclusion)으로 설정하였다. 그리고 문헌연구의 초록구조는 배경(background), 목적(purpose), 연구 설계 및 방법(design/methods), 결론(conclusion)으로 설정하여 연구의 성격을 표현할 수 있는 형태로 초록 구조를 최적화하는 방안을 제시하였다.

구조적 초록의 주요 연구자 Hartely는 구조적 초록의 효과성을 사회과학 분야(1997), 비의학 분야(1998), 심리학 분야(2003)에서 조사하였는데 학문적 특수성에 상관없이 비구조적인 초록에 비해 구조적 초록의 가장 큰 개선점은 정보가 더 많이 포함된다는 것이다.

90년대부터 2000년대에 수행된 구조적 초록을 분석한 연구를 리뷰한 결과 전통적 초록과 비교하였을 때 구조적 초록의 효과는 다음과 같다(Hartely, 2004).

- 정보량의 증가
- 가독성 향상
- 검색 편이성 개선
- 학술대회논문집(conference proceedings)의 경우 동료심사(peer review)의 편이성 개선 검색 재현율 개선
- 독자와 저자 호응도 향상

이러한 구조적 초록의 효과는 초록을 연구하는 여러 연구자들에게 다양한 측면에서 분석되고 측정되었다. Sharma와 Harrison은 90년대에 전통적 초록 형태에서 구조적 초록 형태로 바꾼 학술지와 전통적 초록을 유지하는 학술지를 선정하여 구조적 초록을 바꾼 시기를 기점으로 전후 1년 치 학술논문을 추출하여 28개 평가기준으로 적용하여 질적 평가를 하였다(Sharma & Harrison, 2006). 평가 결과 전통적 초록에서 구조적 초록으로 변화한 학술지에서 주목할 만한 정보의 질적 개선이 나타난 것으로 밝혀졌다. 전통적 초록 형태의 학술지와 구조적 초록 형태의 학술지 비교에서도 구조적 초록 형태의 학술지 초록이 정보를 제공하는 측면에서 더 좋은 평가를 받았다.

구조적 초록의 질적 평가로 시작된 연구는 구조적 초록의 기능성에 대한 연구로 점차 확

장되어 왔다. Stevenson과 Harrison은 치의학 저널에서 구조적 요약이 임상실험 자료를 검색하는 성능을 개선시킬 수 있는지 조사하였는데 조사결과 정확도나 재현율을 직접적으로 개선하고 있지 못한 것으로 나타났다(Stevenson & Harrison, 2009). 그러나 교육 분야의 학술지에서 구조적 초록이 이용자의 정보 습득에 미치는 영향을 파악한 연구에서는 전통적인 초록을 포함한 학술논문이 이용자에게 선택되지 않은 경우가 구조적 초록이 포함된 논문보다 두 배 가까이 더 많은 것으로 나타나 구조적 초록이 정보 검색 과정에서 적합한 정보를 선정하는데 영향을 미치고 있는 것으로 분석되었다(Hahs-Vaughn & Onwuegbuzie, 2010).

현재까지 구조적 초록에 대한 연구는 구조적 초록의 효과와 의의에 초점을 맞추어 수행되어 왔고 구조적 초록은 전통적 초록의 기능과 질적 수준을 개선시키고 있는 것으로 판명되어 왔다. 그러나 구조적 초록을 구성하는 요소간의 관계와 같이 구조적 초록을 심층적으로 분석한 사례나 구조적 초록을 이용해서 다른 정보처리 성능을 개선시키는 등 다양한 측면에서 적용하여 구조적 요약의 활용성을 분석한 사례는 없는 것으로 조사되었다. 다만, 초록이 아닌 본문의 지식구조를 분석하여 온톨로지로 표현하려는 시도(고영만, 송인석, 2011)가 있었을 뿐이다. 이와 같은 맥락으로 이 연구에서는 구조적 초록을 구성하는 각 요소간의 상관관계를 조사하여 구조적 초록을 세분화하여 활용할 수 있는 방안을 모색하고자 하였으며 구조적 초록을 문서 클러스터링에 적용하여 구조적 초록의 세부 요소가 문서의 주제성을 대표하는 성능에 미치는 영향을 고찰하고자 하였다.

최근 문서 클러스터링 연구에서는 문서 클러스터링에 적합한 자질을 선정하는데 다양한 접근 방법을 제시하고 있다. 이는 클러스터링 대상이 되는 문서에서 단어를 추출하여 비교할 때 유사도 산출대상이 많아지면서 계산량이 증가해짐에 따라 속도가 저하되고 노이즈에 속하는 단어들이 포함됨으로써 오히려 정확성이 떨어지게 되기 때문에 클러스터링의 성능을 저하시키지 않으면서 클러스터링 성능을 개선시킬 수 있는 주제성이 있는 자질을 선정하는 것이 관건이 되기 때문이다. 이러한 문제점을 해결하기 위해서 클러스터링 자질이 문서의 주제를 대표하는 정도를 향상시키기 위하여 클러스터링 자질로 추출한 단어와 WordNet을 결합하여 문서 클러스터링 성능을 개선시키려는 시도도 있었으며(Chen, Tseng, & Liang, 2010), 학술논문과 신문기사를 구성하는 각 필드 특성을 적용하여 필드 독립적인 클러스터링을 수행하여 성능평가를 한 사례도 있다(Zhu, Takigawa, & Mamitsuka, 2009). 문서의 단어 속성이 클러스터링에 미치는 영향을 연구한 최근 국내사례에서는 복합명사가 문서 클러스터링에 미치는 영향력을 조사하였는데 이는 복합명사가 특정 문서에 대한 주제적 변별력을 많이 내포한다는 가정을 적용한 사례이다(조현양, 최성필, 2004). 또 다른 국내 사례에서는 문서 클러스터링을 위하여 보다 적절한 클러스터 자질을 추출하여 문서 클러스터링의 품질을 개선하기 위하여 개체명을 이용하는 방법을 제시하였다(윤보현, 오효정, 2010). 이 연구에서는 ETRI에서 개발된 질의응답을 위한 개체명 인식기(NER: Named Entry Recognizer)를 사용하여 사람, 지역, 기관 등에 해당하는 개체명을 이용하여 상품리뷰

와 경제문서집단을 대상으로 개체명이 클러스터링에 미치는 영향을 분석하였고 개체명의 유형별로 조합하여 클러스터링 효과를 분석하였다. 이와 같은 시도는 문서내 모든 키워드를 동일하게 취급하지 않고 키워드의 속성을 반영하여 문서 클러스터링에 개체명이 미치는 영향을 파악하고자 한 것이다.

최근 문서 클러스터링 연구에서 나타났듯이 문서 내 모든 용어를 차별없이 통합하여 클러스터링에 적용하기 보다는 세부 특성을 분석하여 클러스터링에 반영하려는 시도가 다각적으로 나타나고 있다. 구조적 초록의 경우에도 초록 구조에 따라 독립된 세부 초록 필드에서 나타나는 용어들 간에 학술논문의 주제를 대표하는 정도에 차이가 있을 수 있으며 이와 같은 차이가 문서 클러스터링에 영향을 미칠 수 있다는 가정이 성립될 수 있으므로 이 연구에서는 세부 초록 필드를 적용하여 문서 클러스터링 연구를 수행하고자 하였다.

3. 연구 개요

3.1 데이터 수집 및 실험집단 구성 개요

학술논문의 초록은 일반적으로 연구 목표, 방법, 결과, 의의 등 특정 사항들이 포함되어 있다. 영국의 학술데이터베이스 Emerald에서 제공하는 학술지 Library High Tech와 Library Review는 이러한 학술논문의 특성을 반영하여 저자들로 하여금 학술논문의 초록을 구조적으로 작성하도록 하고 있다. 초록을 구성하는 세부 필드간의 상관관계와 세부 필드의 특성을 문서 클러스터링에 적용하여 초록의 세부 요소들이 문서 주제 클러스터링에 미치는 영향을 조사하기 위하여 구조적 초록을 제공하는 두 저널에 게재되어 있는 최근 5년간 학술 논문을 수집하였다. 실험집단으로 Library Hi Tech에서 292건, Library Journal에서 210건을 수집하여 총 502건의 학술논문을 수집하였다. 이 중 리뷰성 논문과 북리뷰는 제외하고 구조적 초록을 포함한 논문 445개를 후보 집합으로 추출하였다. 추출된 후보집합의 연도별 각 학술지 논문 수는 〈표 1〉과 같다.

수집된 구조적 초록이 있는 학술논문 중에서 문서 주제를 기준으로 초록 필드의 상관성 분석과 문서 클러스터링 실험을 수행하기에 적합한 주제 문서들을 선택하기 위하여 각 학술논문에 부여된 주제어를 분석하였다. 수집된 445건의 문서를 LISA 데이터베이스에서 검색하여 각 문헌에 부여된 주제어로 통제어휘인 디스크립터와 저자가 부여한 키워드를 파악하여 주제문서 클러스터를 구축하는 기준으로 이용하였다.

〈표 1〉 5년간 두 학술지에 개재된 논문 중 구조적 초록을 포함한 논문 수

학술지명	2007	2008	2009	2010	2011	총계
Library High Tech	49	51	47	47	41	235
Library Review	53	46	46	27	38	210
총계	102	97	93	74	79	445

총 445개 학술논문의 주제어 중 10회 이상 부여된 용어를 추출한 결과 총 28종의 주제어가 도출되었다. 학술논문이 중복 주제 카테고리에 소속되는 현상을 최소화하기 위하여 주제어의 출현빈도에 비해서 다른 주제어와 같은 학술논문에 부여되는 빈도가 낮은 주제어를 선택하였다. 주제어의 경우에도 빈도수가 높은 주제어들 중에는 범용적으로 사용되는 주제어가 있어 이 경우 주제가 다른 여러 문헌에 동시 부여되는 경우가 많다. 이러한 주제어들을 대상으로 문서 클러스터링 실험을 위한 주제별 문서집합을 구축하게 될 경우 복수 주제 카테고리에 소속되는 문서 수가 증가되어 클러스터링의 성능을 비교할 수 없게 된다.

이와 같은 문제점을 방지하기 위하여 이 연구에서는 28종 중 다른 키워드와 동일 문서에 출현하는 빈도를 출현빈도로 나눈 값을 평균 중복도로 설정하여 측정하여 주제어 당 문서의 중복도를 분석하였다. 그 결과 평균 중복도가 1 이하, 즉 다른 주제어와 문서가 중복되는 정도가 문서 1건 당 평균 1회 이하인 주제어를 문서 클러스터링 실험을 위한 주제 문서집단을 형성하는데 적합한 주제어로 선정하게 되었다. 이와 같은 주제어 선정과정을 통해 주제적 배타성이 상대적으로 높은 총 15개의 주제어가 선정되었다.

15개 주제어가 부여 되어 있는 학술 논문 수는 총 262인데 이 경우에도 15개의 주제어 중 주제어가 둘 이상 부여되어 있어 동일 문서가 주제별 문서클러스터에 중복되어 나타나 있는 경우가 발생하여 해당되는 중복 출현한 문서를 제외하고 선정된 주제어가 1개씩만 부여된 문헌을 선별한 결과, 〈표 2〉 실험 문서 집단 주제 클러스터와 같이 총 161건이 최종적으로 실험 집단으로 구축되었다.

〈표 2〉 실험 문서 집단 주제 클러스터

문서 클러스터 번호	클러스터 주제명(주제어)	클러스터 주제	논문 수
C01	Ethics	윤리	8
C02	Collections management	장서관리	7
C03	Knowledge management	지식관리	17
C04	Reading	독서, 독서지도	6
C05	Reference services	참고봉사	6
C06	Information literacy	정보활용교육	16
C07	Serials	연속간행물, 학술저널	9
C08	Communication technologies	정보통신기술	12
C09	Library users	도서관 이용자	23
C10	Electronic books	전자책	7
C11	Archives	기록, 기록보존소	6
C12	Higher education	대학, 대학도서관	6
C13	Digital preservation	전자 기록 보존	18
C14	Education	도서관 교육	14
C15	Copyright law	저작권	6
총 계			161

3.2 구조적 초록의 세부 구성요소

실험집단으로 구축한 Library High Tech와 Library Review의 초록 세부 구성요소는 연구 목적(purpose), 연구 설계(Design/methodology/approach), 연구 결과(Findings), 연구한계 및 시사점(Research limitation/implications), 실용적 시사점(Practical implications), 독창성 및 가치(Originality/value) 총 6개로 구성되어 있다. 이 중 학술논문이 공통적으로 포함하고 있는 초록 필드는 연구 목적, 연구 설계, 연구 결과이며 시사점 필드와 가치 필드는 논문의 유형별로 포함여부가 달라진다. 이 연구에서는 기본 초록 필드는 각각 독립적으로 적용하여 처리하였으며 시사점과 가치 필드는 학술논문 성격에 따라 부가적인 정보를 제공하는 부가정보로 규정하여 한 필드로 통합하여 처리하였다. 따라서 초록 필드간 상관관계를 분석한 실험과 문서 클러스터링 실험에 적용한 초록 필드는 다음과 같이 총 4개 필드로 설정하였다.

- 연구 목적(Purpose: 이하 p)
- 연구 설계(Design: 이하 d)
- 연구 결과(Findings: 이하 f)
- 가치/의의(Value/Implications: 이하 vi)

3.3 구조적 초록 필드 분석 및 클러스터링 평가 과정

이 연구에서는 다음과 같이 총 3가지 과정을 통해 구조적 초록을 구성하는 세부 필드를 분석하고 문서 클러스터링에서 활용성을 분석, 평가하였다. 구조적 초록의 필드간 상관관계를 분석하여 구조적 초록을 구성하는 요소의 특성을 분석하였고, 각 세부 필드를 배제하면서 문서 클러스터링을 수행하여 각 필드를 배제하였을 때 클러스터링 성능이 변화하는 정도를 파악하였다. 이 과정에서 특정 필드에서 집중적으로 출현하지만 주제성이 떨어지는 단어가 파악되어 이러한 단어들을 활용하여 클러스터링 성능을 개선시킬 수 있는 방안을 모색하기 위하여 필드 종속적 단어를 제거하고 클러스터링을 수행한 후 성능을 평가하였다.

첫째 초록을 구성하는 요소간의 상관관계를 측정하는 실험을 위하여 총 4개로 설정된 초록 필드는 각 초록 필드에 출현한 단어 중 불용어를 처리한 후 색인을 생성하였다. 산출된 색인어를 기반으로 필드 간 상관관계를 측정하는데 SPSS Version 19와 피어슨 상관계수를 이용하였다.

둘째, 구조적 초록의 각 필드가 문서 클러스터링에 영향을 미치는 것을 측정하기 위하여 먼저 초록 필드를 통합하여 초록 전체를 대상으로 클러스터링하였고, 이후에는 초록 전체에서 각 세부 필드를 하나씩 번갈아가면서 제외하고 클러스터링 하였다. 이 실험에서 비교한 케이스는 총 5개로 아래와 같으며 초록 전체를 적용하여 문서 클러스터링을 한 결과와 각 세부 필드를 제외한 결과를 비교하였다.

- 초록 전체를 클러스터링한 케이스(ALL)
- 초록 전체에서 연구 목적을 제외하고 클러스터링한 케이스(not_p)
- 초록 전체에서 연구 설계를 제외하고 클러스터링한 케이스(not_d)
- 초록 전체에서 연구 결과를 제외하고 클러스터링한 케이스(not_f)
- 초록 전체에서 가치/의의를 제외하고 클러스터링한 케이스(not_vi)

셋째, 각 필드별로 색인어를 추출하여 문서 클러스터링을 실험하는 과정에서 특정 용어들이 특정 필드에 집중되어 출현하는 현황이 관찰되었다. 특정 필드에 종속되어 나타나는 단어는 실제 논문의 주제를 표현하기 보다는 특정 필드의 기능을 표현하는 것으로 분석되었다. 따라서 이 연구에서는 초록의 필드를 직접적으로 클러스터링에 적용한 연구에서 확장하여 초록의 필드를 적용하여 주제성이 약한 단어를 문서 클러스터링에서 배제하는 실험을 수행하였다. 이는 주제 클러스터링에 노이즈가 되는 불용어를 초록의 필드를 이용해서 선정하여 주제 클러스터링의 성능을 개선시킬 수 있는지를 조사하기 위한 것이다.

필드에 종속적인 용어, 즉 필드 종속어를 선정하기 위하여 단어가 출현한 문헌빈도를 필드별로 분석하였다. 먼저 전체 실험집단에서 단어의 문헌빈도를 네 필드별로 산출하여 각 단어에 대해서 가장 많은 문헌에서 출현한 필드를 파악했다. 전체 문헌 집합에서 해당 단어가 가장 많이 출현한 필드에서의 문헌빈도로부터 두 번째로 많이 출현한 필드에서의 문헌빈도를 빼서 1위와 2위의 차이를 산출하였다. 즉, 각 단어별로 문헌빈도가 1위인 필드와와 2위인 필드에서의 문헌빈도의 차이를 산출해낸 것이다. 이렇게 산출된 1, 2위 필드별 문헌빈도의 차이가 9회 이상인 단어를 필드 종속적인 단어로 선정하였다. 9회는 전체 문서 수(161)의 약 5%에 해당하는 수치로 두 단어의 필드별 문헌빈도 차이가 전체 논문 수의 5%를 넘어가게 되면 해당 단어가 1위 필드에 특화되어 나타나고 있는 필드 종속어라고 가정하였다. 결국 두 번째로 많이 등장하는 필드에서의 문헌빈도에 비해서 뚜렷하게 많이 출현한 필드가 있다면 해당 단어는 1위 필드에서 기능적인 역할을 하는 필드 종속어라고 간주한 것이다. 예를 들면 단어 A가 연구 목적, 설계, 결과, 가치/의의까지 네 필드에서의 문헌빈도가 각각 5, 15, 2, 0일때, 1위와 2위 필드별 문헌빈도 차이가 10이 된다. 이런 경우 단어 A는 연구 설계 필드에서 유난히 많이 등장하므로 연구 설계를 설명하기 위한 기능어일 가능성이 높다. 이와 같은 과정을 거쳐 주제성이 약하고 기능적인 역할을 주로 하는 필드 종속어를 파악하였다.

필드제외 클러스터링 실험과 필드 종속어 제외 클러스터링 실험에서 수행한 문서 클러스터링에 적용된 기법은 문서 클러스터링에 가장 일반적으로 적용되는 Ward 기법과 집단내 평균연결(Within Group Average) 기법이다. 흔히 평균연결 기법이라고 불리는 것은 집단간 평균연결(Between Group Average) 기법이며 집단내 평균연결 기법은 생성될 클러스터에 속한 케이스 사이의 거리 평균이 최소화되도록 병합할 클러스터 쌍을 선택한다. 집단내 평균연결 기법은 집단간 평균연결 기법에 비해서 생성되는 클러스터의 크기가 상대적으로 고르다는 특징이 있다. 사전 실험에서 SPSS 프로그램이 제공하는 여러 계층적 클러스터링 기법을 적용해 보았으나 Ward 기법과 집단내 평균연결 기법 이외의 다른 기법들은 모두 현저하게 낮은 성능을 보였으므로 이 논문에서는 두 기법을 사용한 경우만 보고하였다.

3.4 클러스터링 성능평가 척도

문서 클러스터링의 성능 평가를 하기 위하여 기준이 되는 주제 집단은 3.1절에서 기술하였

듯이 총 15개의 주제집합으로 구축하였고 클러스터링 평가 척도는 정영미와 이재윤이(2002) 제안한 비편향적인 단일 클러스터링 평가 척도인 WACS와 기준 집단(수작업 분류)과 클러스터링 유사도를 측정하는 CSIM을 포함하여 총 8개 척도를 적용하여 평가하였다. 평가에 적용한 척도는 아래와 같다.

D: 전체 문서 수
m: 미리 규정한 기준 주제 클러스터 수
n: 자동생성한 클러스터 수
Mi: 미리 규정한 기준 주제 클러스터 i
Ci: 자동 생성한 클러스터 I

• 카이 제곱(Chi-square)

$$x^2 = \sum_{i=1}^{i=R} \sum_{j=1}^{j=C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

R: 클러스터링 결과 A에서 나온 클러스터 수
C: 클러스터링 결과 B에서 나온 클러스터 수
O_{ij} : 관찰된 빈도
E_{ij} : 기대된 빈도

• 엔트로피(Entropy)

$$\text{Entropy(Ci)} = \sum_{i=1}^{n} -p_i \log_2 p_i$$

pi: 클러스터 pi에서 할당한 클러스터 C의 문서 비율

• F척도(F measure), 정확률(precision), 재현율(recall)

$$\text{정확률} = \frac{1}{D} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{|M_i \cap C_j|}{|C_j|}$$

$$\text{재현율} = \frac{1}{D} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{|M_i \cap C_j|}{|M_i|}$$

$$F(p,r) = \frac{2pr}{p+r}, \quad (\text{p는 정확률, r은 재현율})$$

• WACS

$$\text{WACS} = \frac{1}{D} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{2|M_i \cap C_j|^2}{|M_i| + |C_j|}$$

• 상호정보량(Mutual Information: MI)

$$I(M:C) = \frac{1}{D} \sum_{i} \sum_{j} |M_i \cap C_j| \log_2 \frac{D|M_i \cap C_j|}{|M_i||C_j|}$$

• CSIM

$$\text{CSIM} = \frac{2a}{2a+b+c}$$

: a는 두 문헌이 수작업 분류와 클러스터링 결과에서 모두 동일 클러스터에 속한 경우의 수
: b는 수작업 분류에서는 한 범주에 속한 두 문헌이 클러스터링 결과에서는 다른 클러스터에 속한 경우의 수
: c는 수작업 분류에서는 다른 범주에 속한 두 문헌이 클러스터링 결과에서는 같은 클러스터에 속한 경우의 수

이중에서 카이제곱통계량과 엔트로피는 범위가 고정되어 있지 않아서 상대적인 평가에 불리하다. 엔트로피는 다른 평가 척도와 달리 값이 작을수록 좋은 성능을 의미한다. 상호정보량은 평가 대상 클러스터링 결과의 클러스터 크기가 수작업 분류의 범주 크기와 현저하게 차이나는 경우를 과대 평가하는 경향이 있다. 이런 경향은 상호정보량과 카이제곱통계량도 가지고 있다. CSIM은 크기가 1인 클러스터가 있는 경우에는 문헌 쌍이 없으므로 평가가 부정확하게 된다. 재현율은 극단적으로 큰 클러스터가 있는 경우에 심하게 편향되는 방향으로 평가가 이루어진다. 따라서 재현율과 정확률의 조합인 F척도도 클러스터 크기의 차이에 영향을 받는다. 이

실험에서는 여타 기법에 비해서 클러스터의 크기가 비교적 고르게 도출되는 Ward 기법과 집단내 평균연결 기법을 사용한 덕분에 크기가 1인 클러스터는 발생하지 않았다. 따라서 사용한 평가 척도 중에서 비교적 클러스터 크기의 차이에 민감하지 않은 지표는 WACS와 CSIM이라고 할 수 있다. 따라서 이후의 평가에서는 모든 지표를 제시하되 성능 비교는 주로 WACS와 CSIM을 위주로 설명하였다. 또한 정확률과 재현율은 상보적인 특성을 지니며 F척도를 산출하기 위한 중간값에 해당하므로 산출결과를 표에 제시하기는 하지만 그림으로 비교할 때에는 혼란을 막기 위해서 제외하였다.

4. 구조적 초록 필드간 상관관계 분석

구조적 초록의 각 필드별 특성을 먼저 필드별 색인단어 수로 분석하였다. 필드별 단어를 형태소 분석한 뒤 불용어를 제거하고 각 필드별로 비교하니 〈표 3〉과 같이 나타났다. 연구 설계가 다른 초록 필드에 비해 가장 색인 대상이 되는 용어 수가 가장 적은 것으로 나타났으며 가치와 의의가 가장 많은 것으로 밝혀졌다. 또한 연구 목적은 연구 결과에 비해 약 30% 정도 적은 길이로 나타나 연구 목적이 연구 결과에 비해 핵심 키워드로 학술논문의 주제를 함축적으로 표현하고 연구 결과는 상대적으로 확장된 용어로 학술논문의 주제를 구체적으로 표현할 가능성이 있는 것으로 조사되었다.

구조적 초록의 각 필드 간 상관관계를 분석한 결과 〈표 4〉와 같이 연구 목적 필드는 연구 설계, 연구 결과, 가치/의의와 모두 0.01 수준에서 상관관계가 있는 것으로 나타났다. 따라서 연구 목적은 결과나 가치 등 주제적인 측면이나 연구 방법과 같은 비 주제적 측면의 정보를 모두 내포하고 있는 것으로 분석되었다. 반면 다른 세 필드는 연구 목적 필드 이외에는 긍정적인 상관관계를 보이는 필드가 없으므로 제시하는 정보가 고유한 정보를 제시하는 것으로 나타났다. 특히 주목할 만한 사항은 연구 설계와 연구 결과의 관계인데 두 필드 사이의 상관계수는 -0.024로서 유의수준 95%에서 부정적 상관관계가 있는 것으로 나타난 것이다. 연구 설계와 연구 결과는 학술논문의 주제를 표현하는데 있어 중복되지 않고 상이한 정보를 포함하고 있기 때문에 한 문서의 초록을 완성한다는 측면에서는 상호 보완적인 구성요소일 수 있다. 그러나 문서간 주제 유사성을 비교하는 목적으로 보면 두 문서의 상이한 필드에 포함된 정보가 해당 필드에서만 유사하다고 해서 전체 주제적으로도 유사한 문서라고 판단하기는 어렵다고 생각된다.

구조적 초록의 각 필드 간 관계를 〈그림 1〉과

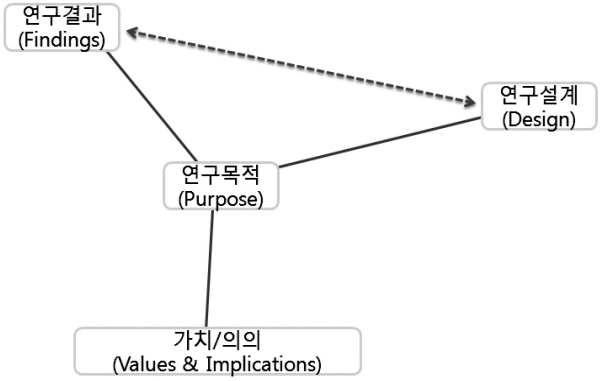
〈표 3〉 구조적 초록 필드별 색인 단어 수

필드	연구 목적 (Purpose)	연구 설계 (Design)	연구 결과 (Findings)	가치와 의의 (Values & Implications)
평균 길이 (색인된 단어 수)	16.2	15.2	24.1	25.7

〈표 4〉 구조적 초록 필드 간 색인 단어 빈도의 상관관계

		초록 전체	연구 목적 (Purpose)	연구 설계 (Design)	연구 결과 (Findings)	가치/의의 (Values & Implications)
초록 전체	Pearson 상관계수	1	.600**	.402**	.549**	.608**
	유의확률(양쪽)		.000	.000	.000	.000
	N	8400	8400	8400	8400	8400
연구 목적 (Purpose)	Pearson 상관계수	.600**	1	.067**	.136**	.163**
	유의확률(양쪽)	.000		.000	.000	.000
	N	8400	8400	8400	8400	8400
연구 설계 (Design)	Pearson 상관계수	.402**	.067**	1	-.024*	.001
	유의확률(양쪽)	.000	.000		.028	.915
	N	8400	8400	8400	8400	8400
연구 결과 (Findings)	Pearson 상관계수	.549**	.136**	-.024*	1	.016
	유의확률(양쪽)	.000	.000	.028		.133
	N	8400	8400	8400	8400	8400
가치/의의 (Values & Implications)	Pearson 상관계수	.608**	.163**	.001	.016	1
	유의확률(양쪽)	.000	.000	.915	.133	
	N	8400	8400	8400	8400	8400

**. 상관계수는 0.01 수준(양쪽)에서 유의함
*. 상관계수는 0.05 수준(양쪽)에서 유의함



〈그림 1〉 다차원척도법(PROXSCAL)으로 나타낸 필드 간 상관관계

같이 도식화 하면 구조적 초록을 구성하는 요소 간의 상관관계가 보다 명확히 나타나고 있다. 위 그림에서 실선으로 표시된 관계는 95% 유의수준에서 긍정적인 상관관계이며 화살표가 붙은 점선으로 표시된 관계는 95% 유의수준에서 부정적인 상관관계가 나타난 경우이다. 그림에서 나타나 있듯이 연구 목적은 구조적 초록을 구성하는 가장 핵심적인 요소로 다른 세부 요소들과 의미 있는 상관관계를 가지고 있다. 따라서 구조적 초록에서 연구 목적을 제외하게 될 경우 초록

의 내용을 효과적으로 파악하기 어려울 수 있다는 해석이 가능하다. 또한 연구 결과와 연구 설계는 연구 목적의 효과성을 보완할 수 있는 구성요소이지만 연구 결과와 연구 설계간 상호 보완은 이루어지지 않으므로 이러한 특성을 반영하여 초록의 세부 필드 활용을 고려해야 한다.

5. 구조적 초록 필드를 활용한 문서클러스터링

5.1 구조적 초록 필드의 영향력 분석

구조적 초록의 각 필드가 문서 클러스터링에 미치는 영향력을 측정하기 위하여 초록 전체를 가지고 문서 클러스터링을 한 경우와 각 초록 필드를 하나씩 제외시키면서 문서 클러스터링 수행 한 결과를 비교하였다. 그 결과 각 필드의 상관관계에서도 관찰되었듯이 연구 목적이 주제 클러스터링에 가장 큰 역할을 하고 있는 것으로 나타났다. 초록 전체를 클러스터링 한 결과와 연구 목적을 제외시키고 클러스터링을 비교한 결과를 비교하면 〈표 5〉에서 나타나듯이 Ward 기법에서 클러스터링 성능이 모든 평가척도에서 크게 떨어진 것으로 나타났으며 집단내 평균연결 기법에서는 카이 제곱과 재현율, WACS에서 성능이 떨어진 것으로 나타났다. 초록 전체를 사용하여 클러스터링 하였을 때 Ward 기법보다 집단내 평균연결 기법을 사용한 경우에 성능이 나쁘게 나타난 이유는, 대부분의 평가척도가 클러스터의 크기 차이가 심한 경우에 유리하게 판단하는 편향성을 가지고 있는데 실험에서 집단내 평균연결은 수작업 분류와 비슷한 크기의 클러스터를 형성한 반면 Ward 기법은 매우 큰 클러스터와 매우 작은 클러스터가 뒤섞인 결과를 산출했기 때문이다. 그럼에도 불구하고 일부 필드를 제외한 경우에는 집단내 평균연결 기법의 성능이 Ward 기법의 성능보다 여러 경우에서 좋게 나타난 이유는 Ward 기법이 이상치(outlier)에 둔감한 반면 집단내 평균연결기법은 이상치에 민감하게 반응하기 때문이다. 문헌에서 일부 필드를 제외하면 각 문헌을 다른 문헌과 차별화하는 독자적인 분류 자질수도 줄어들게 되어 이상치가 상대적으로 덜 발생한다. 따라서 모든 필드를 포함하여 클러스터링하는 경우는 이상치가 더 자주 발생하므로 이에 민감한 집단내 평균연결기법은 상대적으로 불리할 수 있다.

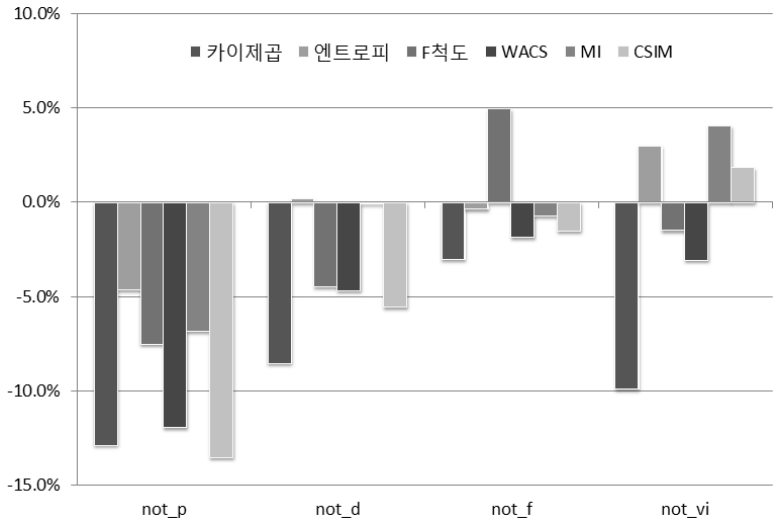
기법의 성향 없이 클러스터링의 성능에 영향을 미친 정도를 살펴보기 위하여 두 기법을 적용한 클러스터링 결과를 평가하여 성능이 변화한 정도를 평균을 내어 분석한 결과는 〈그림 2〉에서 제시되었다. 〈그림 2〉에서 보면 평균적으로 연구 목적을 제외시키고 클러스터링을 했을 때 가장 큰 성능 저하가 있는 것으로 나타났다.

연구 목적 다음으로 클러스터링에 영향을 미치고 있는 초록 필드는 연구 설계 필드로 연구 설계를 제외하고 클러스터링을 한 결과 〈표 6〉에서 나타났듯이 최대 19.2%(Ward 기법 적용 결과 CSIM 척도 기준) 성능 저하가 일어났다. 평균 변화율을 살펴보아도 엔트로피 척도와 상호정보량 척도를 제외한 대부분 척도에서 성능이 상당히 저하된 것으로 나타났다. 반면 연구 결과와 가치/의의 필드를 각각 제외하고 클러스터링 성능을 평가한 결과를 보면 일부 척도에서는 오히려 클러스터링 성능이 개선된 경우가

〈표 5〉 구조적 필드별 제외 문서 클러스터링 실험 평가 결과

기법	범위	카이제곱	엔트로피	F척도	정확률	재현율	WACS	MI	CSIM
WARD	ALL (초록 전체)	641.2	2.2104	0.3448	0.3389	0.3508	0.2952	1.5286	0.2083
	not_p (연구목적 제외)	506.4 (-21.0%)	2.4692 (-11.7%)	0.2919 (-15.3%)	0.2901 (-14.4%)	0.2937 (-16.3%)	0.2281 (-22.7%)	1.2699 (-16.9%)	0.1477 (-29.1%)
	not_d (연구설계 제외)	524.5 (-18.2%)	2.3143 (-4.7%)	0.2952 (-14.4%)	0.2906 (-14.3%)	0.3001 (-14.5%)	0.2548 (-13.7%)	1.4248 (-6.8%)	0.1683 (-19.2%)
	not_f (연구결과 제외)	605.9 (-5.5%)	2.3027 (-4.2%)	0.3668 (6.4%)	0.3270 (-3.5%)	0.4177 (19.0%)	0.2791 (-5.4%)	1.4363 (-6.0%)	0.1904 (-8.6%)
	not_vi (연구가치/의의 제외)	564.2 (-12.0%)	2.1831 (1.2%)	0.3302 (-4.2%)	0.3129 (-7.7%)	0.3496 (-0.4%)	0.2783 (-5.7%)	1.5560 (1.8%)	0.2053 (-1.4%)
집단내 평균연결	ALL (초록 전체)	564.1	2.1353	0.3010	0.2893	0.3138	0.2819	1.6038	0.2043
	not_p (연구목적 제외)	537.6 (-4.7%)	2.0828 (2.5%)	0.3020 (0.3%)	0.3040 (5.1%)	0.3000 (-4.4%)	0.2786 (-1.2%)	1.6563 (3.3%)	0.2084 (2.0%)
	not_d (연구설계 제외)	570.6 (1.1%)	2.0267 (5.1%)	0.3174 (5.4%)	0.3136 (8.4%)	0.3213 (2.4%)	0.2942 (4.4%)	1.7123 (6.8%)	0.2208 (8.1%)
	not_f (연구결과 제외)	561.1 (-0.5%)	2.0609 (3.5%)	0.3116 (3.5%)	0.3053 (5.5%)	0.3182 (1.4%)	0.2868 (1.8%)	1.6782 (4.6%)	0.2157 (5.6%)
	not_vi (연구가치/의의 제외)	520.0 (-7.8%)	2.0334 (4.8%)	0.3048 (1.3%)	0.3111 (7.5%)	0.2988 (-4.8%)	0.2808 (-0.4%)	1.7057 (6.4%)	0.2147 (5.1%)

* 괄호 안은 필드 제외를 하지 않은 경우와 비교한 성능 향상율, 밑줄 친 곳은 성능이 향상된 경우



〈그림 2〉 구조적 필드별 제외 문서 클러스터링 평균 향상율

발생하였다. 연구 결과 필드를 제외하고 Ward 기법을 적용한 결과에서는 재현율 척도 기준으로 19%가 클러스터링 성능이 개선된 것으로 나타났으며 집단내 평균연결 기법을 적용했을 때

〈표 6〉 두 기법 성능을 평균하였을 때의 향상율(밑줄친 곳은 성능이 향상된 경우)

	카이제곱	엔트로피	F척도	정확률	재현율	WACS	MI	CSIM
ALL	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
not_p	-12.9%	-4.6%	-7.5%	-4.7%	-10.3%	-11.9%	-6.8%	-13.5%
not_d	-8.5%	0.2%	-4.5%	-2.9%	-6.0%	-4.7%	0.0%	-5.6%
not_f	-3.0%	-0.3%	5.0%	1.0%	10.2%	-1.8%	-0.7%	-1.5%
not_vi	-9.9%	3.0%	-1.5%	-0.1%	-2.6%	-3.1%	4.1%	1.8%

는 카이 제곱을 제외한 모든 평가척도에서 클러스터링 성능이 개선된 것으로 나타났다. 가치/의의 필드를 제외하고 Ward 기법으로 클러스터링을 하였을 때는 엔트로피와 상호정보량 척도에서 클러스터링의 성능 개선이 일어난 것으로 분석되었고 집단내 평균연결 기법을 사용하였을 때 카이 제곱, 재현율, WACS 척도를 제외한 5개의 척도에서 클러스터링 성능이 향상된 것으로 측정되었다.

클러스터링 기법에 따른 구조적 초록 필드가 미치는 영향을 살펴보자면 Ward 기법에서는 일부 척도에서 연구 설계 필드와 가치/의의 필드를 제거하였을 때 클러스터링 성능 개선에 약간 긍정적 영향을 미친 것으로 나타났다. 그러나 개선된 것으로 평가한 지표가 재현율과 F척도(연구 설계 필드 제외), 그리고 엔트로피와 상호정보량(연구 결과 필드 제외)으로서 모두 클러스터 크기 차이가 심한 경우에 편향된 평가를 내리는 척도인 반면, 비편향적인 WACS와 CSIM 척도 기준에서는 모두 저하된 것으로 평가되었다. 연구 목적 필드는 WACS와 CSIM을 비롯한 대부분 척도에서 클러스터링 성능 저하에 영향을 크게 미치고 있는 것으로 파악되어 문서 클러스터링의 성능에 더욱 직접적인 영향을 미치는 것으로 나타났다.

집단내 평균연결 기법에서는 각 필드를 제외하였을 때 클러스터링의 성능이 대부분의 평가척도에서 향상되는 것으로 나타났다. 특히 연구 설계 필드는 문서 클러스터링에서 제외될 때 8개의 평가척도에서 모두 클러스터링이 개선된 것으로 나타나 집단 내 평균연결 기법을 적용하여 클러스터링을 할 때에는 연구 설계 필드는 문서의 주제를 인식하는데 저해요소로 작용하는 것으로 해석되었다. 일반적으로 연구 설계 부분에서는 주제 보다는 방법론에 치우친 표현이 포함되기 마련인데, 주제가 다른 문서도 유사한 연구 설계를 하는 경우도 있기 때문에 이 필드가 주제 클러스터링에서 문제를 발생시키는 것으로 추측된다.

5.2 필드 종속적 용어의 영향력 분석

구조적 초록의 세부 필드가 문서 클러스터링에 미치는 영향력을 평가하는 실험을 수행한 결과, 특정 필드는 클러스터링을 위해 문서의 주제를 파악하는데 독립적으로 기여를 크게 하지 못하고 있는 현상이 분석되었다. 특히, 문서의 주제를 표현하는데 크게 기여하지 못하는 필드에서 출현한 용어들 중에는 논문의 주제가 아닌 특정 필드의 기능을 표현할 가능성이 있는 용어들이 포함되어 있는 것으로 분석되었다. 예를 들어 연구 설계 필드는 집단내 평균연결 방식에

서 문서 클러스터링에서 제외되었을 때 클러스터링 성능이 개선되는 대표적인 필드로, 이러한 필드에는 주제와 크게 관련없이 특정 용어가 반복적으로 출현하는 현상이 나타났다. 이와 같은 현상을 구체적으로 검증하기 위하여 3.3절에서 설명한 방식으로 초록의 용어가 필드에 나타나는 빈도를 산출하여 필드 종속어를 선정한 결과 〈표 7〉과 같이 총 9개 용어가 추출되었다. 필드 종속어로 선정된 9개의 용어와 이 용어들이 종속되어 있는 필드를 조사한 결과 연구 설계 필드와 가치/의의 필드가 각각 4개씩으로 나와 필드 종속어가 출현하는 주요 필드인 것으로 분석되었다.

〈표 7〉 필드 종속어와 해당 필드

선정된 필드 종속어	1-2위 차*	종속된 초록 필드
survey	18	연구 설계
questionnaire	15	연구 설계
interest	11	가치/의의
collect	10	연구 설계
access	10	가치/의의
understand	10	가치/의의
public	9	연구 결과
similar	9	가치/의의
combination	9	연구 설계

* 1위 필드에서의 출현한 문서 수에서 2위 필드에서 출현한 문서 수를 뺀 값

필드 종속어로 선정된 용어들을 구체적으로 살펴보자면 필드 종속성이 가장 큰 것으로 나타난 용어는 survey와 questionnaire로서 연구 방법 유형에 속하는 대표적인 단어이다. 이 단어들은 연구 설계 필드에서 집중적으로 출현하고 있는데 실제 연구 주제의 특정성을 표현하지 못하는 용어들이다. 또한 interest나 collect와 같은 용어들도 주제를 표현하기 보다는 보편적인 방법이나 행위를 표현하는데 적합한 용어들이라서 실제 문서 주제를 파악하는데 크게 기여하지 못할 수 있다. 이와 같은 현상은 필드를 배제하여 클러스터링한 실험에서 나타난 결과와도 일관된 사항으로 연구 설계 필드는 클러스터링에 배제되었을 때 오히려 클러스터링 성능이 개선된 결과를 보여준 바가 있다. 따라서 이 연구에서는 문서 클러스터링에서 이러한 용어들을 불용어로 처리할 수 있도록 초록의 필드 속성을 활용하였고 그 결과는 〈표 8〉과 〈표 9〉와 같다.

필드 종속어를 불용어로 처리하여 문서 클러스터링에 적용한 결과 〈표 9〉에서 보이듯이 평균적으로는 다 개선된 것으로 나타났으나 두 기법 간에 차이가 큰 것으로 분석되었다. 〈표 8〉을 보면 Ward 기법에서는 엔트로피와 상호정보량 척도에서는 클러스터링의 성능에 긍정적 변화가 있는 것으로 나타났으나 나머지 6종

〈표 8〉 필드 종속적인 용어를 제외한 클러스터링 결과

기법	범위	카이제곱	엔트로피	F척도	정확률	재현율	WACS	MI	CSIM
WARD	ALL	641.2103	2.2104	0.3447	0.3388	0.3508	0.2951	1.5286	0.2082
	필드 종속어 배제	618.7615 (-3.50%)	2.1552 (2.50%)	0.3383 (-1.86%)	0.3381 (-0.22%)	0.3385 (-3.50%)	0.2836 (-3.89%)	1.5838 (3.61%)	0.1956 (-6.05%)
집단내 평균연결	ALL	564.1159	2.1352	0.3010	0.2892	0.3138	0.2818	1.6037	0.2042
	필드 종속어 배제	593.8784 (5.28%)	1.9880 (6.90%)	0.3279 (8.94%)	0.3276 (13.26%)	0.3282 (4.61%)	0.3060 (8.57%)	1.7510 (9.18%)	0.2356 (15.33%)

〈표 9〉 필드 종속적인 용어를 제외한 클러스터링 평균 향상율

	카이제곱	엔트로피	F척도	정확률	재현율	WACS	MI	CSIM
필드 종속어 배제	0.61%	4.66%	3.17%	5.99%	0.33%	2.19%	6.46%	4.54%

〈표 10〉 집단내 평균연결 기법 - 초록 필드 제외 vs. 필드 종속어 제외

기법	범위	카이제곱	엔트로피	F척도	정확률	재현율	WACS	MI	CSIM
집단내 평균 연결	ALL	564.1	2.1353	0.3010	0.2893	0.3138	0.2819	1.6038	0.2043
	not_p	537.6 (-4.7%)	2.0828 (2.5%)	0.3020 (0.3%)	0.3040 (5.1%)	0.3000 (-4.4%)	0.2786 (-1.2%)	1.6563 (3.3%)	0.2084 (2.0%)
	not_d	570.6 (1.1%)	2.0267 (5.1%)	0.3174 (5.4%)	0.3136 (8.4%)	0.3213 (2.4%)	0.2942 (4.4%)	1.7123 (6.8%)	0.2208 (8.1%)
	not_f	561.1 (-0.5%)	2.0609 (3.5%)	0.3116 (3.5%)	0.3053 (5.5%)	0.3182 (1.4%)	0.2868 (1.8%)	1.6782 (4.6%)	0.2157 (5.6%)
	not_vi	520.0 (-7.8%)	2.0334 (4.8%)	0.3048 (1.3%)	0.3111 (7.5%)	0.2988 (-4.8%)	0.2808 (-0.4%)	1.7057 (6.4%)	0.2147 (5.1%)
	필드 종속어 배제	593.8784 (5.28%)	1.9880 (6.90%)	0.3279 (8.94%)	0.3276 (13.26%)	0.3282 (4.61%)	0.3060 (8.57%)	1.7510 (9.18%)	0.2356 (15.33%)

척도에서는 모두 성능이 저하된 것으로 나타났다. 반면 집단내 평균연결 기법에서는 〈표 10〉에서 나타나듯이 필드 종속어를 문서 클러스터링에서 배제한 것이 주제성이 약한 필드인 연구 설계 필드나 가치/의의 필드 전체를 제외시킨 것보다 클러스터링 성능 개선에는 더 효과적인 것으로 분석되었다. 예를 들어 초록 필드를 제외시킨 실험에서 클러스터링 성능 향상율이 가장 높았던 경우는 연구 설계 필드를 제외한 것을 CSIM 척도로 평가한 것이었는데(8.1%), 필드 종속어를 제외하고 클러스터링 한 결과를 동일 척도인 CSIM 척도로 평가하였을 때와 비교하면 클러스터링 향상율이 15.33%로 크게 증가되었다.

6. 결 론

구조적 초록의 세부 필드는 클러스터링 기법에 따라 편차가 있지만 클러스터링의 성능에 영향을 미치고 있는 것으로 파악되었으며 이러한 세부 필드의 특성을 반영하여 클러스터링 성능을 저해하는 용어를 산출하여 클러스터링의 성능 개선에 반영할 수 있는 것으로 나타났다. 이 연구에서는 구조적 초록을 구성하는 세부 필드 간 상관관계를 분석한 후 각 필드를 선택적으로 배제하여 문서 클러스터링을 수행한 후 성능을 평가하였다. 필드 배제 클러스터링 과정에서 파악된 주제성이 약한 필드 종속어를 클러스터링 불용어로 처리하여 클러스터링 한 결과 클러스터링 성능이 개선된 것으로 파악되었다. 분석결

과 도출된 사항은 다음과 같다.

첫째, 구조적 초록을 구성하는 요소 중 연구 목적은 초록을 구성하는 다른 필드인 연구 설계, 연구 결과, 가치/의의와 모두 상호보완적인 관계가 있는 것으로 나타났고 연구 설계와 연구 결과는 부정적 상관관계가 있는 것으로 나타났다.

둘째, 구조적 초록의 각 필드를 하나씩 배제하여 초록 전체를 클러스터링 한 결과와 비교하였을 때 Ward 기법에서는 연구 목적을 배제하였을 때 가장 성능이 저하되는 것으로 나타났으며 연구 결과와 가치/의의 필드는 클러스터링 성능 저하에 크게 영향을 미치지 못하거나 배제하였을 경우 일부 척도에서 오히려 성능이 개선된 것으로 분석되었다. 각 필드별로 색인된 평균 용어 수를 비교하였을 때에는 가치/의의 필드나 연구 결과 필드가 연구 목적보다 양적으로 많았던 것을 고려하면 가치/의의나 연구 결과 필드는 정보량에 비해 주제성은 없는 단어를 포함하고 있는 성향이 있는 것으로 해석되었다. 또한 가장 정보량이 적었던 연구 목적을 제외하였을 때 클러스터링의 성능 저하가 가장 크게 발생하였기 때문에 구조적 필드에서는 정보량과 주제성이 비례하고 있지 않은 것으로 파악되었다. 연구 목적은 정보량과 반비례하여 주제성은 높은 정보 필드인 것으로 판명되었다.

셋째, 집단내 평균연결 기법을 적용하였을 때는 필드를 배제하였을 때 많은 척도에서 성능이 개선된 것으로 평가되었다. 특히 연구 설계 필드를 배제하였을 경우 모든 평가 척도에서 클러스터링의 성능이 개선된 것으로 나타났다. 이 실험에서 고찰된 것에 기반을 두면 Ward는 초록의 세부 필드 배제에 크게 영향을 받지 않는 것으로 나타났으며 집단내 평균연결 기법은 세부 필드 배제에 영향을 받는 것으로 분석되었다. 이와 같은 현상은 필드 종속어를 제거한 후 클러스터링 한 경우에도 일관적으로 나타났다.

넷째, 특정 필드에 종속적으로 출현하는 용어를 분석하여 문서의 주제 클러스터링에 노이즈가 되는 불용어로 처리하여 클러스터링 한 결과 집단내 평균연결 기법을 적용하였을 때에는 클러스터링 성능이 주목할 만큼 개선된 것으로 나타났다. 특히, 필드 종속어로 주제성이 없는 정보를 제거하는 방식은 주제성이 약한 필드 전체를 제외시키는 것보다 효과적인 것으로 나타났다. 이 연구에서 수행된 클러스터링 분석 결과, 필드 종속적인 단어는 학술논문의 주제를 대표하는 주제성이 낮아 클러스터링 성능 저하를 일으키는 요인이 될 수 있으므로 초록의 세부 필드를 활용하여 불용어를 파악하는 방안은 추후 좀 더 심층적으로 연구되어야 한다.

구조적 초록은 학술논문의 주제성을 표현하는 중요한 정보이다. 이 연구에서는 두 가지 클러스터링 기법을 적용하여 구조적 초록의 세부 필드가 문서 클러스터링 성능을 개선시키는데 활용될 수 있는 가능성을 확인하였다. 이 실험에서 도출된 사항을 일반화하려면 다양한 클러스터링 기법을 적용하는 것을 고려해야 한다. 구조적 초록의 세부 필드 주제성도 좀 더 다각적인 측면에서 측정할 것을 제언하며 필드 종속어의 특성을 반영하여 문서 클러스터링 외에 주제 색인 및 학술 논문 검색 성능 개선 등 다양한 정보처리 및 검색 분야에서 시도되어야 할 것이다.

참 고 문 헌

고영만, 송인석 (2011). 연구문헌의 지식구조를 반영하는 의미기반의 지식조직체계에 관한 연구. 정보관리학회지, 28(1), 145-170. doi: 10.3743/KOSIM.2011.28.1.145

윤보현, 오효정 (2011). 개체명 기반 웹 문서 클러스터링에서 자질 조합 분석. 한국정보기술학회논문지, 9(3), 199-206.

정영미, 이재윤 (2001). 클러스터링 성능 평가를 위한 비편향적 척도의 개발. 제8회 한국 정보관리학회 학술대회 논문집, 167-172.

조현양, 최성필 (2004). 계층적 결합형 문서 클러스터링 시스템과 복합명사 색인방법과의 연관관계 연구. 한국문헌정보학회지, 38(4), 179-192.

Chen, C., Frank, S. C., & Tseng, T. (2010). An integration of WordNet and fuzzy association rule mining for multi-label document clustering. Data & Knowledge Engineering, 69(11), 1208-1226. doi: 10.1016/j.datak.2010.08.003

Choi, Sang Hee (2010). Document clustering using reference titles. Journal of the Korean Society for Information Management, 27(2), 241-252. doi: 10.3743/KOSIM.2010.27.2.241

Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2009). Quality of abstracts in articles submitted to a scholarly journal: A mixed methods case study of the journal Research in the Schools. Library & Information Science Research, 32(1), 53-61. doi:10.1016/j.lisr.2009.08.004

Hartley, J. (1997). Is it appropriate to use structured abstracts in social science journals? Learned Publishing, 10(4), 313-317.

Hartley, J. (1998). Is it appropriate to use structured abstracts in non-medical science journals? Journal of Information Science, 24(5), 359-364.

Hartley, J. (1999). Applying ergonomics to Applied Ergonomics. Applied Ergonomics, 30(6), 535-541.

Hartley, J. (2000). Clarifying the abstracts of systematic reviews. Bulletin of the Medical Library Association, 88(4), 332-337.

Hartley, J. (2003). Improving the clarity of journal abstracts in psychology. Science Communication, 24(3), 366-379.

Nakayama, T, Hirai, N., Yamazaki, S., & Naito, M. (2005). Adoption of structured abstracts by general medical journals and format for a structured abstract. Journal of the Medical Library Association, 93(2), 237-242.

Sharma, S., & Harrison, J. E. (2006). Structured abstracts: Do they improve the quality of information in abstracts? American Journal of Orthodontics and Dentofacial Orthopedics, 130(4), 523-530.

Stevenson, H. A., & Harrison, J. E. (2009). Structured abstracts: Do they improve citation retrieval from dental journals? Journal of Orthodontics, 36(1), 52-60.

Zhu, S., Takigawa, I., & Mamitsuka, H. (2009). Field independent probabilistic model for clustering multi-field documents. Information Processing and Management, 45(5), 555-570. doi:10.1016./j.jpm.2009.03.005

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Hyun-Yang, & Choi, Sung-Pil (2004). The experimental study on the relationship between hierarchical agglomerative clustering and compound nouns indexing. The Journal of Korean Society for Library and Information Science, 38(4), 179-192.

Chung, Young Mee, & Lee, Jae Yun (2001). Development of an unbiased measure for clustering performance. Proceedings of the 7th Conference of Korean Society for Information Management, 167-172.

Ko, Young-Man, & Song, Inseok (2011). A study on the knowledge organizing system of research papers based on semantic relation of the knowledge structure. Journal of the Korean Society for Information Management, 28(1), 145-170. doi: 10.3743/KOSIM.2011.28.1.145

Yun, Bo-Hyun, & Oh, Hyo-Jung (2010). Analysis of feature combination in document clustering using named entities. The Journal of Korean Institute of Information Technology, 9(3), 199-206.