과학기술분야 학위논문 내용목차에 따른 주제어 출현빈도에 관한 연구

A Study on Frequency of Subjects on Contents of Thesis in Field of Science and Technology

이혜영(Hye-Young Lee)*
곽승진(Seung-Jin Kwak)**

초 록

일반적으로 문헌을 검색하고 접근하기 위하여 주제색인과 같은 주제어를 활용하곤 한다. 그렇다면 문헌의 내용과 문헌의 주제어는 분명히 어떤 밀접한 상관관계가 있을 것으로 예측해볼 수 있다. 본 연구는 이러한 의문점에서 출발하여, 디지털콘텐트의 본문내용이 비교적 짜임새 있게 정형화되어 있는 석사 학위논문을 연구문헌으로 한정하여 학위논문 전문에서 나타나는 학위논문의 주제어 분포도를 연구하였다. 학위논문의 주제어는 논문 저자가 부여한 주제어를 사용하되, 학위논문 전문은 '목차', '서론', '이론배경', '본론', '결론', '참고문헌'의 내용위치로 분할하여 내용위치에 따른 주제어의 출현율을 확인하였다. 연구대상 학위논문 전문은 1226.3개의 용어, 5152.3번의 용어 출현을 보였다. 학위논문 저자가 부여한 주제어는 12~13개 용어로 구성되어 있었다. 연구결과, 전문 내용위치에 따른 주제어의 출현율은 '목차' 11.4%와 '서론' 11.2%에서 가장 높았으며(11%), 다음 순위는 내용위치 '결론' 9.8%이었다.

ABSTRACT

We would generally use subject terms such as subject indexing for searching and accessing documents. So then, there must be any relationship between document's full-text and its subject terms. This study is started in this question. Master's theses in field of science and technology are worked with because full-text is relatively formatted. This study is to study locations of subject term on Thesis, distribution patterns of subject terms on content of full-text; 'Contents', 'Introduction', 'Theory', 'Main subject', 'Conclusion' and 'References'. Thesis were averagely composed of 1226.3 terms. And Subject terms were averagely compose of 12~13 terms. As a result, 'Contents' and 'Introduction' have had the most frequency of subject.

키워드: 디지털콘텐트, 내용목차, 주제어, 주제어 분포도, 전문 내용위치
digital content, table of contents, subject term, frequency of subject, division of full-text

---

* 한국과학기술원 학술정보처 학술정보개발팀(hye@kaist.ac.kr) (제1저자)
** 충남대학교 사회과학대학 문헌정보학과 조교수(sjkwak@cnu.ac.kr) (공동저자)

■ 논문접수일자 : 2008년 2월 18일 ■ 게재확정일자 : 2008년 3월 14일
■ 情報管理學會誌, 25(1): 191-210, 2008. [DOI:10.3743/KOSIM.2008.25.1.191]

1. 서 론

1.1 연구의 필요성

지식정보는 인류의 역사적 산물을 이해하게 하고 또한 더 나은 인류발전을 위한 영향력 있는 기본적인 요소라고 할 수 있다. 오랜 역사 속에서 축적된 수많은 지식정보는 문헌학적 연구를 통하여 고대 자료부터 디지털 자료에 이르기까지 지식체계, 정보 분류, 서지학, 정보서비스 등의 연구 및 서비스분야를 통하여 인류발전에 기여해 왔다. 디지털정보가 급속히 증가하는 현 디지털사회에서 이들 연구 분야에서는 뉴스, 신문기사와 같은 단편적인 문서에 한정된 연구를 탈피하여 학위논문, 연구보고서와 같은 내용분량이 많은 문헌에까지 그 연구 영역을 확대할 필요가 있다.

문헌은 일반적으로 내용목차(table of contents)를 가지고 있다. 문헌의 저자는 내용목차의 순서에 따라 이야기하고자 하는 내용을 기술하기 때문에, 문헌의 본문 내용을 모두 읽지 않더라도 내용목차만으로도 간략하게나마 손쉽게 문헌의 내용을 추측할 수 있다. 우리의 두뇌는 문헌의 내용목차만으로 문헌의 내용구조를 형상화 할 수 있다. 이것은 문헌의 내용목차에 따른 문헌의 내용분석에 있어서 어떠한 의미 있는 역할을 기대해 볼 수 있게 한다. 이것은 문헌의 내용목차에 따른 사용되는 용어, 용어의 출현횟수, 문장 개수 그리고 문헌의 내용목차에 따라 표현하고자 하는 내용 측면에서 어떠한 차이점을 기대해 볼 수 있게 한다. 따라서 문헌의 내용을 분석할 때, 문헌의 내용목차에 따른 용어, 용어 출현개수 등의 특징을 활용한다면 그렇지 않았을 때와 비교하였을 때 보다 효과적인 결과를 얻을 수 있을 것으로 기대된다.

뉴스 기사, 이-메일과 같은 단편적인 문서의 주제어 자동추출 연구를 학위논문과 같은 디지털콘텐트까지 연구영역을 확장하기 위해서는 무엇보다 디지털콘텐트 전문에 대한 데이터 분석이 필요하다 하겠다. 본 연구는 디지털콘텐트의 내용목차에 따라, 각 내용목차의 내용의 미에 따라 전문에서 사용되고 있는 용어를 분석 하였다. 용어 분석은 디지털콘텐트 저자가 부여한 주제어가 다수 발견되는 내용목차 항목을 선별하는 것이다. 분석결과는 디지털콘텐트의 주요 내용 중심위치가 내용목차에 따라 변화하고 있음을 보여 줄 것이다. 본 연구결과를 디지털콘텐트의 주제어 자동추출 또는 문서분류에 응용한다면 더 효과적인 연구결과를 기대해 볼 수도 있을 것이다.

1.2 연구의 목적 및 범위

본 연구의 목적은 학위논문 전문(full-text)에서 학위논문의 주제어가 학위논문의 내용목차에 따라 어떻게 분포하고 있는 가를 연구하는 것이다. 연구 대상 자료는 디지털콘텐트이되, 전문의 내용목차 구성이 비교적 정형화되고 일반화된 학위논문으로 선정하였다. 문헌의 내용목차에 따른 주제어 출현빈도 및 분포도를 조사하기 위하여, 연구대상 자료는 각각의 디지털콘텐트 내용목차 항목구성이 유사할 필요가 있다. 개별적인 디지털콘텐트의 내용목차 항목을 가능한 수용하면서, 연구결과를 일반화시키기 위해서는 모든 디지털콘텐트의 내용목차 항목구성이 유사해야 한다. 이럴 때, 내용목

차에 따른 모든 디지털콘텐트의 본문내용 분할이 가능할 것이다.

본 연구는 K대학교의 석사 학위논문 디지털콘텐트를 연구 대상 자료로 하였다. 과학기술인재양성이라는 목표를 가지고 있는 K대학교는 이공계 중심의 학과로 구성되어 있어 학문의 주제 분야가 제한적일 뿐 아니라, 특히 석사 학위논문의 경우 일반적으로 실험과 비교평가를 통해 연구가 이루어지기 때문에 학위논문의 구성내용이 비교적 유사할 것으로 예상된다. 또한 K 대학교는 학위논문 작성규정을 통해 학위논문의 작성 및 내용 전개 순서를 제시하고 규정을 지킬 것을 권고하고 있다. 이러한 과학기술분야 학위논문은 타 콘텐트에 비해 텍스트 내용이 상당히 길고, 거의 유사한 내용 짜임새를 가진 내용목차를 따르는 디지털콘텐트이기 때문에 연구범위를 과학기술분야 학위논문으로 한정하였다.

학위논문은 일반적으로 표지정보(서명, 저자, 지도교수 등), 목차, 초록, 연구목적, 연구필요성, 본론, 배경 이론, 실험, 실험결과, 결론, 이력, 감사의 글 순으로 기술되는 경향이 있다. 이러한 순서에 따른 내용전개 방식은 과학기술분야 학위논문의 특징이기도 하다. 이렇게 학위논문은 본문 내용이 비교적 짜임새 있게 정형화되어 있다는 특징을 가지고 있기 때문에 전문의 내용목차에 따라 본문내용을 분할하기에 용이하다는 장점을 지닌다. 이러한 이유로 K대학교의 석사 학위논문을 선택하여 학위논문 내에서의 학위논문의 주제어 출현율 또는 분포도를 연구하였다.

특히 K 대학교의 모든 학위논문은 모두 메타데이터를 구축하고 있는데 메타데이터에서는 학위논문의 저자가 직접 선정한 국문과 영문의 학위논문 주제어를 포함하고 있다. 학위논문의 주제어는 학위논문의 중심 내용을 전달하는 주 용어로써 본 연구에서는 이 점에 중점을 두어 학위논문의 본문 내용에서의 주제어 분포도를 밝혀 디지털콘텐트의 본문 내용에서의 주요 내용의 위치를 연구하고자 한다.

우리는 때때로 학위논문에 할당되어 있는 주제어만으로 학위논문의 내용을 대략적인 이해가 가능하다. 따라서 연구를 통하여 학위논문의 주제어가 빈번하게 사용되고 있는 학위논문의 내용위치를 밝혀낸다면, 학위논문의 일부 본문내용만으로 문헌의 중심내용을 파악할 수 있는 근거를 마련할 수 있을 것이다. 학위논문의 내용을 함축하는 대표어인 학위논문의 주제어는 학위논문 작성자가 직접 선정하여 입력한 용어이다. 학위논문 작성자가 선정하고 부여한 주제어는 학위논문의 주제어를 선정하는 어떤 방법보다 가장 신뢰도가 높다고 할 수 있을 것이다.

본 연구는 학위논문 전문을 내용목차에 따라 내용위치를 분할하고, 분할한 내용위치에서 사용되는 용어를 분석하는 과정으로 진행하였다. 특히, 각 학위논문의 내용위치에서 사용되는 용어가 학위논문의 주제어인가를 확인하여, 학위논문 내용위치에 따라 학위논문의 주제어 사용빈도를 분석하였다. 본 연구는 학위논문의 본문내용과 학위논문의 주제어 간의 상관관계 또는 내용위치에 따른 주제어의 분포도를 밝히는 것이다.

2. 관련 연구

2.1 주제색인

용어(terminology)는 '일정한 전문 분야에서 주로 사용되는 말'이고, 어휘(vocabulary)는 '일정한 범위 안에서 쓰이는 단어의 총체'로 정의된다. 그런데 언어를 구성하는 중요한 의미단위는 어휘소이고, 실제 어휘집들이 어휘를 분류하기 위하여 채택한 기준은 모두가 개별 어휘소들의 의미이다(김광해 1993). 그러므로 용어를 분류한다는 것은 용어 또는 어휘의 내재적인 의미에 따라 분류하는 것을 말한다고 할 수 있다.

특별히 주제색인(subject indexing)은 저자명색인, 표제색인 등의 비주제색인과는 달리 문헌의 내용을 나타내기 위한 것으로 색인대상 문헌에 대한 정확한 내용 파악이 필요하다. 문헌의 주제 분석을 통해 추출된 특정 개념을 표현하는 용어 또는 색인어를 선정한다. 주제색인은 색인어의 선정뿐 아니라 선정된 색인어의 다양한 어의적 관계, 단일어로서 표현할 수 없는 복합주제에 대한 다단어 색인어의 논리적 조합, 그리고 배열에 관한 구문적 관계 등의 복잡한 요소를 내포하고 있다. 주제색인은 우선 색인어의 통제 유무에 따라 통제언어(controlled vocabulary) 색인과 자연언어(natural language) 색인으로 구분된다. 통제언어 색인은 색인어 선정에서 시소러스 등의 언어정보를 이용하여 본문 내에 표현된 언어를 기준으로 정한 용어로 바꾸어 색인하는 하는 것을 말한다. 자연언어 색인은 본문에 등장한 용어 그 자체를 색인어로 선정하는 것이다. 그러나 자연어색인은 여러 개의 다른 용어로 같은 개념을 표현할 수 있는 용어의 의미에 있어서의 특성을 배제하고 있기 때문에 문헌 검색의 재현율 저하를 가져오고, 동형이의어의 용어의 특성을 배제하게 되어 문헌 검색의 정확율 저하를 가져오게 된다.

국립중앙도서관(2003)은 데이터의 종류, 양, 분야, 이용자의 요구 등에 따라 다양한 요구 및 주제명표목표에 포함되어야 할 정보와 주제명표목표가 가져야 할 기능에 대한 새로운 요구에 대응할 수 있는 주제명표목표를 개발하였다. 주제명표목표는 상황분류, 복수개의 상위개념, 세목의 배제, 다양한 분류방법의 수용, 느슨한 범주화, 주제명표목표의 용도의 일반화, 참조정보의 수용, 개념구조와 특정성 조절이라는 7가지 특징을 가지고 있다. 주제명표목표의 형식은 시소러스를 따랐으며, 주제명표목표의 기술구축과 시소러스형 주제명표목표의 특징을 가능한 상세하게 기술하고 있다.

정보검색시스템에서 색인용어를 통제하여 효율적인 색인작업과 검색의 효율성을 높이기 위해 사용되는 용어통제는 크게 주제명표목표와 시소러스(thesaurus)로 구분 할 수 있다. 주제명표목표는 사전체 목록시스템을 채택하고 있는 도서관에서 주제명목록 작성을 위해 전통적으로 사용하고 있는 통제어휘집으로 MESH (Medical Subject Headings)와 LCSH(Library of Congress Subject Headings)가 대표적이다. 시소러스는 색인작업에서 적절한 색인어의 선택과 색인어의 통제를 위해 필요한 것으로, 검색에서 적절한 탐색어의 선택을 위해서도 필요하다. ANSI/NISO Z39.19에서는 시소러스의 기능에 대하여 네 가지로 나누고 있다. 즉 첫째, 저자, 색인자, 이용자가 사용한 자연언어

를 색인작성과 검색에 사용할 통제어휘로 번역하는 수단을 제공한다. 둘째, 색인어 부여의 일관성을 보증한다. 셋째 용어간의 의미관계를 지시한다. 넷째, 문헌 탐색에서 탐색보조도구가 된다.

2.2 문서의 용어선택 자동화

문서의 특징을 자동 추출하는 것은 일반적으로 문서의 자동 범주화에서 매우 중요하다. 전산학의 문서 분류에 관한 연구에서 이러한 처리단계를 '용어선택(Feature Selection)을 한다'라고 부른다(Yang and Pedersen 1997). 용어선택(Feature Selection)의 중요성이 부각되는 또 다른 이유는 문서 분류에서 관계없는 용어, 불필요한 용어들로 인하여 문서 분류 속도 또는 분류 정확도를 자주 저해할 수 있기 때문이다. 따라서 용어선택을 위한 자동화 방법은 말뭉치의 통계와 보다 높은 차원의 용어에 보다 낮은 차원의 용어(feature)를 수직적으로 결합한 용어의 구조에 따라 의미 없는 용어를 제거하는 것을 포함한다.

Novovicova(2004)에 따르면, 용어선택 방법은 크게, 가장 개별적인 용어 선택방법(BIF: best individual features)과 선행적 선택방법(SFS: sequential forward selection)으로 구분된다. BIF 방법은 용어선택에 있어서 주어진 평가기준에 따라 모든 단어를 개별적으로 평가하여 정렬한 후, 우선순위 k개 단어를 선택하는 방식으로 보다 빠르고 효율적이며 간단하기 때문에 가장 빈번하게 사용된다. 문서 빈도량(Document Frequency), 용어 빈도량(Term Frequency), 정보 획득량(Information Gain), 기대상호정보량(Mutual Information), 카이제곱량(X^2-test)이 이에 해당한다. SFS 방법에 따른 용어선택은 주어진 기준에 따라 평가하여 가장 적합한 단어 하나씩 선정하여 선정한 단어의 개수를 충족시킬 때까지 단어를 선정하는 방법이다. BIF에 비해 효율적이지는 않지만 단어 간의 의존성을 가지게 된다. 이러한 상호정보량(MI)에 따른 SFS 방법은 Battiti(1994)와 Kwak(1999)의 연구논문을 통해 확인 할 수 있다.

Yang과 Pedersen(1997)은 문서 분류를 위한 용어선택방법으로서 문서 빈도량, 정보 획득량, 기대상호정보량, 카이제곱량 등의 성능을 비교 연구하였는데, 일반적인 통계학적 문서분류에 있어서 정보획득량과 카이제곱량이 가장 효율적인 것으로 나타났다. 또한, 저널 논문을 대상으로 문서 범주화에 관한 Moens(2000)의 연구에서도 세 가지 분류기법을 비교 연구한 결과에서 베이지언 분류, Rocchino 알고리즘 보다 카이제곱량이 가장 효과적인 것으로 나타났다.

2.3 저자가 부여한 학술저널논문의 주제어에 관한 연구

Gil-Leiva와 Alonso-Arroyo(2007)는 디스크립터 내에서의 키워드의 존재유무를 확인하기 위하여 과학분야 논문의 저자와 논문에 할당된 디스크립터를 분석하였다. 디스크립터(descriptor)는 컴퓨터분야에서 정보의 분류 및 색인에 쓰는 어구로서 기술어(記述語)이다. 640개의 INSPEC(Information Service for Physics, Engineering, and Computing), CAB

(Current Agriculture Bibliography) 초록, ISTA(Information Science and Technology Abstracts), LISA(Library and Information Science Abstracts) 데이터베이스를 연구대상으로 하였다. 그의 자세한 비교결과, 논문 저자가 부여한 키워드는 연구대상인 데이터베이스에서의 매우 의미 있는 출현을 보이고 있음을 알게 되었다. 모든 키워드의 약 25%가 디스크립터에서 동일하게 나타나고 있었고, 또 다른 21%에서는 대체로 유사하게 나타나고 있는 것으로 나타났다. 이것은 키워드의 약 46%가 디스크립터에 나타난다는 것이다. 이들 연구는 과학분야 논문에서 논문 저자가 부여한 키워드가 지능적인 용어 색인기를 통하여 생성된 디스크립터에 얼마나 직접적으로 정확하게 또는 최소화한 일반화 과정을 통하여 이루어진 간접적으로 유사하게 등장하는 가에 대하여 연구한 것이다. 이들은 연구를 위해 2005년도의 학술저널과는 달리 보다 국제적으로 사용되고 있는 학술데이터베이스를 대상으로 하고 있다.

Gil-Leiva와 Alonso-Arroyo(2007)은 몇 가지 측면에서 의미를 찾을 수 있다. 과학분야 논문에서 저자가 부여한 키워드는 논문 텍스트의 즉각적인 인덱스 작업에서 일정 부분에서 역할을 할 수 도 또는 역할을 하지 못할 수도 있다는 것인데, 이것은 이미 여러 곳에서 효율성이 확인되고 있다. 다음으로는 연구된 실험대상이 매우 크지는 않지만 3 종류의 인덱스 작업정책인 발견된다는 것이 또 다른 의미이다. 즉 INSPEC, LISA과 같이 사람의 판단에 따라 디스크립터의 할당이 변화될 수 있다는 것이고, 때로는 너무 많은 디스크립터를 제한하기 위하여 디스크립터 개수를 제한하기도 한다는 것이다. CAB 이 대표적인 예이다. 그리고 마지막으로 인덱스 작업의 경제적 효율성을 위하여 오직 4개의 디스크립터만 보유하게 한 것이다. 가장 중요한 연구 의미는 저자에 의해 부여된 키워드는 데이터베이스 디스크립터 안에서 매우 중요한 출현을 보인다는 것이다.

3. 연구 방법

3.1 연구 설계 및 절차

본 연구의 연구대상은 디지털콘텐트이다. 그동안 주요 연구대상이었던 디지털콘텐트의 메타데이터, 혹은 초록데이터를 뛰어넘어 디지털콘텐트 자체 전문을 연구대상으로 한다. 연구대상 자료를 특정 분야의 문헌, 과학기술분야 학위논문으로 한정하였는데, 다양한 정보그룹의 혼재로 인하여 발생할 수 있는 디지털콘텐트의 본문내용 전개순서의 다양성을 배제하기 위함이다. 연구대상 정보그룹이 혼재할 경우, 디지털콘텐트 본문내용의 내용위치 분할 기준을 일률적인 적용이 불가능하고 디지털콘텐트 전문의 분할이 모호해질 수 있기 때문이다.

학위논문 전문에서의 주제어 분포도에 관한, 본 연구는 크게 자료 수집, 데이터 추출, 데이터 분석, 결과 도출의 일련의 과정을 걸쳐 이루어졌다. 첫째, 연구대상 자료수집 작업으로서 학위논문 디지털콘텐트와 논문저자가 부여한 주제어를 수집하였다. 둘째, 학위논문 전문의 색인용어를 추출하였다. 색인용어 자동추출기는 KLT(국민대학교 2007)을 활용하였다. 불용어 등 키워드 추출에 필요한 모든 제반환경은

KLT 시스템을 따랐다. 색인용어 자동추출기는 학위논문 본문내용을 내용위치에 따라 분할한 텍스트의 색인용어 추출에도 동일하게 적용하였다. 셋째, 데이터분석 작업으로 학위논문 전문에 대한 색인용어 분석 및 학위논문의 주제어를 분석하고, 학위논문의 색인용어와 학위논문의 주제어를 매핑하여 학위논문 전문에서의 주제어 출현빈도 및 분포도를 비교분석하였다. 이러한 학위논문 전문에서의 주제어 위치정보를 토대로 통계학적 수치분석을 활용하여 학위논문 내의 주제어 출현율을 계산하여 분석결과를 도출하였다.

학위논문의 내용위치로써 초록데이터 또한 내용위치로 포함하여 다른 학위논문의 내용위치와 함께 주제어 분포도를 비교 분석하였다. 학위논문의 초록데이터가 일반적으로 학위논문의 요약정보로 사용되는 것에 관하여 학위논문의 초록데이터에서 학위논문의 본문내용보다 더 많은 학위논문의 주제어가 발견될 가능성이 있기 때문이다. 과연 초록데이터에서 디지털콘텐트 전문 전체 또는 전문의 주요 내용위치보다 더 높은 주제어의 출현율을 보이는지를 확인하고자 하였다.

3.2 자료 수집

디지털콘텐트 내용목차에 따른 전문에서의 주제어 출현율에 관한, 본 연구는 과학기술분야 석사학위논문으로 연구대상 자료를 한정하였다. 연구대상 자료는 대체로 유사한 내용 전개, 내용목차를 채용하는 문헌이어야 한다. 디지털콘텐트 전문의 내용위치를 선정하고 전문을 분할한 후 문헌 내의 주제어 위치를 찾아내어 주제어가 다수 발견되는 일반적인 문헌의 내용위치를 밝혀야 하기 때문이다. 만일, 전문의 내용 전개가 디지털콘텐트마다 가변적이라면 전문의 내용위치를 동일하게 선정하기 어려울 수 있을 것이다. 혹여 발생할 수 있는 문제점을 사전에 방지하기 위하여 실험 위주의 연구방법을 취하는 과학기술분야 석사 학위논문으로 한정하여 연구하였다. 과학기술분야의 석사 학위논문은 내용 기술에 있어서 비교, 분석에 따른 사실적인 기술에 중점을 두기 때문에 용어 사용에 의한 문제점 즉 은유, 비유, 함축적인 용어 사용에 따른 연구결과 분석의 모호성을 상당히 감소시킬 수 있을 것으로 예측된다.

이러한 이유로 객관적인 사실 기술에 중점을 두어, 실험, 증명, 검증을 통해 논제를 서술해가는 과학기술분야 인재를 배출하는 국내 K대학교의 학위논문을 사용하였다. K대학교는 이공계 중심의 학과로 구성되어 있어 학문의 주제분야가 제한적일 뿐 아니라, 특히 석사학위논문의 경우 일반적으로 실험과 비교평가를 통해 연구가 이루어지기 때문이다. K대학교는 국가산업진흥과 국가경쟁력 배양을 위한 과학기술인재 양성을 목표로 하여, 과학기술과 경영의 통합적인 지식과 사고에 근간한 문제해결능력을 갖춘 경영인재 양성에도 앞장서고 있다. 경영대학도 과학기술에 근간한 실험 위주의 학위논문을 많이 배출하고 있다. 연구대상 자료수집에 있어서, K대학교의 단과대학 중 최근 신설된 문화과학대학을 제외한 자연과학대학, 공과대학, 경영대학에서 배출된 석사 학위논문을 수집하였다. 또한 K대학교는 학위논문 작성규정을 통해 학위논문의 작성 및 내용 전개 순서를 제시하고 규정을 지킬 것을 권고하고 있다.

따라서 학위논문 전문의 내용전개가 거의 유사하다 할 수 있어, 전문 내용목차에 따른 데이터 분석에 적합하다.

본 연구는 학위논문의 학문 분야에 따른 전문에서의 주제어 쓰임과 출현횟수, 전문 내용위치에 따른 주제어 분포도도 확인하였다. 학문분야에 따라 주된 관심 주제가 상이한 것은 당연하기 때문이다. 학문분야 즉 단과대학을 중심으로 주제어를 분석하였을 때, 주로 사용되는 주제어가 상이할 것이다. 데이터 분석대상인 학위논문 전문에서 사용되는 주제어의 위치정보를 가지고 학문분야에 따른 디지털콘텐트 전문의 내용적 분석을 접근 하고자 하였다. 연구대상 자료는 최근 3년간 K대학교에서 수여한 석사학위논문을 단과대학과 학위년도에 따라 등분하여 선정한 학위논문 386건이다.

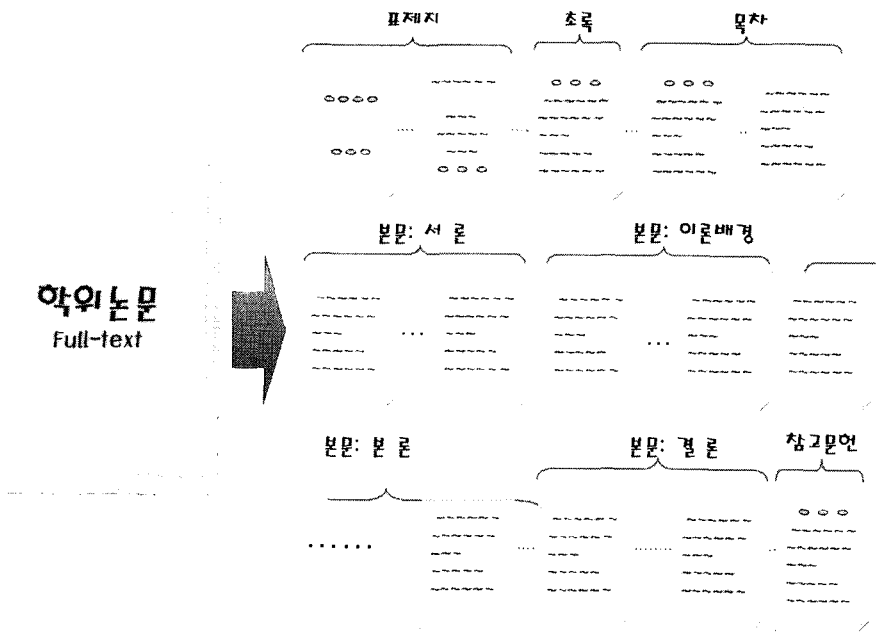
3.3 학위논문 내용목차 및 내용위치 선정

연구대상 자료인 K대학교의 학위논문의 내용은 겉표지, 속표지, 학위논문 제출승인서, 학위논문 심사완료 검인, 논문 초록(영문), 목차, 본문, 국문요약(본문이 외국어인 경우), 참고문헌, 감사의 글, 이력서, 감사의 글 순으로 작성 할 것을 지시하고 있다. 학위논문의 본문내용은 대부분 크게 서론, 본론, 결론으로 구분하고 있으며, 통상적으로 서론은 연구목적, 연구목표, 연구배경 등을 기술하고, 본론은 이론, 선행연구, 실험방법, 실험재료, 결과, 고찰 등을 기술하고, 마지막에 결론을 기술하여 논문을 완성하고 있다.

〈그림 1〉과 같이 학위논문의 제본 바인드를 제거하여 낱장의 페이지를 펼쳐 배열하면 학위논문의 페이지 수 만큼의 학위논문의 텍스트를 얻을 수 있다. 학위논문의 본문내용으로서 사용하기 어려운 '표제지', '감사의 글'에 해당하는 페이지를 제외한 학위논문의 내용목차 항목을 기준으로 전문의 내용위치를 설정하였다. 사실상 학위논문의 각 내용목차는 모두 상이하다. 학위논문 작성요령 및 순서에 관한 규정이 있다고 하나, 학위논문 저자의 의도에 따라 내용목차 항목을 자의적으로 선정하고 학위논문을 작성하기 때문에 학위논문의 내용목차 항목이 학위논문마다 목차항목 개수가 상이하다.

〈표 1〉 연구대상자료

학위년도	단과대학	학위논문 수	소 계
2005	자연과학대학	36	144
	공과대학	29	
	경영대학	79	
2006	자연과학대학	30	128
	공과대학	33	
	경영대학	65	
2007	자연과학대학	24	114
	공과대학	20	
	경영대학	70	
합 계			386



〈그림 1〉 학위논문의 내용목차별 텍스트

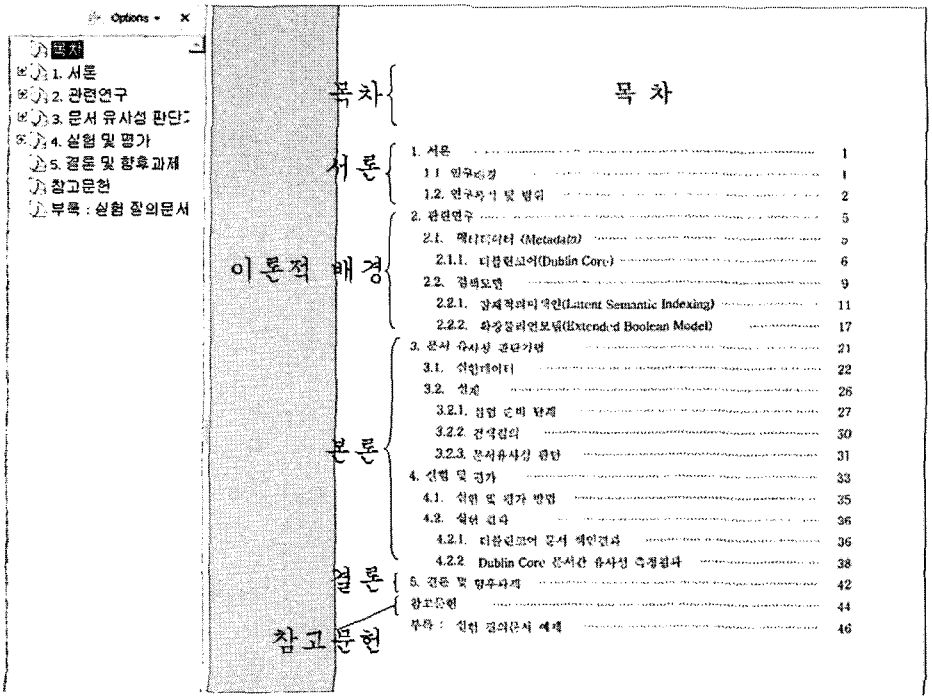
또한 내용목차 항목의 제목도 매우 다양하다. 그러나 과학기술분야 학위논문은 학위논문 내의 실제 연구내용과 연구방법이 거의 유사하기 때문에 다음의 내용목차 항목으로 그룹화 할 수 있다.

학위논문은 내용목차 항목 순서에 따라 "서론", "이론배경", "본론", "결과" 순의 4개 내용위치이다. "서론", "이론배경", "본론", "결과" 순의 4개 내용위치에 따라 학위논문의 내용목차 항목의 제목을 참조하여 학위논문의 전문을 분할하였다. 본 연구는 위 4개 항목의 전문의 텍스트 내용위치에 더하여, 학위논문의 "목차", 학위논문의 "참고문헌", 학위논문의 "초록" 데이터 또한 전문 내용위치로 간주하여 주제어의 출현율과 주제어 분포도를 함께 비교 분석하였다.

4. 연구결과 분석

4.1 학위논문의 주제어 분석결과

학위논문의 주제어에 있어서 단과대학에 따른, 즉 학문분야에 따른 디지털콘텐트의 평균 주제어 개수와 주제어 사용빈도는 별다른 차이점이 발견되지 않았다. 〈표 2〉와 같이, 모든 학문분야에 걸쳐 학위논문의 주제어는 평균 12~13개가 할당되어 있었다. K대학교는 학위논문 저자가 주제어 부여할 때, 규정상 주제어 개수의 제한은 없으나 국문 주제어와 영문 주제어를 각기 입력하도록 되어 있다. 이것은 학술저널 논문작성법에서도 밝히고 있는 주제어 입력규정 "주제어 표기언어별 5~6개" 입력과 매우 흡사한 결과이다.



〈그림 2〉 학위논문의 내용위치

〈표 2〉 주제어 개수

학문분야	학위논문 개수	주제어 개수	학위논문 당 평균 주제어 개수
자연과학	90	992	11.02
공학	82	1,154	14.07
경영학	214	2,827	13.21
(전체)	386	4,973	12.88

〈표 3〉은 학위논문 주제어로써 가장 많이 사용된 용어를 분석한 것이다. 물리, 화학, 생물 등 순수과학을 연구하는 자연과학분야 학위논문과 건설, 기계, 항공우주 등의 공학분야 학위논문, 그리고 경영공학, 정보기술, 경영미디어 등의 경영학분야 학위논문은 각 자연과학, 공학, 경영학의 학문분야에 따라 자주 사용되는 주제어에 다소 차이가 있었다.

학문분야별 최다 사용되는 주제어는 학위년도와는 무관하였고, 일반적으로 사용되는 보통명사라는 특징이 있었다. 자연과학에서는 '단백질', '세포', '분자' 와 같은 보통명사가 자주 사용되었고, 공학에서는 '네트워크', '항공기', '센서', '제어'와 같은 보통명사가, 경영학에서는 '신용', '스캐줄', '모형', '위험'과 같은 보통명사가 자주 사용되고 있었다.

〈표 3〉 학문분야별 최다 사용된 주제어

학위년도		자연과학	공학	경영학
2005년도	1위	전달	제어	모형
	2위	mmp-9	네트워크	model
	3위	organic	현실	변동
	4위	단백질	control	변동성
	5위	반응	reality	volatility
2006년도	1위	단백질	제어	신용
	2위	세포	항공기	credit
	3위	nmr	owl	scheduling
	4위	분광법	무인항공기	스케줄링
	5위	분자	복합	위험
2007년도	1위	반응	센서	model
	2위	변동	인식	신용
	3위	변동성	sensor	credit
	4위	Dibromopropane	네트워크	design
	5위	ALKB	상황	모형

4.2 학위논문 전문의 색인용어 분석결과

〈표 4〉는 학위논문 전문의 색인용어 개수와 초록데이터의 색인용어 개수를 나타내고 있다. 전문의 경우 평균 5152.3개의 단어로 구성되어 있으며, 중복 사용되는 단어 개수를 제외하면 실제 사용된 용어 개수는 1226.3개 이었다. 학위논문 전문에서 개별 용어 평균 4.2회 반복사용하고 있는 것이다. 학위논문 전문은 경우에 따라 영어 또는 한글, 한 가지 언어로 작성되어 있었다. 초록데이터는 국문초록과 영문초록으로 표기언어에 따라 구분되어 있었다. 따라서 학위논문 내용위치로서 국문초록, 영문초록, 국문과 영문초록을 합한 전체 초록의 3개 내용위치를 설정하였다. 각 내용위치의 초록데이터는 국문초록의 경우 65.5개 용어를 104.6회, 즉 용어 평균 1.6회 사용하고 있었다. 영문초록은 55.4개 용어를 75.9회, 용어 평균 1.4회 사용하고 있었다. 국문초록과 영문초록을 합한 전체 초록은 112.1개 용어를 170.9회 사용하고 있었으며, 이것은 용어 평균 1.5회 사용하는 것이다.

학위논문 전문과 초록데이터에서 사용되는

〈표 4〉 학위논문 전문 및 초록데이터의 자동색인 결과

구 분	학위논문 전문	초 록		
		전체(국문+영문)	국 문	영 문
용어 개수	1226.3	112.1	65.5	55.4
	100%	9.14%	5.34%	4.52%
용어 출현개수	5152.3	170.9	104.6	75.9
	100%	3.32%	2.03%	1.47%

용어 비교에서 개별 용어의 반반복사용 횟수의 평균이 각각 전문 4.2회, 초록 1.5회로, 전문에서 개별 용어의 반복사용이 많은 것으로 나타났다. 텍스트의 내용분량 측면에서 초록데이터는 전문의 내용분량의 3.32%, 즉 전문에서 사용되는 용어 출현개수의 3.32%에 해당하는 단어들로 국영문 초록데이터를 구성하고 있는 것이다. 이것은 예를 들어 1000개 단어로 이루어진 학위논문 전문을 332개 단어로 학위논문 내용을 요약하고 있는 셈이다. 그리고, 전문과 초록에서 사용한 개별 용어는 전문에서 사용한 개별 용어의 9.14%가 국영문 전체 초록데이터에서 사용하고 있었다. 이것으로 학위논문 전문이 초록에 비하여 용어의 반복사용이 많음을 알 수 있다.

〈표 5〉는 학위논문 전문의 키워드 자동색인과 전문의 내용위치에 따라 분할 데이터를 자동색인한 결과이다. 색인용어 자동추출은 KLT(국민대학교 2007)을 활용하였다. 앞서 짐작할 수 있듯이, 가장 많은 용어와 가장 많은 용어 출현을 보인 전문의 내용위치는 '본론' 이었다. 내용위치 '본론'에서 사용된 용어의 반복 사용 횟수는 평균 3.9회로서 다른 내용위치의 용어 반복 사용횟수에 비해 다소 높은 것으로 나타났다. 용어의 반복 사용횟수는 '본론' 3.9회를 선두로 하여, '이론적 배경' 3.2회, '결론' 2.8회, '목차' 2.4회, '서론' 2.2회, '참고문헌' 1.8회 순이었다.

4.3 전문 및 초록데이터에서 발견되는 주제어 분석결과

전문의 내용위치에 따른 주제어의 분포도를 확인하기에 앞서 학위논문 전문과 국영문초록에서 발견되는 주제어, 학위논문 저자가 부여한 주제어의 존재여부를 조사하였다. 학위논문마다 부여된 평균 12.2개의 주제어는 〈표 6〉, 〈표 7〉에서와 같이 9.38개 주제어에 해당하는 약 73%의 주제어가 전문에서 사용되고 있었다. 초록데이터에서는 국문초록 4.72개(37%), 영문초록 4.07개(32%), 국영문 전체 초록 8.37개(65%)의 주제어가 사용되고 있었다. 국문과 영문으로 구분되어 있는 초록에서 발견할 수 있는 특징은 언어표기에 따른 주제어의 사용개수이다. 국문초록에서 사용되는 주제어 개수가 영문초록에서 사용되는 주제어 개수보다 1~2개 주제어가 더 많이 사용되는 것으로 보이는데, 이것은 국문초록의 경우 영문으로 표기된 주제어를 혼용하여 사용할 수 있기 때문인 것으로 추측된다. 또한 국문초록에서 사용된 주제어 개수(4.72개)와 영문초록에서 사용된 주제어 개수(4.07개)를 더하면 국영문 전체 초록

〈표 5〉 전문의 내용위치별 용어 자동색인

구 분	학위논문 전문	전문의 내용위치					
		목차	서론	배경	본론	결론	참고문헌
용어 개수	1226.3	93.1	282.7	459.8	445.3	245.3	220.3
	100%	7.59%	23.05%	37.49%	36.31%	20.00%	17.97%
용어 출현개수	5152.3	224.7	613.0	1478.4	1749.1	688.4	396.2
	100%	4.36%	11.90%	28.69%	33.95%	16.36%	7.69%

〈표 6〉 전문 및 초록데이터의 주제어 사용율

구 분	학위논문 전문	초 록		
		전체(국문+영문)	국문	영문
주제어 사용율	73%	65%	37%	32%

〈표 7〉 학위논문 전문 및 초록에서의 주제어 개수

구 분	학위논문 전문	초 록		
		전체(국문+영문)	국문	영문
주제어 개수	9.38	8.39	4.72	4.07
	100%	89.45%	50.28%	43.37%
주제어 출현개수	472.05	26.14	15.35	10.23
	100%	5.54%	3.25%	2.17%

에서 사용되는 주제어 개수(8.39개)와 거의 유사함을 알 수 있다.

본 연구에서 사용된 학위논문 전문은 국문 또는 영문의 한 가지 언어로 기술된다. 초록데이터처럼 국문초록과 영문초록을 구분하여 작성하지 않는다. 하지만 연구에 사용되는 학위논문 주제어는 국문 주제어와 영문 주제어, 2개 언어의 주제어를 모두 포함하고 있다. 이것은 학위논문 전문의 사용언어에 따라 주제어와의 매핑률에 많은 영향을 줄 수 있을 것이다. 사용언어에 따른 매핑률 영향력을 조사하기 위하여 사용언어로 구분이 가능한 초록의 자동 색인어와 주제어 간의 매핑률을 비교하여 전문의 사용언어에 따른 주제어와의 매핑률을 유추하였다.

〈그림 3〉은 전문과 초록의 각 영역에서 사용되는 주제어 비율을 나타낸 것으로, 각 영역의 색인용어 중 몇 개의 색인용어가 주제어와 매핑되는가를 표기한 것이다. 특정 데이터 영역에서 발견되는 주제어의 개수 측면에서 보았을 때, 발견되는 주제어의 비중은 특정 데이터 영역에서 비슷한 것으로 나타났다. 국문초록, 영문초록, 국영문 전체초록에서 눈에 띄는 차이점을 보이지 않았다. 다만, 국영문 전체초록에서의 주제어 사용빈도(7.5%)가 단일 언어를 사용한 초록에서 보다 약간 높게(2~3%) 나타난다는 것이다. 각 데이터 영역의 색인용어와의 매핑에서 사용된 주제어는 국문 주제어와 영문 주제어 모두로 동일하게 적용하였다.

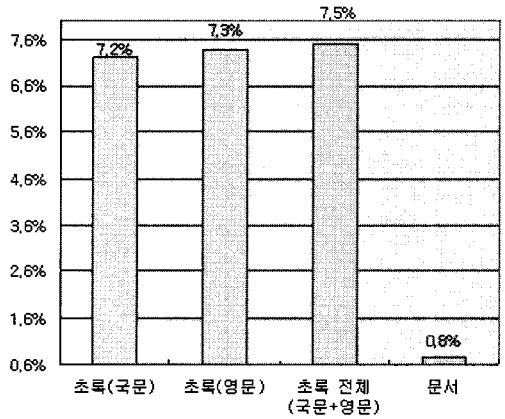
〈그림 4〉는 학위논문 주제어 가운데 각 데이터 영역에서 발견되는 주제어 개수를 비율로 표기한 것이다. 전문과 초록에서의 주제어 사용률에서 학위논문 전문이 월등히 높은 것으로 나타났다. 월등히 많은 색인용어가 추출되는 학위논문 전문에서 월등히 많은 주제어가 사용되고 있었다. 전문에서의 주제어 사용률은 국문초록의 2배, 영문초록의 2.3배에 이른다. 학위논문 전문이 국문 또는 영문 한 가지 언어로 작성된다는 점을 감안하면 많은 색인용어가 추

출된다하더라도 이것은 높은 수치이다. 이것은 연구대상 자료의 특징에서 그 원인을 유추할 수 있을 것으로 보인다. 본 연구에 사용된 과학기술분야 학위논문의 학문분야는 대부분 미국 등 영어권 국가가 주도하고 있는 실정이다. 따라서 학위논문에서 사용되는 전문용어 역시 영문인 경우가 많다. 따라서 국문으로 작성된 학위논문일지라도 내용 전달을 위해 영단어, 영문 주제어를 사용할 수 있을 것이다. 국문초록에서의 주제어 사용률(37%)이 영문초록에서의 주제어 사용률(32%)보다 높다. 그러나 국영문 전체 초록에서의 주제어 사용률(65%)은 국문초록의 주제어 사용률과 영문초록의 주제어 사용률을 합한 수치(69%)보다 떨어진다. 이것은 국문초록에서 영문 주제어가 사용될 수 있을 것이라는 추측을 가능하게 한다.

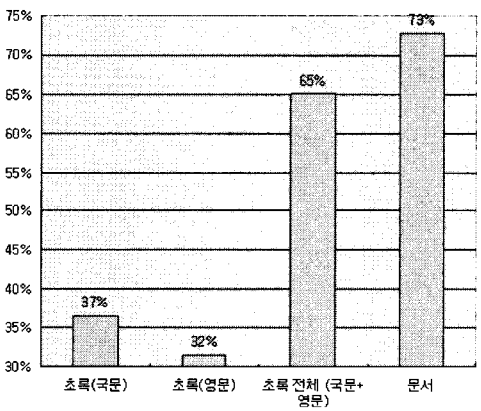
〈그림 4〉에서 발견할 수 있는 또 하나의 사실은, 전문에서의 주제어 사용률(73%)이 국영문 전체 초록에서의 주제어 사용률(65%)보다 무려 8% 가 높다는 것이다. 단일 언어를 사용하는 학위논문 전문과 2개 언어를 모두 사용하는 전체 초록의 비교에서 전문에서의 주제어 사용률이 높다는 것은 학위논문 전문을 기술하는 과정에서 주요 용어를 국영문 혼용하여 사용하고 있음을 강하게 시사한다. 또한 국영문 전체 초록에서조차 사용하지 않는 학위논문 주제어를 학위논문 전문에서는 사용하고 있음을 알 수 있다. 〈그림 4〉는 학위논문 전문에서의 주제어 사용률이 국문초록, 영문초록에 비해 2~2.3배에 이르는 높은 사용률을 보이고 있다.

4.4 학위논문 전문의 내용위치별 주제어 분석결과

학위논문 전문의 내용위치는 '목차', '서론', '이론배경', '본론', '결론', '참고문헌' 순으로 설정하였다. 전문을 각 내용위치에 해당하는 데이터로 분할하여 자동 색인을 한 후 색인용어와 주제어를 매핑·비교하여 일치 여부를 조사하였다.



〈그림 3〉 각 데이터 영역의 색인용어 개수 대비 발견된 주제어 개수



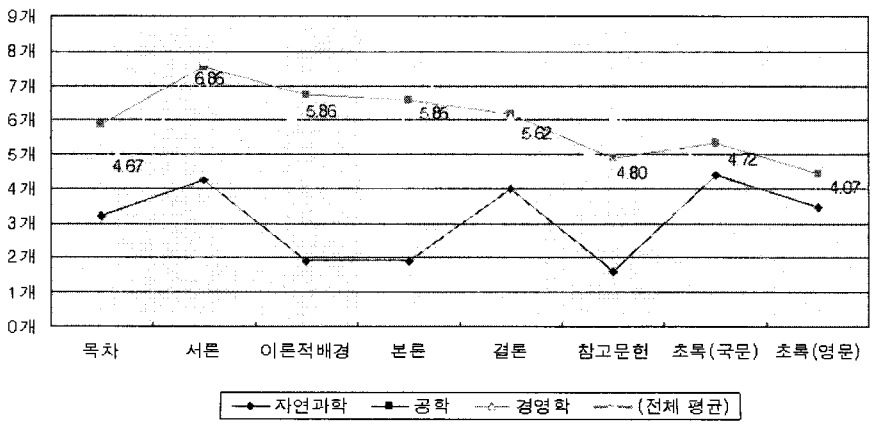
〈그림 4〉 전체 주제어 개수 대비 전문 및 초록에서 발견된 주제어 개수

〈표 8〉과 〈그림 5〉에서 보여주는 것과 같이, 학위논문 전문의 내용위치 가운데 '서론' 에서 가장 많은 주제어, 6.86개가 발견되었다. 이것은 '국문초록' 4.72개보다 높은 수치이다. 주제어가 가장 적게 발견된 전문의 내용위치는 '목차' 로 4.67개이다. 그러나 '목차' 에서 발견된 주제어의 개수는 국문 및 영문초록에서 발견된 주제어 개수보다 높은 수치라는 점은 인상적이다. 각 데이터 영역의 데이터 분량에 따라 좌우하는 주제어 출현개수는 내용위치 '본론'에서 가장 많은 출현을 보였다. 그리고 데이터 분량이 상대적으로 적은 '참고문헌'에서 21.04개, '목차'에서 25.53개의 가장 낮은 주제어 출현을 보였다.

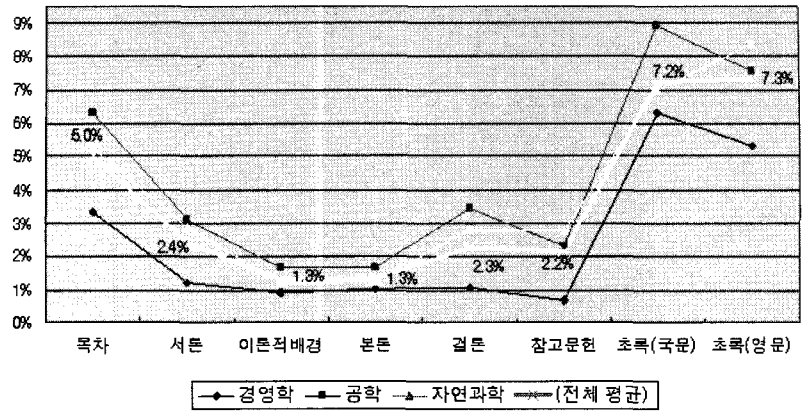
〈그림 6〉과 〈그림 7〉은 학위논문 전문의 내용위치에서 사용된 색인용어 개수와 비교하여 내용위치에서 발견된 주제어 개수를 확인한 것이다. 학위논문 전문은 상당히 많은 단어와 문장들의 집합이기 때문에 많은 색인용어가 추출되기 때문에 전문의 각 내용위치에서의 주제어의 사용률은 1.3%~7.3%로 그리 높지 않은 편이다. 초록을 제외하고 학위논문 전문의 내용위치에 따른 주제어 사용률을 살펴보았을 때, 가장 높은 주제어 사용률은 내용위치 '목차' 5.0%이었다. 물론 초록은 전문의 요약문이라 할 수 있기 때문에 학위논문 전문('목차'~'결론')의 내용위치와 비교하면 국문초록에서 가장 높은

〈표 8〉 학위논문 내용위치에서 발견된 주제어 개수

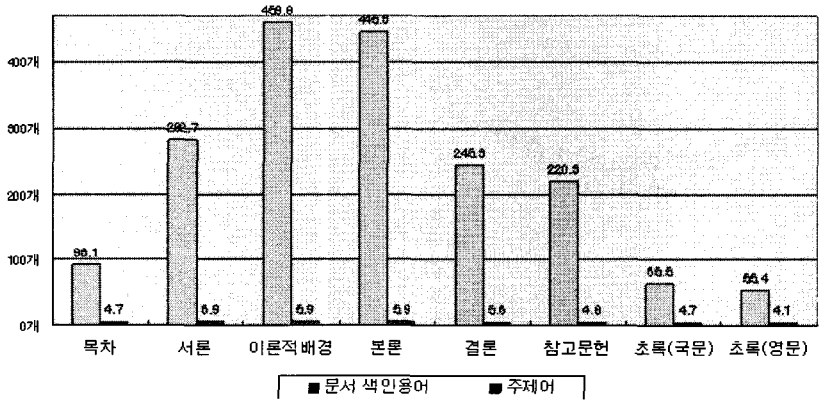
구 분	학위논문 전문의 내용위치							
	목차	서론	이론적 배경	본론	결론	참고 문헌	초록(국문)	초록(영문)
주제어 개수	4.67	6.86	5.86	5.86	5.62	4.80	4.72	4.07
주제어 출현개수	25.53	68.50	123.53	158.61	67.75	21.04	67.75	21.04



〈그림 5〉 학위논문의 내용위치별 발견된 주제어 개수 분포



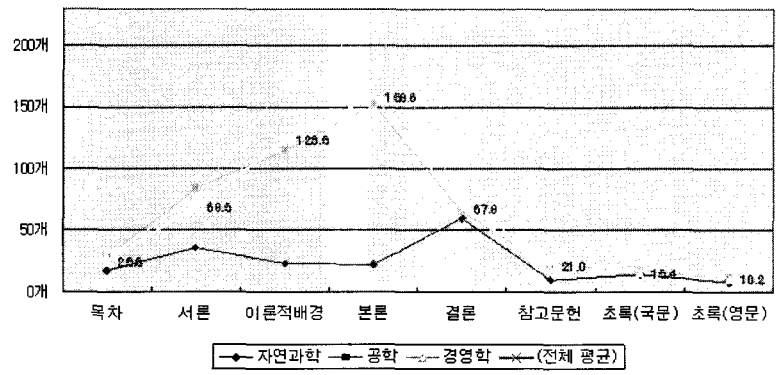
〈그림 6〉 학위논문의 내용위치별 색인용어 대비 주제어 사용률



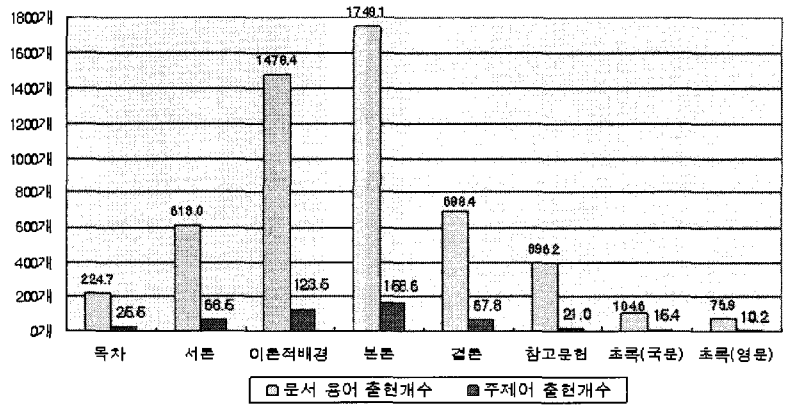
〈그림 7〉 학위논문의 내용위치별 색인용어 및 주제어 개수

주제어 사용률(7.3%)을 보인다. 그러나 본 연구방법에서 밝힌 바와 같이 본 연구에서 초록은 '목차', '서론', '이론적 배경', '본론', '결론'과 같은 전문의 내용위치의 자격이 아니라 이들 내용위치에서의 주제어 사용률을 수치적으로 비교하기 위한 참조자적 자격의 내용위치로 사용하였다. 〈그림 5〉는 내용위치 '목차'를 뒤이어 '서론' 2.4%, '결론' 2.3% 순의 주제어 사용률 순위를 보이고 있다. 내용목차와 그림/표목차 데이터로 이루어진 내용위치 '목차'에서 다른 내용위치에 비해 2배 이상의 주제어 사용률을 보인다는 것은 매우 인상적이다.

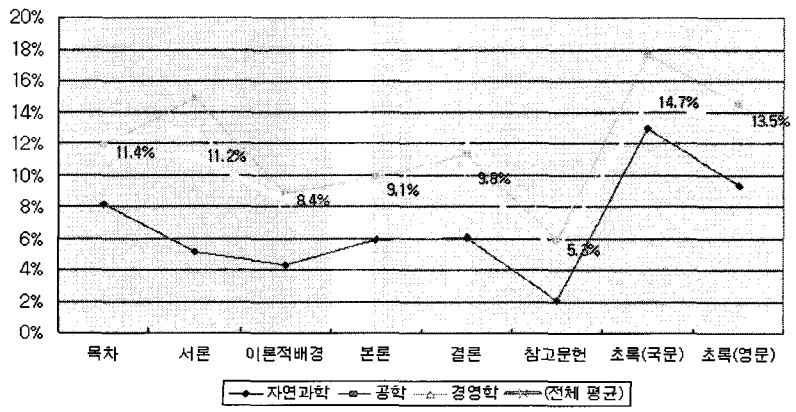
그러나 〈그림 8〉, 〈그림 9〉에서 보여주듯, 전문의 내용위치에 따른 주제어의 출현개수는 내용 텍스트가 가장 많은 '본론' 158.6개로 가장 높았다. 그 다음은 '이론적 배경' 123.5개, '서론' 68.5개, '결론' 67.8개, '목차' 25.5개, '참고문헌' 21.0개 순이었다. 그러나 주제어 출현율은 조금 다른 결과를 보여준다. 〈그림 10〉가 바로 그것으로, 이것은 학위논문 전문의 내용위치에서 사용된 색인용어 출현개수와 비교하여 내용위치에서 발견된 주제어 출현개수를 비율로 표기



〈그림 8〉 학위논문의 내용위치별 주제어 출현개수



〈그림 9〉 학위논문의 내용위치별 색인용어 및 주제어 출현개수



〈그림 10〉 학위논문의 내용위치에 따른 주제어 출현율

한 것이다. 〈그림 10〉은 비교대상에서 초록을 제외하였을 때 전문의 내용위치 '목차' 11.4%, '서론' 11.2%에서 가장 많은 주제어 출현율을 나타나고 있음을 보여 주고 있다. 내용위치 '결론'(10.2%)가 그 다음 순위의 주제어 출현율을 보이고 있다.

5. 결 론

전자도서, 전자저널, 연구보고서, 학위논문 등 수많은 문헌이 디지털형태로 만들어져 쏟아지는 디지털사회에서 방대한 텍스트 내용으로 이루어진 디지털자료에 대한 연구가 필요하다. 이것은 문서 자동요약, 자동분류, 클러스터링 연구분야에서도 제외가 될 수 없다. 그러나 디지털 문헌이 기계가독형이라 할지라도 방대한 텍스트 내용은 문서 자동분류 등의 연구에서 효율성을 떨어뜨리는 요소로 작용하기도 한다. 문헌은 방대한 내용으로 인하여 내용목차를 가지고 있는 것이 일반적이다. 그렇다면 문헌 전문을 내용목차에 따라 가중치를 부여한다면 문서 자동분류와 같은 연구분야에서 더 효과적인 결과를 기대할 수 있을 것으로 보인다.

본 연구는 K 대학교 석사 학위논문 386건의 학위논문 전문을 대상으로 내용목차에 따른 주제어 분포도를 연구한 것이다. 학위논문의 내용을 가장 잘 요약적으로 표현하는 주제어가 학위논문의 내용위치 가운데 가장 두드러지게 분포하는 위치를 확인하여 학위논문과 학위논문 주제어 간의 관계를 밝히는 것이다. 학위논문 전문을 자동 색인하고, 학위논문의 주제어와 동일한 색인용어를 찾는 과정을 통하여 학위논문 내의 주제어 분포도를 확인하였다. 학위논문의 주제어는 학위논문의 저자가 직접 작성한 주제어를 채택하였다. 연구자료로 채택한 학위논문은 아직까지 국내 대학의 여건상 XML과 같이 구조화된 문서로 작성되지 못하고 있지만 일반화된 과학기술분야 연구논문의 작성법이 존재하고, 소속 대학교에서 지정하는 학위논문 작성요령에 따라 작성되는 경향이 있으므로 어느 정도는 내용상 구조화 되어 있다고 볼 수 있기 때문이다. 학위논문의 전문을 내용 전개에 따라 '목차', '서론', '이론적 배경', '본론', '결론', '참고문헌'의 위치 내용으로 분할하여 학위논문의 내용 위치에 따라 주제어가 어떻게 사용되고 있는가를 확인하였다.

연구자료인 학위논문은 평균 1226.3개의 용어가 5152.3회 사용되고 있었고, 주제어는 평균 12~13개가 학위논문 저작자에 의해 부여되고 있었다. 이들 주제어는 국문과 영문으로 작성되어, 표기 언어별 5~6개의 주제어이었다. 학위논문 내용위치와 국문 및 영문초록에서의 주제어 분포도를 비교한 결과, 문서 내에서 사용되는 주제어의 개수는 '서론'(6.9개)에서 가장 많았고, 문서 내의 용어 출현개수에 따른 문서 내에서 사용된 주제어 출현개수는 '목차', '서론'에서의 주제어 출현율(11%)이 국문초록에서의 주제어 출현율(14%) 및 영문초록에서의 출현율(13%)과 유사한 것으로 나타났다. 여기서 나타난 1~2%의 차이는 사용언어에 따른 주제어 출현율을 비교실험에서 상쇄될 수 있는 수치일 것으로 사료된다. 향후 연구주제로는 학위논문 작성에 사용된 언어와 주제어의 사용언어를 일치시켜 언어 표현의 차이점으로 인하여 발생한 오차에 대한 추가 연구가 필요하다.

참 고 문 헌

국립중앙도서관. 2003. 『국립중앙도서관 주제명표목표 개발』. [cited 2007.07.01]. 〈http://www.nl.go.kr〉.

국민대학교. "KLT: Korean Language Technology: (구)HAM." [cited 2007.07.01]. 〈http://nlp.kookmin.ac.kr/HAM/kor/index.html〉.

김광해. 1987. 『유의어, 반의어 사전』. 서울: 한샘.

백지원, 최석두. 2002. 용어분류의 비교연구. 『제9회 한국정보관리학회 학술대회논문집』, 19-26.

안희국, 노희영. 2005. 문서 분류를 위한 문장 응집도와 주어 주도의 주제어 추출 『한국컴퓨터종합학술대회 논문집』, 32(1): 463-465.

유영준. 2003. 문헌정보학의 지식 구조에 관한 연구. 『정보관리학회지』, 20(3): 277-297.

이강일, 이창환. 2005. 주제어와 미분류 문서들을 이용한 문서의 자동 분류 방법. 『한국컴퓨터종합학술대회』, 32(1B): 592-594.

이경찬, 강승식. 2002. 범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템. 『한국정보과학회 봄 학술발표논문집』, 29(1): 475-477.

이영숙 외. 2001. 계층적 분류체계를 위한 자동분류 기법에 관한 연구. 『제8회 한국정보관리학회 학술대회 논문집』.

이창범, 김민수, 이기호, 이귀상, 박혁로. 2002. 주성분 분석을 이용한 문서 주제어 추출. 『정보과학회논문지: 소프트웨어 및 응용』, 29(10): 747-754.

이혜영, 곽승진. 2007. 학위논문의 주제어 분포에 관한 연구. 『제14회 한국정보관리학회 학술대회 논문집』.

이혜영. 2003. 『잠재적의미색인을 이용한 더블린코어 메터데이터 유사도 판단기법』. 석사학위논문, 충남대학교 대학원, 컴퓨터과학과.

한광록, 오삼권, 임기욱. 2004. 주제어구 추출과 질의어 기반 요약을 이용한 문서 요약. 『정보과학회논문지: 소프트웨어 및 응용』, 31(4): 488-497.

황재영, 이응봉. 2003. 자동문헌분류를 위한 대표색인어 추출에 관한 연구. 『제10회 한국정보관리학회 학술대회 논문집』, 55-64.

Amini, M. R. and P. Gallinari. 2002. "The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization." *Procedding of ACM SIGIR'02*, 105-112.

Battiti, R. 1994. "Using Mutual Information for Selection Features in Supervised Neural Net Learning." *IEEE Trans. Neural Networks*, 5: 537-550.

Chuang, W. T. and J. Yang. 2000. "Extracting Sentence Segments for Text Summarization: A Machine Learning Approach." *Proceeding of ACM SIGIR'00*, 152-159.

Craven, Timothy C. 2000. "Abstracts Produced Using Computer Assistance." *Journal of the American Society for Information Science*, 51(8): 745-756.

Damerau, Fred J. 1993. "Generation and evaluating domain-oriented multi-word terms from texts." *Information Processing and Management*, 29(4): 433-447.

Gil-Leiva, I. and A. Alonso-Arroyo. 2007. "Keywords given by authors of scientific articles in database descriptors." *Journal of the American Society for Information Science and Technology*, 58(8): 1175-1187.

Heery, Rache. 1996. "Review of Metadata Formal." *Program*, 30(4): 345-373.

Kwak, N. and C. Choi. 1999. "Improved Mutual Information Feature Selector for Neural Networks in Supervised Learning." *Int. Joint Conf. on Neural Networks (IJCNN'99)*: 1313-1318.

Lange, Holley R. and B. Jean Winkler. 1997. "Taming the Internet Metadata, A Work in Progress." *Advances in Librarianship*, 21: 47-72.

Marshakova-Shaikevich, Irina. 2005. "Bibliometric Maps of Field of Science." *Information Processing and Management*, 41: 1534-1547.

Moens, Marie-Francine. and Jos. Dumortier. 2000. "Text categorization: the assignment of subject descriptors to magazine articles." *Information Processing & Management*, 36(6): 841-861.

Novovicova, Jana., Antonin. Malik, and Pavel. Pudil. 2004. "Feature Selection Using Improved Mutual Information for Text Classification." *LNCS*, 3138: 1010-1017.

Sano, Hikomaro. 1990. "Facet Tabulation of Index Terms." *Information Processing and Management*, 25(4): 543-548.

Silvester, June P. 1993. "An Operational System for Subject Switching Between Controlled Vacabularies." *Information Processing and Management*, 29(1): 47-59.

Soucy, Pascal. and Buy W. Mineau. 2003. "Feature Selection Strategies for Text Categorization." *LNAI*, 2671: 505-509.

Yang, Yiming. and Jan O. Pedersen. 1997. "Comparative Study on Feature Selection in Text Categorization." *Proceedings of the 14th ICML97*: 412-420.