구문·통계적 기법을 이용한 한국어 자동색인에 관한 연구

An Experiment in Automatic Indexing with Korean Texts: A Comparison of Syntactico-Statistical and Manual Methods

서 은 경(Eun-Gyoung Seo)*

□목 차□

초 록

본 논문은 자연어 형태의 한국어 텍스트로 부터 주제를 대표할 수 있는 색인어를 자동으로 추출하는 실험적인 구문·통계적 자동색인 시스템을 구현하였다. 구문·통계적 자동색인 시스템은 형태소 분석과 단어가중 기법을 이용하여 단일어와 명사구를 동시에 선택하는 자동색인 시스템을 말한다. 시스템의 성능을 측정하기 위하여, 300 개의 우리말 학술 및 학위논문 초록에서 선택된 단일·복합어 색인어를 수작업 색인과 비교하였다. 이와 같은 실험 결과를 가지고 아직 미흡한 연구상태인 우리말 자동색인 개발에 있어서 필요한 기초자료를 제시하였다.

ABSTRACT

This study was undertaken in order to develop practical automatic indexing techniques suitable for Korean natural language texts. It has taken a modest step toward this goal by developing an automatic syntactico-statistical indexing method and evaluating the method by comparing the resutls with manual indexing. For this experimental study, the Korean text database was constructed manually based on 300 abstracts covering business subject. The experimental results showed that the performance of the automatic syntactico-statistical indexing system was comparable to that of other studies which have compared automatic indexing with manual indexing.

* 한성대학교 문헌정보학과

Ⅰ. 서 론

1.1 연구의 배경

정보 검색 시스템에서 문헌과 탐색문은 보통 용어(term), 키워드(keyword), 디스크립터(descripter)라고도 불리우는 색인어(index term)로 변환된다. 이와 같은 변환 작업—한 문헌의 내용을 분석하여 그 중심주제를 표현하기에 가장 적합한 용어, 즉 색인어를 선정하여 문헌에 부여하는 일—을 색인 작업이라 한다(Salton, 1989). 특히 이 작업이 컴퓨터에 의하여 행해졌을 때 보통 우리는 자동색인이라 부르며, 정영미(1987)는 '자동색인'을 컴퓨터에 입력된 문헌의 텍스트를 분석한 뒤, 자동적으로 한 문헌의 주제내용을 대표할 수 있는 단어나 단어구를 추출해 내는 것이라고 정의 내렸다. 자동색인은 룬(Luhn)이 1957년 커뮤니케이션 이론에 근거하여 문헌에 출현한 단어들은 문헌의 내용 분석을 위하여 사용할 수 있으며 단어의 출현 빈도가 그 단어의 주제어로서 중요성을 측정하는 기준이 된다는 가설 아래 자동색인에 관한 연구를 발표 한 후, 계속해서 활발히 연구되고 있는 분야이다. 최근에는 자동색인에 인공지능 분야의 자연어 처리 기법, 전문가 시스템, 또는 언어학 이론 등을 도입한 연구들이 활발히 진행되고 있으며 자동색인은 정보학의 중심 주제 분야로 각광 받고 있다.

자동색인의 기본 원리는 문헌을 구성하는 단어들을 일정한 기준에 의해 주제어와 비주제어, 의미어와 무의미어, 또는 전문어와 비전문어로 구분하고 주제어, 의미어, 또는 전문어로 평가된 단어들로 부터 보다 적합한 색인어를 인간의 중재없이 선정하는 것이다(정영미 & 이태영, 1982). 따라서 자동색인을 연구하는 학자들은 색인어를 선정 또는 구분하는 기준을 파악하려고 노력해 왔다. 초기의 자동색인에 관한 연구는 단어의 출현빈도를 근거로 하여 주제어로서의 중요도를 측정한 다음 색인어를 선정하는 통계적 방법을 이용한 것으로 단어의 의미를 통계적으로 측정할 수 있는 다양한 기법을 제시하였다. 이와 같은 통계적 기법의 대다수는 문헌에서 단어의 상대적 중요도에 따라 일정한 값을 부여하는 단어가중 기법(term weighting technique)으로, 대표적인 기법으로는 문헌내 단어빈도(Within Document Frequency, WDF), 역문헌빈도(Inverse Document Frequency, IDF), 문헌분리가(Term Discrimination Value, TDV), 확률이론(Probablistic Theory) 등이 있다 (Luhn, 1957 ; Sparck Jones, 1972 & 1973 ; Salton et al., 1973 & 1975 ; Bookstein & Swanson, 1974 ; Harter, 1975a /b ; Bookstein & Kraft, 1977). 따라서 초기의 자동색인 시스템은 불용어 사전을 이용하여 색인어 대상이 될 모든 단일어를 추출한 뒤, 위의 기법 중 한가지를 이용하여 가중치를 산출한 다음 가중치의 값이 일정한 한계치의 범위에 드는 단어를 색인어로 선택하는 것을 말한다. 이와 같은 단어가중 기법을 이용한 단일어 색인 이론은 실용적인 면에서, 검색효율면에서 매우 효과적인 것으로 나타났다(Salton, 1985).

통계적 기법을 이용한 단일어 자동색인 시스템이 정보검색 시스템에서 효율적인 것으로 나타난 것은 사실이나, 문헌에서 추출하여 문장밖으로 떼어 놓여진 단일어는 실제 문헌의 내용/주제의 오직 일부분만 설명해 주거나, 의미를 정확히 전달하지 못하는 단점을 가지고 있다. 다시 말하자면 같은 단어라도 그 단어가 어떤 문맥 속에서 사용되었느냐에 따라 그 의미가 달라질 수 있기 때문에 자연어 텍스트에

서 추출된 색인어는 때때로 의미어로서 적합하지 않을 경우도 있다. 이러한 단점을 극복하고 자동색인의 결과를 향상시킬 수 있는 여러 방법들이 계속해서 제안되었다.

그 중 대표적인 방법으로 디소오러스를 이용하는 방법, 관련된 용어들을 연결시켜 주는 연관어 지도(Term Association Maps)를 이용하는 방법, 또는 보다 정확한 내용을 담고 있으며서 한 단어로 처리될 수 있는 명사구, 전치사구 등을 추출하여 색인어로 이용하는 방법 등이 있다(Salton, 1986). 먼저, 디소오러스는 색인 작업시 적절한 색인어의 선택과 색인어의 통제를 위해 사용될 뿐 만 아니라 검색시에는 적절한 탐색어의 선택을 위해 사용된다. 따라서 주제 전문가에 의하여 잘 구성된 디소오러스는 주제 색인어의 통제와 탐색어 확장을 통하여 검색효율의 향상에 기여하게 된다. 디소오러스를 이용하여 자동색인 시스템을 개발한 살톤과 레스크(Salton & Lesk, 1968)는 그들의 시스템이 단어가중 기법만을 이용한 자동색인 시스템 보다 높은 재현율을 보였다고 발표하였다. 그러나, 디소오러스의 작성은 한 특정 주제에 대한 심오한 지식과 신중한 작성과정을 요구하므로 디소오러스의 작성은 무척 어려우며 과학적 작업이라 보기 보다는 하나의 예술적 작업이라 할 정도로 매우 주관적인 작업이 바로 디소오러스 작성 작업이다. 따라서 자동색인 시스템을 위하여 디소오러스를 새롭게 구축하는 경우는 매우 드물고 보통 기존에 있는 디소오러스를 이용하여 자동 색인어의 어휘를 통제·확대시키는, 즉 추출된 자연어 색인어를 통제 어휘로 대체시키는 시스템이 많이 구축되었다(Van der Meulen & Janssen, 1977 ; Brenner et al., 1984 ; Vleduts-Stoko-lov, 1987). 마티네(Martinez et al., 1987)는 자연어를 디소오러스의 디스크립터로 변환시킬 때 생기는 문제점을 상세히 논의하면서 그 어려운 점을 알려 주고 있다. 많은 학자들이 디소오러스를 이용하는 자동색인 시스템이 높은 효율성을 가진다는 것은 인정하나, 정확하며 완벽한 디소오러스의 작성의 어려움으로 인하여 많이 이용되고 있지 않는 실정이다.

연관어 지도를 이용하여 색인어를 확장시키는 방법을 보통 연관색인(associative indexing)으로 불리워 진다(Doyle, 1961). 연관어 지도를 이용하는 자동색인 시스템은 탐색시 연관된 여러 색인어를 제공함으로써 재현율을 높여주는 시스템을 말한다(Lesk, 1969). 그러나 용어의 의미는 전혀 고려하지않고 용어의 통계적 특성(용어의 동시출현 빈도)만을 이용하여 구성된 연관어 지도는 단순히 한 용어와 다른 용어의 연관 관계만을 제시해 주기 때문에 때때로 잘못된 연결을 유도하기 한다. 따라서 살톤(1986)은 연관어 지도를 이용한 자동색인 시스템의 성능이 항상 좋을 것이라는 가설은 입증될 수 없다고 했다.

색인 시스템의 재현율을 높이는 방법이외에, 정확률을 높이는 기법도 연구되었다. 그 하나 방법이 색인어로 단어구를 이용하는 것이다. 단어구를 이용하는 목적은 단어구가 보다 뚜렷한 '개념(concept)'을 전달시키며 색인어의 특정성을 높일 수 있기 때문이다(Fagan, 1988). 따라서 단어구의 이용은 탐색시 정확율을 높여주므로써 문헌정보검색 시스템의 효율성을 향상시켜 줄 것이라는 기대를 받았다. 자연어 텍스트에서 단어구를 선택하려는 첫번째 시도는 단어의 동시출현 빈도에 근거하여 단어구를 형성하는 통계 연관법(Statistical association methods), 확률이론을 이용한 확률적 단어 의존 모델(Probablistic

term dependency)과 같은 통계적 방법을 기초로 하여 이루어졌다(Salton et al, 1981). 그러나 이러한 통계적 방법을 이용한 색인시스템의 실험의 결과는 실망적이었다: 한 실험에서 20%의 검색효율이 향상 되었으나 다른 실험에서는 어떤 차이도 발견 할 수가 없었다(Salton & Buckly, 1988). 또한 여러 연구에서 통계적 기법으로 추출된 단어구가 색인어로서 적합하지 못하며, 자연어 문장에서 추출된 단어구는 오직 그 문헌에서 유용할 뿐이라는 결과도 나왔다(Cleverdon et al., 1966 ; Salton & Lesk, 1968 ; Lesk, 1969 ; Cagan, 1970). 따라서 통계적 데이타를 이용하여 단어구를 추출하는 방법의 한계점이 확인되자, 이외에 다른 방법 즉 언어학적 기법에 의한 단어구 선택 방법론이 1970년대 후반 부터 활발히 연구되기 시작하였다.

비통계적인 기법인 언어학적 기법은 문헌의 의미분석에 기초한 것으로서, 단어의 문법적인/구문적인 이용이 문헌내용의 결정에 도움을 줄 수 있다는 가정에 기초한 것이다(허미숙, 1990). 언어학적 기법에 속하는 여러 방법 중 주로 단어의 구문적인 형태를 분석하며 단어구를 추출하는 구문분석 기법이 주류를 이루고 있다. 구문분석 기법에는 문장을 문법적으로 분석하여 전치사구나 명사구 등 한 단어로 처리되는 단어군을 추출하는 부분적 문장분석 기법과 문장의 모든 구문적 규칙을 컴퓨터에 소장하여 단어구의 유형과 구조를 자동적으로 식별하여 색인어를 선택시키는 완전구문분석 기법 등이 있다. 존스(Jones et al., 1990)는 구문분석을 근거로 하여 선택된 단어구는 문헌의 내용을 나타내는 지표로 중요한 역할을 수행할 뿐만 아니라 검색시스템의 재현율과 정확률을 높혀 준다고 하였다. 또한 화간(Fagan, 1988)는 구문분석 방법과 비구문분석 방법을 비교하는 논문에서 존스의 결론을 지지 하였으나, 어의적 처리를 하지 못하는 구문분석 방법이 색인어로 적합한 단어구를 선정하는데 한계점을 가지고 있고 이를 극복하기 위하여 어의 분석이 필요하다고 강조하였다.

최근에 여러 연구자들은 자동색인을 위한 의미분석에 대하여 심층적으로 연구하고 있으며 몇 개의 새로운 기법을 제안하였다. 그 대표적인 예로 의미분석기, Semantic Analyser SEMAN(Jonak 1984 ; Smetacek, 1984) ; 잠재적 의미분석기법, Latent Semantic Indexing (Lochbaum & Streeter, 1989 ; Deerwester et al., 1990) ; 인공지능 기법을 이용한 전문가 시스템 또는 지식기반 시스템 (Humphrey & Miller, 1987 ; Driscoll et al., 1991) 등이 있다. 이와 같은 연구 또는 방법론은 자동색인의 새로운 장을 열고 있으나 아직까지는 분석, 수행된 주제의 범위가 매우 한정되어 있고 의미론에 대한 언어학자 간의 이론이 정립되지 않는 상태여서 개발 및 수행과정이 무척 복잡하고 그 연구 진행 과정이 무척 느린 상태이다. 정영미(1987)는 어의 분석이 포함된 언어학적 기법은 그 복잡성에 비추어 그다지 큰 효과가 기대되지 않는다고 했고, 화간(1989)은 그의 최근 논문에서 자동색인 시스템을 위한 완전구문 분석과 의미분석의 잠재적 가치가 실제 실용적 시스템에서 어느 정도 응용이 될지에 관해서는 확신이 서지 않는다고 말하고 있으며, 랑카스타(Lancaster, 1991)는 '인간의 중재 없이 완벽하게 자동으로 색인어를 선정할 수 있는 단계는 아직 오지 않았다'고 말하고 있어 그 어려운 점을 알려주고 있다.

이와 같이 과거 30년동안 자동색인 시스템 또는

색인 기법을 개발시키고자 막대한 노력을 해왔으나, 기존에 개발된 시스템은 각자 장점과 그 한계점을 뚜렷이 가지고 있으며 아직까지도 자동색인 기법이 지닌 근본적 문제점(예 : 자연어 텍스트를 분석하는 가장 좋은 방법은? 색인어를 선정하는데 필요한 기준의 이론적 근거는 무엇인가? 어떤 종류의 색인 기법이 계속해서 좋은 결과를 산출하는가? 등)을 해결하지 못하고 있다. 더구나 해결 방법에 대한 여러 학자들의 의견 또한 일치되고 있지도 않는 실정이다. 따라서 본 연구는 이런 문제점의 해결을 위한 또하나의 시도로서 자연어 형태의 한국어 텍스트에 적합하며 효율적으로 색인어를 자동으로 추출할 수 있는 방법을 제시하고자 한다.

1.2 연구의 목적 및 연구문제

본 논문은 자연어 형태의 한국어 텍스트로 부터 주제를 대표할 수 있는 색인어를 자동으로 추출하는 실험적인 구문·통계적 자동색인 시스템을 구현하는데 목적이 있다. 색인어의 구문적인 특성과 통계적인 특성을 동시에 이용한 구문·통계적 자동 색인 시스템은 형태소 분석과 단어가중 기법을 이용하여 단일어와 명사구를 동시에 선택하는 자동 색인 시스템을 말한다. 특히 본 연구는 수작업 색인의 결과를 이용하여 본 연구에서 구현된 자동색인 시스템을 평가 하였고 부수적으로, 구현된 자동색인 시스템에 단어가중 기법과 초록의 길이가 어떻게 영향을 미치는 가를 조사하였다. 따라서 시스템은 두가지의 단어가중 기법, 문헌내 단어빈도 기법(WDF)과 역문헌 빈도 기법(IDF)을 각각 사용하여 색인작업을 수행했으며 실험문헌은 두 종류의 초록, 학술잡지 논문 초록(짧은 초록)과 학위논문 초록(긴 초록)으로 구성되었다.

본 논문에서 조사하고자 하는 연구 문제는 다음과 같다.

1. 구문·통계적 기법을 이용한 자동색인 시스템이 과연 "주제를 대표할 수 있는" 색인어를 추출할 수 있는 가?
2. 단어가중 기법(WDF 와 IDF)이 본 시스템에서 구문분석만을 이용하여 추출된 명사·명사구의 중요도를 측정하여 이들로부터 색인어로서 보다 적합한 명사·명사구를 선택할 수 있는 가?
3. 역문헌 빈도 기법(IDF)을 이용한 자동색인 시스템이 문헌내 단어빈도 기법(WDF)을 이용한 시스템보다 더 좋은 결과를 가질 수 있는 가? 자동색인 시스템의 효율성을 선정률과 정확률로 평가된다면, IDF시스템이 WDF시스템보다 더 높은 선정률과 정확률을 얻을 수 있는 가?
4. 초록의 길이가 본 연구에서 구현된 자동색인 시스템의 효율성에 영향을 미치는가? 즉 짧은 초록을 이용한 자동색인 시스템이 긴 초록을 이용한 시스템보다 더 높은 선정률과 정확률을 얻을 수 있을까?

1.3 연구의 범위 및 제한점

색인어는 정보원과 정보 입수자 사이에 위치하여 특정한 주제의 문헌들을 선별해 주고 선별된 자료의 소재를 지시하여 주는 기능을 수행한다. 따라서 효과적인 정보검색은 적절한 색인의 사용을 전제로 하고 있으며 색인의 성능이 결국 정보시스템의 성능을 좌우하게 된다. 일반적으로 정보시스템의 성능을 높일 수 있는 좋은(Good) 색인어에 대한 적절한 정의는 존재하지 않고 있다. 존 오코너 (John O'connor, 1961)는 "좋은 색인어란 좋은 검색 결과를 유발시키는 용어"라고 정의

했고 많은 연구자들이 이 정의를 인용하나, 색인어를 검색효율만을 가지고 평가한다는 한계점을 지니고 있다. 즉 검색효율이 실제 색인어가 지닌 어의적 적합성을 평가해주지 못하기 때문이다. 본 연구에서 정의 내린 좋은 색인어란 색인 전문가나 주제 전문가에 의하여 그 문헌에 적합하며 문헌의 내용을 나타내는 주제어라고 판정될 수 있는 용어를 뜻하며, 본 연구는 이와 같은 용어를 선정할 수 있는 방법 즉 구문·통계적 자동색인 시스템을 구현하였다.

본 연구는 실험적 연구로써 그 제한점으로는 완벽한 형태소 분석을 못한 점(형태적 중의성을 자동적으로 처리 못한 점), 입력 데이타의 모든 표준화 처리를 자동적으로 처리 못한 점, 제한된 데이타 규모(300 개의 초록)를 이용한 점, 주제 영역을 경영/경제학으로 제한한 점, 단어가중 기법을 수치 산출이 용이한 두가지 방법으로 제한한 점 등을 들 수 있다.

Ⅱ. 연구 방법

본 연구의 주 목적은 한국어 텍스트에서 색인어를 추출시키는 구문·통계적 자동색인을 개발한 다음 수작업 색인과 비교하여 구현된 시스템의 효율성을 평가하는 것이다. 따라서 본 2장에서는 구문·통계적 자동색인 시스템을 설계 및 구현하는 방법과 구현된 시스템을 수작업색인 시스템과 비교·평가하는방법을 제시하였다.

2.1 자동색인 시스템 설계 및 구현

2.1.1 시스템 개요

본 시스템은 문헌의 내용(주제)를 나타내는 단일어(단일 명사) 또는 명사구는 특정 구문 범주에 속한다는 가정을 이용하여, 자연어 형태의 한국어로 표현된 문장에서 구문적으로 적합한 명사 및 명사구를 자동으로 추출한 다음 통계적 가중치를 이용하여 이 중 색인어로서 부적합한 명사·명사구를 제거하고 가중치가 높은 명사·명사구를 색인어로 선정하는 자동색인 시스템을 말한다.

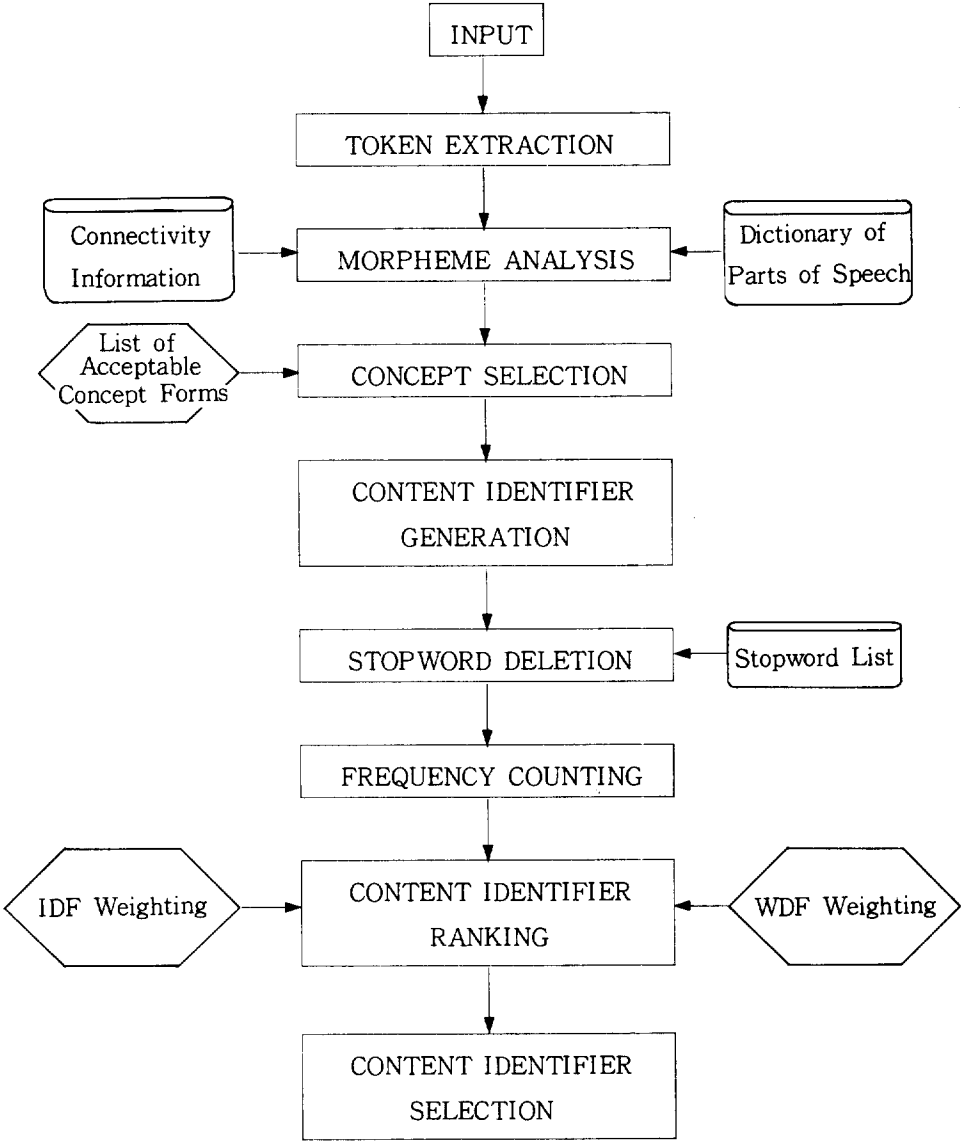
시스템의 대략적인 내용을 살펴보면 다음과 같다. 시스템은 먼저, 형태소 분석기를 이용하여 텍스트에 있는 각 단어를 형태소 코드로 변환한 다음, 단어의 구문적 범주(형태소)를 미리 작성된 구문패턴 사전과 비교하여, 구문패턴 사전에 있는 패턴과 일치한 단일 명사 또는 명사구를 추출한다. 구문패턴 사전은 한 문장에서 색인어로 적합한 주제어나 주제구를 추출하기 위하여 작성된 것으로, 색인어가 보편적으로 갖고 있는 구문적 패턴의 리스트를 말한다.

다음, 시스템은 출현빈도로 이용한 단어가중 기법을 근거하여 모든 문장에서 추출된 명사나 명사구에 주제어로서의 중요도 값(가중치)을 부여한 후, 가중치 값을 이용하여 보다 색인어로 적합한 명사·명사구를 선택한다. 즉 단어가중 기법을 이용하여 한 문헌에 나타난 모든 명사·명사구의 순위를 매긴 다음 10위안에 든 명사·명사구를 최종적으로 각 초록에 해당하는 색인어로 선택하는 것이다. 본 연구는 두 종류의 단어가중 기법을 사용하였으므로 실험 결과 또한 두 종류로 형성되었다. 시스템의 전체적인 구성은 〈그림 1〉과 같다.

2.1.2 시스템 설계

1) 입력 데이타 처리 모듈

본 실험에서는 경영/경제학 분야의 100 개의 학위논문과 200 개의 학술잡지 논문을 실험 대상 문



〈그림 1〉 시스템의 전체 구성도

헌으로 선정하여 그 표제와 초록으로 부터 색인어를 추출하도록 하였다. 100 개의 긴 초록은 1988년에서 1992년 사이에 제출된 연세대학교 대학원의 석·박사 학위논문(경영학 전공)으로 구성되었고 짧은 초록은 6개의 학술잡지에서 실린 200 개의 최신 논문으로 구성되었다. 선정된 경영/경제학 분야의 학술잡지는 다음과 같다 : 한국개발 연구/한국개발연구소 ; 경영학 연구/한국 경영학회 ; 증권학회지/한국 증권학회 ; 경영 연구/경영연구소, 고려대학교 ; 경제 논집/경제연구소, 서울대학교 ; 경영과학/한국 경영과학회.

모든 초록은 우리말로 쓰여진 것으로, 긴 초록은 150 개 이상의 어절을 가진 초록을 말하며 짧은 초록은 50 개에서 100 개의 어절을 가진 초록을 뜻한다. 따라서 여기에 해당하지 않는 초록(너무 짧은 학위논문 초록이거나 너무 긴 학술잡지 논문의 초록)은 실험 문헌으로 선정되지 못했다. 이와 같이 선정된 실험 데이타는 시스템 실험을 위하여 수작업으로 입력되었고, 각 문헌의 입력물은 문헌번호, 저자명, 서지사항, 문헌의 표제 그리고 초록이며, 입력시 다음과 같은 통제를 가했다.

첫째, 동일 단어의 출현 빈도의 분산을 막기위하여 실험 데이타는 한글 맞춤법 문법에 맞게 고쳐서 입력시켰고, 띄어쓰기는 "단일어 띄어쓰기"를 기준으로하여 모든 문장을 그 기준에 통일시켰다. 복합명사는 사전에 표제어로 수록된 것만을 인정하였다.

둘째, 실험 문헌 중 국한문 혼용문의 경우는 한자를 한글로 변환하여 입력시켰다.

2) 형태소 분석 모듈

형태소 분석 모듈에서는 자연어 문장의 형태소를 분석한다. 형태소 분석은 어절단위로 이루어지기 때문에 시스템은 먼저 문장에서 공백(빈칸)이나 문장부호를 이용하여 어절을 분리시키고, 분리된 어절은 한국어 어절내 형성되는 형태소 결합구조를 근거로 하여 형태소로 다시 분리되고 각 형태소는 방대한 품사사전에 근거하여 품사를 부여받는다.

본 연구는 형태소를 분석하기 위하여 "형태소 분석기"를 개발하는 대신 연세대학교 한국어 사전 편찬 센타에서 초기에 개발된 프로토타입 형태소 분석기를 사용하였다. 이 형태소 분석기는 연세 말뭉치 I을 구성하고 있는 모든 단어를 기초로 하여 형성된 품사사전과 형태소 결합정보를 이용하여 어절을 분석한다. 이 분석기는 기능어(조사, 어미, 접사)를 먼저 처리하는 우에서 좌로 분석하는 방법을 이용하며, 분석 도중 중위성을 가지는 어절에 대해서 가능한 모두 분석한다.

3) 명사·명사구 선택 모듈

이 모듈의 목적은 한 문장에서 미리 정해진 구문패턴과 일치하는 단어·단어구를 식별하여 색인어로 선택하는데 있다. 따라서 의미있는 단어 및 단어구(색인어)의 형태소 구성에 대한 정의를 내리는 것이 중요하다. 자연어 형태의 한국어로 표현된 문장을 대표하는 색인어는 거의가 단일 명사나 명사구로, 명사구에는 명사 나열형 명사구, 관형형 또는 관형구가 포함된 명사구, 조사 및 전치사가 포함된 명사구, 접속부사가 첨가된 명사구 등이 있다는 허미숙(1992)의 조사에 근거하여, 본 시스템에서는 단일 명사, 명사 나열형 명사구, 조사 '의' 첨가형 명사구, 관형형 '-적' 첨가형 명사구와 같은 형태를 지닌 명사구를 주제를 나타내는 대표적인 색인어 형태로 보았다. 명사구의 정의가 상당히 제한적이기는 하지만 구문적으로

부적합한 명사구의 선택을 피한다는 장점을 지닌다.

본 시스템에서 설정한 명사구의 정의를 문맥자유문법을 이용하여 표현하면 다음과 같다.

명사구={명사+[의|-적]}*+명사

('{ }*' : 반복, '|' : 선택, '[]' : 생략가능)

이와 같은 명사구의 정의에 따라, 본 연구는 최대 네 개까지의 단어로 구성된 구문패턴을 규정지었고, 자연어 문장에서 나타난 일련의 형태소의 조합형식이 미리 정의된 구문패턴과 일치된다면 색인어로써 추출된다. 본 연구에서 설정한 색인어의 구문패턴은 표 〈1〉 과 같다.

구문패턴의 리스트는 색인어로 적합한 주제어를 추출하기 위하여 구성된 것으로, 4개의 태그를 이용한 256 조합 방법 중에서 색인어가 지닌 가장 보편적 구문패턴인 58개로 구성되었다. 시스템은 최장 일치의 원칙을 적용하였고, 구문 형태가 확인되어 추출되면, 또다른 형태(짧은 패턴)으로는 추출되지 않는다. 따라서 구문패턴과 일치하여 추출된 명사·명사구는 그 문장에서 주제를 나타내는 "색인어"로 인정되어 조사와 불용어를 제거하기 위해서 다음 모듈로 넘어가며, 구문패턴과 일치하지 않은 모든 어절은 자동적으로 제거된다.

4) 조사 및 불용어 제거 모듈

이 모듈은 전 단계에서 추출된 명사·명사구의 끝에 붙어있는 조사를 제거한 다음, 불용어 사전에 있는 단어 또는 어절을 탈락시킨다. 본 논문에서의 불용어 사전은 수작업으로 구성되었고 불용어를 유형별로 구분해 보면 다음과 같다 : 불완전명사 따위의 기능어(common function word) ; 내용 규정에 관계가 없는 명사(common word) ; 구문론적으로 잘못 분석된 명사(ambiguous noun : 명사 형태소 태그를 지녔으나 실제로 명사가 아닌 단어)들이 있다.

5) 색인어 순위부여 모듈

구문패턴을 이용하여 한 초록에서 추출된 명사·명사구의 수는 보통 25 개 이상으로 이 모든 것이 색인어로 사용될 수 없기 때문에 추출된 명사·명사구 중 어느 것이 보다 색인어로 적합한 가를 결정할 필요가 있다. 본 연구는 한 문헌 또는 문헌집단에서 추출된 색인어의 상대적 중요성을 나타내는 값, 즉 가중치를 이용하여 추출된 명사·명사구의 주제어로서의 중요도 순위를 부여하였다. 따라서 시스템은 그 값을 기준으로 하여 일정한 한계치의 범위 안에 드는 단어·단어구를 쉽게 선택할 수 있다.

시스템은 여러 단어가중 기법 중에서 보편적으로 가장 많이 사용되며, 가중치를 용이하게 산출할 수 있는 두가지 방법을 이용하였다. 사용된 단어가중 기법은 룬이 발표한 문헌내 단어빈도 가중기법과 스파크 존스의 역문헌빈도 가중기법이다. 문헌내 단어빈도 가중기법의 공식은 $WDF_{ki}=tf_{ki}$ 로 표현 될 수 있고 'WDF_{ki}'는 주어진 문헌 i에서 나타난 명사·명사구(k)의 단어빈도 가중치를 의미하고, 'tf_{ki}'는 주어진 문헌 i에서 명사·명사구(k)의 단어빈도수를 뜻한다.

다음 역문헌빈도 가중기법은 $IDF_{ki}=Log_2n-Log_2df_k+1$로 나타내며, '$IDF_{ki}$'는 주어진 문헌i에서 명사·명사구(k)의 역문헌빈도 가중치이고, 'n'는 전체 문헌의 수 (본 연구에서는 300), 'df_k'는 명사·명사구(k)가 할당된 문헌의 수를 뜻한다.

〈표-1〉 58개의 구문패턴

구문형식	구문패턴
= 단일명사	
1. N1	3, 4, 5
= 명사 나열형 명사구	
2. N1, N2,	
3. N1, N2, N3	33, 34, 35
4. N1, N2, N3, N4	333, 334, 335
= 조사 '의' 첨가형 명사구	
5. N1+AP , N2	43, 44, 45
6. N1, N2+AP, N3	343, 344, 345
7. N1+AP, N2, N3	433, 434, 435
8. N1+AP, N2+AP, N3	443, 445
9. N1, N2, N3+AP, N4	3343, 3344, 3345
10. N1, N2+AP, N3, N4	3433, 3434, 3435
11. N1, N2+AP, N3+AP, N4	3443, 3445
12. N1+AP, N2, N3, N4	4333, 4334, 4335
13. N1+AP, N2, N3+AP, N4	4343, 4345
14. N1+AP, N2+AP, N3, N4	4433, 4435
= 관형형 '-적' 첨가형 명사구	
15. N1+M, N2	13, 14, 15
16. N1+M, N2, N3,	133, 134, 135
17. N1+M, N2, N3, N4	1333, 1334, 1335
= 조사 '의'와 관형형 '-적' 첨가형 명사구	
18. N1+M, N2+AP, N3	143, 144, 145
19. N1+M, N2, N3+AP, N4	1343, 1344, 1345
20. N1+M, N2+AP, N3, N4	1433, 1434, 1435
21. N1+M, N2+AP, N3+AP, N4	1443, 1445

'1' : N+M (명사+적)

'3' : N (명사)

'4' : N+AP (명사+의)

'5' : N+P (명사+조사)

6) 색인어 선정 모듈

시스템은 두가지 기법을 이용하여 추출된 명사·명사구에 각각 그 가중치 값을 부여하여 순위를 매긴 다음 10위안에 든 명사·명사구를 최종적으로 각 초록에 해당하는 색인어로 선택하였다. 따라서 각 초록 당 두개의 색인어 리스트가 구성되었다.

2.2 자동색인 시스템 평가

본 연구에서 사용된 자동색인 시스템 평가방법은 색인작업은 언어학적 과정이라고 보는 보편적인 개념을 근거로 하여 구성되었다. 즉 구문·통계적인 자동색인 기법에 의하여 선택된 각 색인어의 검색능력(Retrieval Power)을 평가하는 것이 아니라, 구축된 시스템이 문헌의 주제를 나타내는 좋은 색인어를 어느정도 선택할 수 있는 지를 평가하였다. 이를 위하여 구현된 자동색인 시스템은 수작업 색인 시스템과 비교하였다.

수작업 색인어 리스트와의 비교는 자동적으로 추출된 색인어의 질(Quality)을 평가하는 방법으로 많이 이용되고 있으나, 수작업 색인어의 불일치성, 색인자의 실수나 편견으로 인한 부정확성 때문에 비판을 받고 있기는 하지만(O'Connor, 1964 ; Rolling, 1981) 검색할 만한 큰 규모의 데이타베이스가 존재하지 않고 있는 현실과, 검색효율 보다는 색인어의 선정 능력을 평가하는 것에 목적이 있는 본 연구에서는 수작업 색인어와의 비교 방법이 보다 적합하다고 본다.

2.2.1 수작업 색인 시스템

일반적으로 수작업 색인은 색인자의 주제 지식을 이용하여 색인자가 직접 각 문헌에 해당하는 색인어를 할당/책정하는 작업을 뜻한다(Kemp, 1988). 수작업 색인작업은 두명의 색인자(주제 전문가와 색인 전문가)에 의하여 수행되었고, 색인어의 불일치성과 색인자의 편견·실수를 가급적 최소화시키기 위하여 두명의 색인자들은 각각 독립적으로 300 개의 초록으로부터 색인어를 추출한 다음, 합의를 걸쳐 각 초록에 적합한 10여개의 색인어를 선택하였다. 이와 같은 과정을 걸쳐 구축된 수작업 색인어 리스트가 주어진 한 문헌의 주요 내용(주제)을 나타내는 색인어로 구성되었다는 가정하에 자동 색인어의 평가와 자동색인 기법의 효율성을 측정하기 위한 객관적 기준으로 이용되어진다.

2.2.2 평가 측정 방법

구문·통계적 자동 색인 시스템의 성능은 다음과 같은 4 가지의 수치적 계수에 의하여 평가되었다 : 1) 자동 색인어와 수작업 색인어의 유사성을 측정하는 유사계수(Similarity Coefficient) ; 2) 적합·부적합(relevant / nonrelevant) 색인어의 수와 누락(misssing)색인어의 수 ; 3) 모든 적합 색인어들 가운데 선택된 적합 색인어의 비율, 선정률(Assignment) ; 4) 선택된 색인어들이 얼마나 적합한 가를 알려주는 정확률(Accuracy). 수치 계수를 산출해주는 공식은 아래와 같다.

1) 유사계수 $= \dfrac{C}{A+M-C}$

2) 적합 색인어의 수
부적합 색인어의 수
누락 색인어의 수

3) 선정률 $= \dfrac{C}{M}$

〈표-2〉 실험 데이타의 통계적 특성

	긴 초록		짧은 초록		전 체		SUM
	표 제	초 록	표 제	초 록	표 제	초 록	
문헌의 수	100	100	200	200	300	300	300
어절의 수	924	28,305	1,414	18,224	2,356	46,529	48,885
평균어절의 수	9.4	283.1	7.1	91.1	7.9	155.1	162.9
최소어절의 수	5	155	3	54	3	54	
최고어절의 수	17	466	17	98	17	466	

4) 정확률 $= \frac{A}{M}$

A : 자동색인 시스템이 추출한 색인어의 수
M : 수작업으로 추출한 색인어의 수
C : 수작업으로 추출한 색인어와 일치한 자동색인어의 수

또한 단어가중 기법의 효율성과 초록길이의 영향을 평가하기 위하여 본 논문은 통계학적 기법(t-test와 one-way ANOVA in GLM)을 이용하여 두 집단간의 통계학적 차이를 검증하였다.

Ⅲ. 실험결과 및 분석

3.1 실험 데이타 및 색인 데이타의 분석

본 연구의 실험 데이타는 경영/경제학 분야의 100 개의 학위논문 초록과 200 개의 학술잡지 논문의 초록으로 구성되었다. 학위논문 초록은 긴 초록으로 구분되며, 평균 한 초록당 283 개의 어절로 구성되어진 반면, 짧은 초록 즉 학술잡지 논문 초록은 초록당 평균적으로 91개의 어절이 있는 것으로 나타났다.

표 〈2〉에서 보면, 문헌 표제의 길이(표제를 구성하는 어절의 수)는 거의 비슷하나, 통제를 받은 두 종류의 초록의 길이는 뚜렷하게 구별된다. 또한 100 개 이하의 어절로 구성된 초록만을 선택하여 이루어진 짧은 초록군의 길이는 거의 비슷하였으나, 긴 초록의 길이의 범위는 폭 넓게 나타났다 : 155 개의 어절로 구성된 짧은 초록도 있지만 466 어절를 포함한 긴 초록도 나타났다.

이와 같은 특성을 지닌 실험 데이타를 근거로 하여 본 연구는 최종적으로 세가지 색인어 리스트를 형성하였다 : WDF 색인어 리스트, IDF 색인어 리스트, 수작업 색인어 리스트. 표 〈3〉은 각 색인어 리스트의 통계적 특성을 보여주고 있다.

표 〈3〉에서 보면, 300 개의 초록에서 색인자들은 전체 3,122 개의 색인어를 추출한 반면, WDF시스템은 3,843 개, IDF시스템은 3,685 개의 색인어를 선택하여 수작업 시스템 보다는 많은 색인어를 선택하였다. 자동색인 시스템에서 한계치(cutoff point)를 '10'으로 결정을 하여 10위 까지 순위에 든 모든 색인어를 선택했으므로 WDF와 IDF시스템에서 최종적으로 선택한 색인어 수는 다르게 나타났다. WDF시스템이 평균적으로 한 초록당 가장 많은 색인어(13)을 추출한 반면, 수작업 시스템은

〈표-3〉 색인어 데이타의 통계적 특성

	WDF			IDF			MANUAL		
	긴초록	짧은초록	전 체	긴초록	짧은초록	전 체	긴초록	짧은초록	전 체
1.	100	200	300	100	200	300	100	200	300
2.	1,411	2,432	3,843	1,432	2,253	3,685	1,101	2,021	3,122
3.	14.11	12.16	12.81	14.32	11.27	12.28	11.01	10.11	10.41
4.	15	12	16	7	13	13	10	12	12
5.	7	6	6	10	5	5	7	5	5
6.	21	17	21	16	17	17	16	16	16
7.	715	967	1,682	435	794	1,229	508	827	1,335
8.	7.15	4.84	5.16	4.35	3.97	4.10	5.08	4.13	4.45
9.	696	1,465	2,161	997	1,459	2,456	593	1,194	1,787
10.	6.96	7.33	7.20	9.97	7.30	8.19	5.93	5.97	5.96

주) 1. 초록 수 ; 2. 총 색인어의 수 ; 3. 초록당 평균 색인어의 수 ; 4. 초록당 색인어 수의 범위 (range) ; 5. 최소 색인어의 수 ; 6. 최대 색인어의 수 ; 7. 선정된 총 단일어의 수 ; 8. 선정된 단일어의 평균 ; 9. 선정된 총 명사구의 수 ; 10. 선정된 명사구의 평균

평균적으로 10개의 색인어를 선택한 것으로 나타났다. 또한 세 개의 색인 시스템은 단일 명사보다는 명사구를 보다 많이 선택하였으며, 예상대로 IDF시스템이 가장 많은 명사구(67%)를 선택한 반면, WDF시스템이 명사구를 가장 적게(56%)를 선택하였다. 특히 IDF시스템은 단일 명사보다 두 배나 많은 수의 명사구를 선택하였다. 또한 긴 초록에서 보다 많은 색인어가 추출되는 현상도 발견할 수 있다. 일반적으로 색인 데이타의 통계적 특성은 색인 시스템의 특성을 반영하여 나타난 것으로 보인다.

3.2 색인 실험의 결과

본 연구에서 구축된 시스템은 내용 즉 주제를 나타내는 단어나 구절이 특정 구문론적 범주에 속하므로 구문론적 패턴을 이용하여 색인어를 선택할 수 있다는 기본 가설 아래 구현되었다. 따라서 본 시스템에서 가장 중요한 단계는 주제를 나타내는 단어/단어구의 구문패턴을 작성하고 구문패턴을 이용하여 명사·명사구를 선택하는 일이었다.

본 자동색인 시스템은 앞장에 정의 내린 명사구의 형태를 기초로 하여 구성된 58 개의 구문패턴을 이용하여 300개의 초록에서 6,824 개의 명사·명사구를 추출하였다. 이때의 명사·명사구는 구문패턴만을 이용하여 추출된 것으로 아직 조사와 불용어가 제거되지 않은 상태를 말한다. 표 〈4〉는 58 개의 구문패턴과 총 출현빈도를 보여주고 있고 표 〈5〉는 출현빈도가 높은 구문과 그 예를 제시해 주고 있다. 일반적으로 단일 명사와 두

〈표-4〉 구문 패턴과 그 출현 빈도

	형 태	구문패턴	출현빈도		형 태	구문패턴	출현빈도
1.	3	N	836	30.	1333	N-M N N N	2
2.	4	N-AP	157	31.	1334	N-M N N N-AP	1
3.	5	N-P	2550	32.	1335	N-M N N N-P	3
4.	13	N-M N	20	33.	1343	N-M N N-AP N	1
5.	14	N-M N-AP	5	34.	1344	N-M N N-AP N-AP	0
6.	15	N-M N-P	171	35.	1345	N-M N N-AP N-P	2
7.	33	N N	267	36.	1433	N-M N-AP N N	0
8.	34	N N-AP	64	37.	1434	N-M N-AP N N-AP	0
9.	35	N N-P	1265	38.	1435	N-M N-AP N N-P	2
10.	43	N-AP N	103	39.	1443	N-M N-AP N-AP N	0
11.	44	N-AP N-AP	2	40.	1445	N-M N-AP N-AP N-P	0
12.	45	N-AP N-P	488	41.	3333	N N N N	9
13.	133	N-M N N	1	42.	3334	N N N N-AP	2
14.	134	N-M N N-AP	0	43.	3335	N N N N-P	26
15.	135	N-M N N-P	41	44.	3343	N N N-AP N	4
16.	143	N-M N-AP N	1	45.	3344	N N N-AP N-AP	0
17.	144	N-M N-AP N-AP	0	46.	3345	N N N-AP N-P	33
18.	145	N-M N-AP N-P	9	47.	3433	N N-AP N N	6
19.	333	N N N	43	48.	3434	N N-AP N N-AP	0
20.	334	N N N-AP	7	49.	3435	N N-AP N N-P	46
21.	335	N N N-P	234	50.	3443	N N-AP N-AP N	0
22.	343	N N-AP N	19	51.	3445	N N-AP N-AP N-P	1
23.	344	N N-AP N-AP	0	52.	4333	N-AP N N N	3
24.	345	N N-AP N-P	200	53.	4334	N-AP N N N-AP	0
25.	433	N-AP N N	21	54.	4335	N-AP N N N-P	20
26.	434	N-AP N N-AP	0	55.	4343	N-AP N N-AP N	1
27.	435	N-AP N N-P	150	56.	4345	N-AP N N-AP N-P	2
28.	443	N-AP N-AP N	0	57.	4433	N-AP N-AP N N	0
29.	445	N-AP N-AP N-P	6	58.	4435	N-AP N-AP N N-P	0

＊1 (N-M) : 명사+적

＊3 (N) : 명사

＊4 (N-AP) : 명사+의

＊5 (N-P) : 명사+조사

〈표-5〉 출현 빈도가 높은 구문 형태와 그 예

순위	구문형태	출현빈도	예
1.	5	2550	회계-P
2.	35	1265	기업 특성-P
3.	3	836	회계
4.	45	488	생산-AP 수요
5.	33	267	고무 제품
6.	35	234	고무 제품-P
7.	345	200	이자 지급-AP 절감-P
8.	15	171	경제-M 요구-P
9.	4	157	회계-AP
10.	435	150	경제-AP 균등 모형-P
11.	43	103	평가절상-AP 효과
12.	34	64	기업 특성-AP
13.	3435	46	제조 기업-AP 회계 시스템
14.	333	43	가격 감소 효과
15.	135	41	경제-M 요구 평가-P
16.	3345	33	제조 기업 특성-AP 분석
17.	3335	26	가격 감소 효과 모형
18.	433	21	연구개발비-AP 회계 선택
19.	13	20	경제-M 요구
20.	4335	20	경제-AP 균등 모형 분석-P
others		89	
TOTAL		6824	

개의 어절로 형성된 명사구가 상당히 많이 추출되었고 세 개 또는 네 개의 단어로 이루어진 16 개의 명사구는 300 개의 초록안에서 한번도 출현되지 않았음을 보여주고 있다.

또한 표 〈6〉는 자동색인 시스템 (WDF와 IDF 시스템)과 수작업 색인 시스템에 의하여 최종적으로 추출된 명사·명사구의 구문형태와 그 구문형태에 속하는 명사·명사구의 수를 가르킨다. 세 개의 시스템에 의하여 구문형태 #14(N-AP N-AP N N)과 #21(N-M N-AP N-AP N)형태의 색인어는 한번도 추출되지 않았으나, 단일 명사(#1)와 두개의 어절로 이루어진 명사구(#2, #5, #15)는 색인어 형태로 가장 많이 추출되었다. 특히 수작업 색인어인 경우 88% 이상이 이 형태에 속한 반면 네 개의 어절로 형성된 명사구는 오직 2%로 색인자들은 특정성이 높은

〈표-6〉 색인 시스템에 나타난 구문패턴과 그 출현빈도

구문 패턴		WDF	IDF	MANUAL
1.	N	1,682	1,229	1,335
2.	N N	1,042	1,220	1,080
3.	N N N	210	243	177
4.	N N N N	33	35	30
5.	N-AP N	368	429	198
6.	N N-AP N	134	182	55
7.	N-AP N N	106	119	30
8.	N-AP N-AP N	4	5	0
9.	N N N-AP N	27	37	7
10.	N N-AP N N	35	45	12
11.	N N-AP N-AP N	2	1	1
12.	N-AP N N N	13	18	4
13.	N-AP N N-AP N	2	3	0
14.	N-AP N-AP N N	0	0	0
15.	N-M N	131	164	141
16.	N-M N N	40	40	28
17.	N-M N N N	6	5	2
18.	N-M N-AP N	7	7	3
19.	N-M N N-AP N	0	1	3
20.	N-M N-AP N N	1	2	0
21.	N-M N-AP N-AP N	0	0	0
others				16
SUM		3,843	3,685	3,122

명사구는 거의 선택하지 않는 경향을 보였다. 또한 색인자들은 시스템이 이용한 명사구의 구문패턴 중 5 개에 해당되는 명사구(#8, #13, #14, #20, #21)는 한번도 선택하지 않은 반면, 시스템에서 설정하지 않은 구문형태에 속하는 명사구, 16개를 선택하였다. 수작업 색인어에서만 나타난 구문형태는 '형용사+명사(A N-P)', '관형형+명사(PP N-P)'와 4 개의 품사 태그의 다른 조합 형태, 즉 '명사-의+명사-적+명사 (N-AP N-M N)'와 '명사+명사-적+명사(N N-M N)' 형태로 나타났는데, 이중 가장 자주 나타난 형태는 '관형형+명사(PP N ; 자동화된 기업)' 형태와 '명사-의+명사-적+명사(N-AP N-M N ; 기업의 사회적 기능)' 형태이다.

〈표-7〉 색인 시스템에 의하여 선정된 색인어의 수

한 초록당 선정된 색인어의 수	WDF		IDF		MANUAL	
	초록수	색인어수	초록수	색인어수	초록수	색인어수
5			3	15	5	25
6	5	30	7	42	6	36
7	7	49	3	21	12	84
8	8	64	8	64	30	240
9	17	153	14	126	41	369
10	32	320	50	500	70	700
11	32	352	61	671	42	462
12	33	396	20	240	40	480
13	37	481	22	286	37	481
14	36	504	14	196	12	168
15	48	720	45	675	3	45
16	22	352	52	832	2	32
17	9	153	1	17		
18	6	108				
19	3	57				
20	1	20				
21	4	84				
TOTAL	300	3,843	300	3,685	300	3,122
	avg.	12.81	avg.	12.28	avg.	10.41

다음 표 7은 각 세 개의 시스템이 각 초록에서 추출된 색인어의 수를 보여주고 있다. 앞서 말했듯이 자동색인 시스템이 수작업 색인 시스템보다 많은 색인어를 선택하였고 가장 많은 색인어를 선정한 WDF시스템의 모드(Mode)는 15인 반면, IDF시스템과 수작업 시스템의 모드는 10으로 나타났다. 또한 색인자들이 보통 한 초록에서 8 개에서 12 개의 색인어로 선정한다는 이전 연구의 결론를 지지하듯이 본 실험에서 색인 작업을 한 색인자도 한초록당 평균 10씩 선택했음을 알 수 있다.

3.3 실험결과 평가 및 분석

3.3.1 자동색인어와 수작업 색인어와의 비교

본 실험에서는 자동색인의 질을 평가하기 위하여, 수작업 색인과 정확하게 일치한 색인어의 수, 즉 일치수(consensus)와 살톤(1968)이 사용한

〈표-8〉 자동색인과 수작업색인과의 비교 : 일치수와 유사계수

		긴 초록	짧은 초록	전 체
WDF	초록수	100	200	300
	일치수	6.51	6.03	6.19
	일치수의 표준편차	2.28	2.02	2.10
	유사계수	0.3622	0.3902	0.3775
	유사계수의 표준편차	0.14	0.14	0.14
IDF	초록수	100	200	300
	일치수	5.57	5.83	5.74
	일치수의 표준편차	1.93	1.99	1.95
	유사계수	0.2954	0.3896	0.3583
	유사계수의 표준편차	0.12	0.14	0.15

〈표-9〉 유사계수를 기준으로 한 WDF와 IDF시스템의 비교

	N	Mean	Std Dev	Std Error	T	Pr〉 T
WDF	300	0.3775	0.1423	0.0082	1.6327	0.1031
IDF	300	0.3583	0.1451	0.0084		

유사계수 공식을 이용하여 자동색인어와 수작업 색인어의 유사정도를 알아보았다. 표 〈8〉은 각 자동색인 시스템에서 생선된 색인어와 수작업 색인어와의 일치수와 유사정도를, 표 〈9〉는 두 시스템이 수작업 색인과의 유사성에서 통계적으로 차이가 있는지를 보여주고 있다.

WDF시스템과 수작업 색인 시스템과의 평균 유사계수 값은 0.38이며 수작업 색인과의 평균 일치수는 6(6.19), 표준편차는 2.10으로 나타났다. 또한 IDF시스템과 수작업 색인 시스템과의 유사정도는 평균 0.36이며, 평균 일치수의 값은 5.74로 WDF시스템보다 다소 낮게 나타났다. 그러나 t-테스트를 이용하여 두 시스템이 지닌 유사계수 값을 비교해본 결과, 두 시스템의 통계적 차이가 나타나지 않았음을 발견하였다 : $t(299)=1.63$, $p > 0.05$.

다음, 각 자동색인 시스템이 어느 정도 적합·부적합 색인어와 누락 색인어를 가지고 있는 지를 분석하여 보았다. 일반적으로 색인어의 적합성(relevance)을 유용성(usefulness), 관련성(relatedness), 적합성(appropriateness) 등과 같은 어휘로 표현되며 여러 가지 개념으로 설명될 수 있으나, 본 연구에서는 수작업 색인어와 완벽하게 일치한 자동색인어를 '적합 색인어 (relevant index term)'로 간주 하였다. 따라서 적합 색인어의 수는 앞에서 계산한 일치수와 같다. (만약 부분일치 비교 방법을 이용하여 부분적으로 일치한 단어도 '적합' 하다고 간주한다면 적합 색인어수와

〈표-10〉 자동색인 시스템의 적합, 부적합, 누락 색인어의 수

		전 체	적 합	부적합	누 락
WDF	초록수	300	300	300	300
	색인어의 수	3,843	1,856	1,987	1,266
	평균 색인어의 수	12.81	6.19	6.62	4.22
	표준편차	2.91	2.10	2.68	2.07
	범위	15	14	15	11
	최소 #	6	1	0	0
	최고 #	21	14	14	10
IDF	초록수	300	300	300	300
	색인어의 수	3,685	1,722	1,963	1,400
	평균 색인어의 수	12.28	5.74	6.54	4.67
	표준편차	2.77	1.95	2.89	2.16
	범위	13	13	14	13
	최소 #	5	1	1	0
	최고 #	17	13	14	12

일치수는 같지 않을 것이다.) 부적합 색인어는 수작업 색인어 리스트와 일치하지 않는 모든 자동색인어로써 자동색인어의 수에서 적합 색인어의 수(일치수)를 빼면 쉽게 계산된다. 마지막으로 누락 색인어는 수작업 색인어 리스트에는 나타나나, 자동색인어 리스트에는 나타나지 않는 색인어로써 수작업 색인어의 수에서 일치수를 감하면 누락 색인어의 수가 계산되어진다. 이 방법은 수작업 색인어 리스트에 있는 색인어 모두가 좋은 즉 적합한 색인어이며 그 문헌에 해당한 색인어 모두를 포함하고 있다는 가정하에 계산될 수 있다.

표 〈10〉는 WDF와 IDF시스템에 의하여 선정된 색인어의 적합·부적합, 누락 색인어의 수를 보여주고 있다. 두개의 색인 시스템 모두 평균적으로 6 개의 적합 색인어, 7 개의 부적합 색인어, 4에서 5 개의 누락 색인어를 가지고 있다. 특히 WDF시스템은 IDF시스템보다 부적합 색인어를 보다 많이 포함한 반면에 (1,987 vs. 1,963 부적합 색인어), IDF시스템은 누락 색인어를 보다 많이 포함하였다 (1,400 vs. 1,266 누락 색인어). 또한 두 개의 시스템 모두 부적합·누락 색인어의 수의 범위(range)가 무척 큰 것으로 나타나(10 이상), 시스템에 의하여 추출되는 색인어의 질이 항상 일정하지 않음을 보여주고 있다.

이와 같은 비교를 기준으로하여 시스템 성능에 대한 결론을 내리기 전에 두가지 요소를 고려해야 할 것이다. 첫째로 수작업 시스템은 한 초록 당 최소 5 개, 최대 12 개 정도의 색인어를 추출한 반면, 자동색인 시스템은 대부분 한 초록에서 10 개 이상의 명사·명사구를 추출했기 때문에 모든

초록에서 부적합 색인어가 나오는 것은 당연한 일이며 많은 수의 부적합 색인어가 보인 것이다. 두번째로, '0' -'1' 값을 가진 완전일치 방법으로 자동색인어의 적합·부적합 분석을 했기 때문에 각 초록에 나타난 적합 색인어의 수는 다른 연구에 비해서 적게 나타났다. 일반적으로 색인어 평가에 부분일치 방법은 많이 쓰인다(그 예로 Weinberg(1982), Yindeemak(1989) 등의 연구를 들 수 있다). 특히 세 개 또는 네 개의 단어로 형성된 명사구를 한 개의 색인어로 처리하는 명사구 색인시스템에서의 수작업 색인과의 완전일치는 단일어 색인 시스템보다 훨씬 어렵다는 것을 고려하여 적합·부적합의 수를 이해해야할 것이다.

3.3.2 자동색인 시스템의 성능 평가

효과적인 정보 시스템을 구축하는데 중요한 역할을 담당하는 색인시스템의 결과는 색인 환경의 차이에 따라 같은 문헌을 대상으로 색인 작업을 한다하더라도 항상 다르게 나타난다. 따라서 자동색인 방법의 차이, 특성이 다른 데이타베이스의 이용은 아마도 자동색인 결과에 영향을 줄 수 있다는 가정을 세울 수 있다. 본 연구는 자동색인 시스템이 이용한 두 종류의 단어가중 기법과 초록 길이가 어떻게 색인 결과에 영향을 미치며, 그 차이가 통계적으로 유의한 지를 조사하였다.

먼저 본 연구는 단어가중 기법의 효율성을 분석하기 위해서 WDF와 IDF시스템을 비교해 보았다. 일반적으로 시스템의 효율성 분석은 재현율/정확률을 이용하나 본 연구에서는 수작업 색인을 이용하여 계산될 수 있는 선정률과 정확률을 사용하였다. 선정률은 시스템이 어느정도 적합한 색인어를 선택할 수 있는 가를, 즉 시스템이 적합 색인어를 선정해내는 능력을 말하며, 정확률은 선정된 자동색인어가 얼마나 적합한 가를 즉 적합한 색인어만을 선정해내는 능력을 표시한다. 표 〈11〉은 각 시스템의 선정률/정확률 및 그 통계적 차이를 보여주고 있다. WDF시스템의 선정률은 평균적으로 0.60이며 정확률은 0.49인 반면 IDF시스템의 선정률 및 정확률은 각각 평균 0.56과 0.48으로 나타났다.

전반적으로 IDF시스템 보다는 WDF시스템에서 높은 선정률과 정확률 값이 보였고, 특히 선정률에서 두 시스템의 차이는 현저하였다(WDF시스템이 적합 선정하는 능력면에서 월등히 좋은 결과를 보였다 : $t\ (299)=2.7926,\ p < 0.01$).

두번째로, 색인 대상의 문헌의 길이가 색인 성능에 영향을 미치는 가를 알아보았다. 우리는 색인 대상의 문헌 즉 초록의 길이가 일반적으로 색인자간 색인어 선택의 일치성 또는 두 색인 리스트의 유사성에 영향을 준다는 것을 알고 있다 : 즉 초록이 짧으면 짧을수록 색인어의 수가 적게 산출되며, 따라서 일치성, 유사성이 높게 산출된다. 따라서 자동색인 시스템의 성능도 사용된 초록의 길이가 길면 길수록 점점 낮아질 것으로 예상할 수 있다.

예상대로 짧은 초록문헌을 이용한 자동색인 시스템이 보다 높은 선정률/정확률을 얻은 것으로 나타났고, 특히 초록의 길이가 IDF시스템에 강한 영향을 주고 있음을 알 수 있다(표 〈12〉 참조). 이를 통계적으로 검증하기 위하여 GLM model에 있는 ANOVA를 이용하여 비교하였다(두개의 표본 집단의 크기가 다르므로 가중값을 주어 크기 차이에서 오는 오류를 줄이는 기법을 이용하였다.)

표 〈13〉은 짧은 초록을 이용한 두 개의 자동색인 시스템 모두 긴 초록을 이용한 시스템 보다 훨

〈표－11〉 WDF와 IDF시스템의 비교: 선정률과 정확률 면에서

	시스템	N	Mean	std Erroe	T	Pr〉T
선 정 률	WDF	300	0.5984	0.1001	2.7926	0.0054**
	IDF	300	0.5589	0.0099		
정 확 률	WDF	300	0.4902	0.0086	0.7331	0.4638
	IDF	300	0.4809	0.0092		

** α=0.01에서 통계적으로 유의함

〈표－12〉 두 종류의 초록 길이를 이용한 자동색인 시스템의 선정률과 정확률

		초록종류	N	Sum	Mean	Std
WDF	선정률	LONG	100	59.93	0.5993	0.1764
		SHORT	200	120.59	0.6029	0.1720
	정확률	LONG	100	46.97	0.4699	0.1600
		SHORT	200	100.06	0.5003	0.1426
IDF	선정률	LONG	100	50.89	0.5089	0.1629
		SHORT	200	116.77	0.5838	0.1709
	정확률	LONG	100	39.55	0.3955	0.1431
		SHORT	200	104.71	0.5235	0.1500

〈표－13〉 긴 초록과 짧은 초록의 비교

	source	df	ss	MS	F－value	F－Pro.
WDF 선정률	Model	1	0.0124	0.0124	0.41	0.5249
	Error	298	9.0935	0.0305		
	total	299				
WDF 정확률	Model	1	0.3739	0.3739	13.11	0.0003***
	Error	298	8.4973	0.0285		
	total	299				
IDF 선정률	Model	1	0.0613	0.0613	2.76	0.0978
	Error	298	6.2956	0.0222		
	total	299				
IDF 정확률	Model	1	1.0921	1.0921	49.67	0.0001***
	Error	298	6.5528	0.0220		
	total	299				

***α=0.001에서 통계적으로 유의함

씬 높은 정확률을 얻었고 그 차이는 통계적으로 유의한 것으로 나타났다(WDF시스템, F (1, 298) =13.11 P < 0.001 : IDF시스템, F (1, 298) =49.67 P < 0.001). 즉 적합한 색인만을 선정해 내는 능력은 초록의 길이의 영향을 직접적으로 받는다고 할 수 있다.

IV. 구문·통계적 자동색인시스템의 고찰

4장에서는 자동색인 시스템의 전반적인 분석을 본 논문의 연구문제와 연결시켜 고찰해 보았다.

4.1 구문 통계적 색인 기법의 전반적 성능

구문·통계적 기법을 이용한 자동색인 시스템이 과연 주제를 대표할 수 있는 색인어를 추출할 수 있는가?

본 연구는 구문·통계적 방법을 이용하여 각 문헌으로 부터 주제를 나타내는 명사·명사구를 추출시킬 수 있는 실험적 자동색인 시스템을 개발하였다. 구축된 시스템은 58 개의 구문패턴을 이용하여 6,824 개의 명사·명사구를 추출하였고 이 중 최종적으로 WDF시스템은 3,843 개, IDF시스템은 3,685 개의 색인어를 선택하였다. 두 자동색인 시스템을 평가하기 위하여 수작업 색인어 리스트가 두명의 색인자에 의하여 구성되었고 300 개의 초록에서 선택된 총 수작업 색인어의 수는 3,122개 였다.

이와 같은 실험결과를 기초로 하여 자동색인 시스템에서 추출한 색인어가 어느 정도 수작업으로 추출한 색인어와 일치하는 지를 조사해 보았다. WDF시스템인 경우, 수작업 색인어와 완벽하게 일치한 자동색인어의 수는 평균적으로 6(6.19)으로 나타났고 평균 유사계수는 0.38이었고, IDF시스템인 경우 수작업 색인어와 일치한 색인어의 수는 6(5.74)이며 평균 유사계수는 0.36으로 WDF시스템보다 다소 낮게 나타났다. 이와 같은 60%의 일치율은 국내에서 연구된 다른 자동색인 시스템 보다 상당히 낮은 것으로 다소 실망적이었다. 그러나, 연구된 대부분의 한국어 처리 자동색인 시스템은 특정 주제분야의 소규모 실험문헌을 이용한 어의 통제 시스템으로 그 결과만 가지고 서로 비교할 수는 없다고 본다.

비교적 낮은 일치율을 얻게된 원인은 두가지로 설명될 수 있다. 수작업 색인어에 비하여 자동색인어가 보편적으로 지닌 높은 특정성(specificity), 잉여성(redundancy), 한 문헌에 나타난 모든 개념을 포함하려는 완벽성(completeness)과 같은 특성이 일치율과 유사계수의 값을 낮게한 것으로 보인다. 따라서 색인어의 특정성과 잉여성을 낮추고 색인어의 불일치성을 방지하기 위한 색인 표준화 작업이 시스템에 추가된다면 일치율은 많이 향상될 것이다. 즉 낮은 일치율은 어의 통제 모듈 또는 어의 네트워크와 같은 하부 시스템의 개발이 필요하다는 것을 제시하고 있다.

일치율이 다른 연구에 비하여 낮은 또 다른 이유는 본 연구에 사용한 완전일치 비교 방법에 있다. 즉 여러 어절로 이루어진 자동색인어가 여러 어절로 이루어진 수작업 색인어와의 완전일치 확률은 상당히 낮기 때문이다. 따라서 대부분의 연구들은 부분일치 비교 방법을 사용하고 있다. 만약 본 연구에서 부분일치 비교방법이 이용된다면 일치율이 상당히 향상될 것이라고 기대할 수 있다. 그러나, 완전일치 비교 방법은 두 색인어 간의 정확한 일치정도와 시스템이 지닌 기본 성능을 제시해 준다는 장점을 가지고 있다.

비록 이와 같은 이유로 본 실험의 결과는 수작업

색인어와의 비교적 낮은 일치정도를 보여주었지만, 시스템의 기본적 가설 즉 주제를 나타내는 명사·명사구는 특정 구문패턴에 의하여 추출될 수 있다는 가설을 입증하였으며 색인자에 의하여 추출된 색인어의 구문패턴이 시스템에서 이용한 구문패턴과 거의 일치하고 있음을 보여주고 있다. 전반적으로 구문·통계적 색인기법은 비교적 만족할 만한 성능을 나타냈으며, 구축된 시스템이 더욱더 향상될 수 있는 잠재력을 가지고 있음을 보여주고 있다.

4.2 단어가중 방법의 성능

단어가중 기법(WDF와 IDF)이 본 시스템에서 구문분석만을 이용하여 추출된 명사 명사구의 중요도를 측정하여 이들로 부터 색인어로서 보다 적합한 명사 명사구를 선택할 수 있는가?

자동색인은 인간의 중재없이 자동적으로 색인어를 추출·선정하는 작업을 말하며 대부분의 자동색인 시스템은 단어의 선택뿐만 아니라 각 단어에 상대적 중요성을 나타내는 값을 할당하여 보다 적합한 색인어를 선택할 수 있는 기능, 즉 단어가중 기능을 포함하고 있다. 본 연구에서 구축된 시스템도 적합한 색인어를 추출하기 위하여 단어가중 기법, 문헌내 단어빈도 가중기법과 역문헌 빈도 가중기법을 이용하였다.

구문패턴을 이용하여 추출된 6,824 명사·명사구에서 문헌내 단어빈도 가중기법은 2,981 명사·명사구(44%)를 배제시켜 3,843을 선택시킨 반면, 역문헌 빈도 가중기법은 3,139 명사·명사구(46%)를 배제시켜 최종적으로 3,685 명사·명사구를 선택시켰다. 또한 자동색인 시스템의 효율성을 근거로하여 단어가중 기법의 성능을 다른 연구와 비교하여 조사해 보았다. 자동색인 시스템의 효율성은 일반적으로 재현율/정확률(또는 선정률/정확률)을 이용하여 평가된다. 본 연구에서 WDF와 IDF시스템은 각각 0.60과 0.56의 선정률을 보였고 각 시스템의 정확률은 각각 0.49와 0.48로 나타났다. 이와 같은 수치들은 수작업 색인어와 비교한 다른 자동색인 시스템의 연구 결과와 비교적 비슷하게 나타났다 : Medlars 시스템의 재현율은 0.577이었고 정확률은 0.504이었고 (Lancaster, 1968), SMART 시스템의 재현율과 정확률은 각각 0.55와 0.65으로 본 연구 결과보다 약간 높았다 (Salton, 1969). 최근 타이(Thai) 문헌을 이용한 자동색인 시스템을 인디막(Yindeemak, 1989)이 구현시켰는데 부분일치 비교방법을 이용한 인디막의 시스템도 다른 시스템과 유사하게 재현율 0.62, 정확률 0.54를 지닌 것으로 나타났다. 따라서 본 시스템에서 사용된 색인기법(구문분석과 단어가중 기법)이 다른 시스템에서 이용된 기법과 그 성능면에서 비슷하며, 특히 단어가중 기법이 색인어로서 적합한 명사·명사구를 잘 선택할 수 있다는 결론을 내릴 수 있다.

4.3 단어가중 기법의 비교

역문헌 빈도기법을 이용한 자동색인 시스템이 문헌내 단어빈도 기법을 이용한 시스템보다 더 좋은 결과를 가질 수 있는가?

본 연구는 두 개의 단어가중 기법의 성능을 비교하기 위하여 선정률과 정확률을 이용하여 WDF 시스템과 IDF시스템을 비교·평가하였다. 그 결과 전반적으로 IDF시스템에서 보다는 WDF시스템에서 높은 선정률과 정확률 값이 보였고, 특히

선정률에서 두 시스템의 차이는 현저하였다. 즉 WDF시스템이 IDF시스템보다 수작업 색인어와 일치한 색인어를 더 많이 추출할 수 있음을 보여주었다. 1983년 수작업 색인에 대한 광범위한 연구를 한 존스(Jones)는 '색인자는 단어의 출현빈도와 문헌의 구조에 영향을 받어 색인어를 선택하고 있다'고 발표하였다. 즉 존스의 방대한 연구는 색인자들이 문헌에 자주 나타나는 단어를 색인어로 선택하는 경향을 확인하여 주었다. 따라서 일반 색인자들이 지닌 특성, 즉 색인자들이 어느 정도 높은 출현빈도와 어느 정도의 특정성을 지닌 단어를 색인어로 선택하는 경향이 한 문헌안에서 나타난 출현 빈도만을 이용한 WDF시스템을 지지한 것으로 볼 수 있다. 왜냐하면 일반적으로 색인자는 희귀하게 나타난 특정성이 높은 단어를 색인어로 추출하지 않기 때문이다.

이와는 반대로 검색효율을 기준으로 하여 자동색인 시스템을 비교하는 여러 연구에서 보면, 문헌내 단어빈도 기법을 이용한 시스템보다는 역문헌 빈도 기법을 이용한 시스템이 더 좋은 결과를 얻은 것으로 나타났다. 이런 이유는 두개의 기법이 지닌 가설이 현저히 틀리기 때문에 나타난 것으로 본다. 즉 역문헌 빈도 기법은 주제를 나타내는 좋은 색인어를 추출하기 위한 알고리즘이 아니라 적합한 문헌이 검색되도록 하는 색인어를 추출시키는 방법으로 IDF시스템의 실제 성능을 평가하기 위해서는 검색 실험을 해야하는 것이 적당하다고 본다.

4.4 초록 길이의 영향

초록의 길이가 본 연구에서 구현한 자동색인 시스템의 효율성에 영향을 미치는가?

일반적으로 색인 시스템에서 실험대상 문헌이 길면 길수록 주제를 나타내는 용어는 더욱 더 많이 추출된다고 한다(Marcus et al, 1971). 따라서 문헌에서 선정된 색인어의 수는 문헌의 길이와 정비례적인 관계가 있으나(문헌의 길이가 길면 길수록 추출된 색인어의 수는 많아진다), 색인자간의 일치성 및 색인어간의 유사성은 반비례의 관계를 가진다(짧으면 짧을 수록 일치성 및 유사성의 값은 높아진다).

본 연구에서 나타난 초록 길이의 영향력에 관한 평가 결과를 보면 위의 보편적 이론을 지지해주고 있다. 즉 자동색인 시스템은 긴 초록에서 보다는 짧은 초록에서 높은 선정률과 정확률을 보여주었다. 이와 같은 결과는 전문용어로 간결하게 쓰여진 문헌(예, 학술잡지 논문의 초록)을 이용한 경우에는 어의 통제없이 자연어 문장에서 추출된 명사·명사구를 색인어로 선택할 수 있으나, 길고 다양한 용어로 서술적으로 쓰여진 문헌을 이용할 경우에는 보다 적합한 색인어을 선택하기 위해서 어의 통제와 같은 부수적 과정이 필요하다는 것을 제시하고 있다. 따라서 본 연구에서 구현된 자동색인 시스템은 짧고 명확한 문헌에 보다 적합한 것으로 보인다.

V. 결 론

1960년대 이후 자동색인에 관한 연구는 끊임없이 계속되고 있으며 다양한 이론 또는 기법들이 제시되어 왔다. 그러나 색인작업의 복잡성(complexity)으로 인하여 아직까지도 완벽한 자동색인 시스템은 완성되지 못하였고 랑카스타 (1991)는 인간의 중재없이 완벽하게 자동으로 색인어를 선정할 수 있는 단계는 아직 오지 않았다고 말하고 있다. 그러나 자

동색인이 지닌 경제적/실용적 유용성 때문에 완전 자동색인에 관한 실험적 연구가 계속해서 시행되어지고 있다. 본 연구 또한 완전 자동색인을 구축하기 위한 또하나의 시도로서 자연어 형태의 한국어 텍스트에 적합하며 실용적으로 응용될 수 있는 자동색인 시스템을 구현하는데 그 목적이 있다.

이에 따라 본 연구는 자연어 형태의 한국어 텍스트로 부터 주제를 대표할 수 있는 색인어(단일어 또는 구절)를 추출하는 구문·통계적 자동색인 시스템을 구현하였고 수작업 색인을 이용하여 비교·평가하였다. 구문·통계적 자동색인 시스템이란 문헌의 내용을 나타내는 명사·명사구는 특정 구문범주에 속한다는 가정 아래 자연어 형태의 한국어로 표현된 문장에서 구문적으로 적합한 명사·명사구를 추출한 다음 통계적 가중치를 이용하여 색인어로서 부적합한 명사·명사구를 제거하고 가중치가 높은 명사·명사구를 색인어로 선정하는 자동색인 시스템을 말한다. 따라서 본 연구에서 구현된 자동색인 시스템은 먼저 58 개의 구문패턴을 이용하여 300 개의 초록에서 6,823 개의 명사·명사구를 추출한 다음, 단어가중 기법을 이용하여 WDF시스템은 3,843 개의 색인어를 그리고 IDF시스템은 3,685 개의 색인어를 각각 최종적으로 선택하였다.

본 시스템에서 추출된 자연어 형태인 자동색인어는 수작업 색인어에 비하여 일반적으로 높은 특정성과 잉여성과 같은 특성을 가지게 됨에 따라 수작업 색인어와의 일치율과 유사계수의 값은 비교적 낮은 값으로 나타났다. 또한 본 시스템은 형태적 중의성을 언어학적 이론을 토대로 하여 자동적으로 해결하지 못한 점, 명사구가 제한된 구문패턴에 의해서만 추출되어 색인어로 적합한 명사구가 누락된 점, 그리고 유의어 통제를 못한 점 등 색인기법의 한계점을 가지고 있으나, 구문·통계적 자동색인 시스템의 전반적인 성능은 다른 자동색인 시스템의 성능과 비교해 볼때 거의 비슷한 것으로 밝혀짐에 따라, 본 연구에서 사용된 구문·통계적 기법이 색인어 추출을 위하여 유용하게 이용될 수 있을 것으로 본다.

참고문헌

정영미. (1987) 정보검색론. 서울 : 정음사.

정영미, 이태영. (1982) "자동색인의 통계적 기법과 한국어 문헌의 실험," 도서관학 9 : 99-118.

허미숙. (1991) 지식베이스를 이용한 자동색인시스템에 관한 연구. 연세대학교 석사학위 논문.

Bookstein, A. and Kraft, D. (1977) "Operations Research Applied to Document Indexing and Retrieval Decision," Journal of ACM. 24 : 418-428.

Bookstein, A and Swanson, D. R. (1974) "Probabilistic Models for Automatic Indexing," Journal of the American Society for Information Science. 25(5) : 313-318.

Brenner, E. H. et al. (1984) "American Petroleum Institute's Machine-aided Indexing and Searching Project," Science and Technology Libraries. 5(1) : 49-62.

Cagan, C. (1970) "A Highly Associative Document Retrieval System," Journal of the American Society for Information Science. 21(5) : 330-337.

Cleverdon, C. W., Mills. J., and Keen, E. M.

(1966) Factors Determining the Performace of Indexing Systems. Cranfield, England : College of Aeronautics.

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T., and Harshman, R. (1990) "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science. 41(6) : 391－407.

Doyle, L. B. (1961) "Semantic Road Map for Literature Searches," Journal of the ACM. 8 : 553－578.

Driscoll, J., Rajala, D., Shaffer, W., and Thomas, D. (1991) "The Operation and Performance of Artificially Intelligent Keywording System," Information Processing and Management. 27(1) : 43－54.

Fagan, J. L. (1988) Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-Syntactic Methods. Ph. D. Dissertation. Ithaca, NY : Cornell University.

Fagan, J. L. (1989) "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval," Journal of the American Society for Information Science. 40(2) : 115－132.

Harter, S. P. (1975a) "A Probabilistic Approach to Automatic Keyword Indexing Part I : On the Disribution of Specialty Words in a Technical Literature," Journal of the American Society for Information Science. 26(4) : 197－206.

Harter, S. P. (1975b) "A Probabilistic Approach to Automatic Keyword Indexing Part II : An Algorithm for Probabilistic Indexing," Journal of the American Society for Information Science. 26(5) : 280－289.

Humphrey, S. and Miller, N. E. (1987) "Knowledge-Based Indexing of the Medical Literature : The Indexing Aid Project," Journal of the American Society for Information Science. 38(3) : 184－196.

Jonak, Zdemek. (1984) "Automatic Indexing of Full Texts," Information Processing and Management. 20(5/6) : 619－627.

Jones, K. P. (1983) "How Do We Index? : A Report of Some ASLIB Informatics Group Activity," Journal of Documentation. 39(1) : 1－23.

Jones, L. P., Gassie, E. W. Jr., and Radhakrishnan, S. (1990) "INDEX : The Statistical Basis for an Automatic Conceptural Phrases-Indexing System," Journal of the American Society for Information Science. 41(2) : 87－97.

Kemp, D. A. (1988) Computer-Based Knowledge Retrieval. London : Aslib.

Lancaster, F. W. (1968) Evaluation of Operation Efficiency of Medlars. The National Library of Medicine Report.

Lancaster, F. W. (1991) Indexing and Abstracting in Theory and Practice. Champaign, IL : GSLIS, University of Illinois at Urbana-Champaign.

Lesk, M. E. (1969) "Word-Word Association in Document Retrieval System," American Documentation. 20(1) : 27038.

Lochbaum, K. E. and Streeter, L. A. (1989) "Comparing and Combining Effectiveness of Latent Semantic Indexing and the Ordinary Vector Space Model for Information Retrieval," Information Processing and Management. 25(6) : 665－676.

Luhn, H. P. (1957) "A Statistical Approach to Mechanized Encoding and Searching of Library Information," IBM Journal of Research and Development. 1(4) : 309－317.

Marcus, R. S. et al. (1971) "The User Interface for the INTREX Retrieval System," In : D. E. Walkwe, ed. Interactive Biblioqraphic Search : the User/Computer Interface. Montvale, NJ : AFIPS Press. pp. 159－201.

Martinez, C. et al. (1987) "An Expert System for Machine－aided Indexing." Journal of Chemical Information and Computer Science. 27 : 158－162.

O'Connor, J. (1961) "Some Remarks on Mechanized Indexing and Some Small Scale Empirical Results," In : Machine Indexing : Progress and Problems. pp. 266－279.

O'Conner, J. (1964) "Mechanized Indexing Methods and Their Testing," Journal of ACM. 11(4) : 437－439.

Rolling, L. (1981) "Indexing Consistency, Quality and Efficiency," Information Processing and Management. 17(2) : 69－76.

Salton, G. (1968) Automatic Information Organization and Retrieval. New York : McGraw-Hill.

Salton, G. (1969) "A Comparison Between Manual and Automatic Indexing Methods," American Documentation. 20(1) : 61－71.

Salton, G. (1985) Another Look at Automatic Text Retrieval System. Technical Report no. TR85－713. Ithaca, N. Y. : Dept. of Computer Science, Cornell University.

Salton, G. (1986) "Recent Trends in Automatic Information Retrieval," In : F. Rabitti, ed. Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval. Pisa, Italy. pp. 1－10.

Salton, G. (1989) Automatic Text Pro-cessing : The Transformation, Analysis, and Retrieval of Information by Computer. Reading, Mass. : Addison-Wesley Publishing.

Salton, G. and Lesk, M. (1968) "Computer Evaluation of Term and Text Pro-

cessing," Journal of ACM. 15(1) : 8－36. 29(4) : 351－372.

Salton, G. and Yang, C. S. (1973) "On the Specification of Term Values in Automatic Indexing," Journal of Documentation. 29(4) : 351－372

Salton, G. and Yang, C. S., and Yu, C. T. (1975) "A Theory of Term Importance in Automatic Text Analysis," Journal of the American Society for Information Science. 26(1) : 33－44.

Salton, G., Wu, H., and Yu, C. T. (1981) "The Measurement of Term Importance in Automatic Indexing," Journal of the American Society for Information Science. 32 : 175－186.

Salton, G. and Buckley, C. (1988) "The Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management. 24(5) : 513－523.

Smetacek, V. (1984) "The Semantic Analysis : A Processing Tool for Text Data Base," International Forum on Information and Documentation. 9(1) : 16－19.

Sparck Johnes, K. (1972) "A Statistical Interpretation of Term Specification and its Application in Retrieval," Journal of Documentation. 28(1) : 11－21.

Sparck Johnes, K. (1973) "Indexing Term Weighting," Information Storage and Retrieval. 9(11) : 619－633.

Van der Meulen, W. A. and Hanssen, P. J. E. (1977) "Automatic Versus Manual Indexing, " Information Processing and Management. 13 : 13－21.

Vleduts-Stokolov, N. (1987) "Concept Recognition in an Automatic Text-Processing System for the Life Science," Journal of the American Society for Information Science. 38(4) : 269－287.

Weinberg, B. H. (1981) Word Frequency and Automatic Indexing. Ph. D. Dissertation. New York, NY : Columbia University.

Yindeemak, L. (1989) Computer Porcessing with Thai Text : Keyword in Context Indexing. Ph. D. Dissertation. Bloomington, IN : Indiana University.