# 자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구

유 춘 식† · 우 선 미† · 유 철 중†† · 이 종 득†††
권 오 봉†† · 김 용 성††

## 요 약

형태소 분석(Morphological Analysis)과 같은 언어학적 처리에 의존하는 기존의 한국어 문헌에 대한 자동색인 기법들은 품사의 애매모호함이나 복합명사의 처리 등으로 인한 부담(overhead)이 크다. 또한 불용어 처리에 사용되는 불용어 리스트가 대상 문헌의 주제 분야별로 따로 구축되어야 하며 그 크기가 방대하다는 문제점이 있다.

이러한 문제점들을 해결하기 위해, 본 논문에서는 각 문헌의 텍스트에 대해 복합명사 처리나 애매모호함에 대한 엄격한 분석을 수행하지 않는 간단한 형태의 형태소 분석을 수행하여 단순명사들을 추출한다. 그런 후 이들 단순명사들을 이용하여 유한 오토마타(Finite Automata)를 구성하고, 구성된 유한 오토마타와 각 명사의 단어빈도(Term Frequency)에 의해 각 색인어 후보들의 중요도를 계산하는 자동색인 기법을 제안한다. 그 결과 품사의 애매모호함에 대한 처리나 복합명사의 처리에 따른 부담을 줄일 수 있었으며, 선정된 색인어들과 수작업으로 선정한 색인어들의 비교 실험에 의해 제안한 자동색인 기법의 성능을 검증하였다.

# A Study on Automatic Indexing System Using Natural Language Processing, Statistical Technique, Relevance Verification

Chun-Sik Yoo† · Sun-Mi Woo† · Cheol-Jung Yoo†† · Chong-deuk Lee†††
Ou-Bong Gwun†† · Yong-Sung Kim††

## ABSTRACT

Typical techniques for automatic indexing on Korean documents are having difficulty in linguistic processing such as morphological analysis because a part of speech is vague and the process of complex noun is not easy. Also, a separate stop-word list have to be developed upon each different subject and the size of a stop-word list is huge to control.

To solve these problems, this theme proposes an automatic indexing technique which performs simple morphological analysis without analysis upon complex nouns and upon process of ambiguity of a part of speech. After constructing finite automata using simple nouns extracted through a morphological analysis, importance of words in text are figured out according to the term frequency of simple nouns and finite automata. Consequently, overhead upon ambiguity of complex nouns and a part of speech is reduced. This theme investigated performance of a proposed automatic indexing technique in comparison with indexing words extracted through a manual indexing process.

* 이 논문은 1996년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.
† 준 회 원 : 전북대학교 컴퓨터과학과
†† 종신회원 : 전북대학교 컴퓨터과학과
††† 정 회 원 : 서남대학교 전자계산학과
논문접수 : 1997년 7월 15일, 심사완료 : 1998년 3월 23일

## 1. 서 론

각 정보 자료의 내용을 분석한 후, 해당 정보 자료의 특성을 표현하는 주요 개념을 추출하여 각 정보 자료를 대표하도록 한 것을 색인(Index)이라고 한다[15, 23, 25]. 색인을 이용하면 방대한 양의 정보 자료로부터 정보 이용자가 원하는 정보 자료만을 선별할 수 있어서 정보 이용자가 직접 접근해야 하거나 탐색해야 하는 정보 자료의 수를 줄일 수 있다[15, 25]. 전통적으로 각 문헌에 대한 색인을 선정하는 색인작업은 색인작업에 대한 전문지식을 갖춘 훈련된 색인자(indexer) 또는 주제 전문가(색인작업의 대상이 되는 문헌들이 다루는 분야에 대한 지식을 갖춘 사람)가 자신의 지식을 기초로 하여 문헌을 분석한 후 임의의 색인어를 부여하거나 통제어휘집을 사용하여 통제된 용어 중에서 적합한 색인어를 선택하는 방법으로 이루어졌다. 그러나 1950년대 말 이후, 방대한 양의 정보 자료의 출현, 전문지식과 충분한 색인경험을 갖춘 전문 색인자의 절대적인 부족 그리고 신속한 문헌 처리를 위한 비용의 엄청난 증가 등의 이유로 적절한 시간 내에 필요한 수의 정보 자료에 대하여 색인작업을 수행하는 것이 불가능해졌다[17]. 또한 사람이 색인작업을 수행하는 경우에 동일한 정보 자료에 대해서 색인자나 색인작업 시점 등에 따라 서로 다른 색인어를 선택하는 색인어 선정의 일관성 결여 문제가 발생하게 된다[17].

따라서 이러한 문제점들을 해결하기 위해 컴퓨터의 기호조작 능력을 이용하는 자동색인(Automatic Indexing) 기법이 출현하게 되었다[15, 16, 17, 21, 25]. 자동색인 기법은 컴퓨터가 입력된 문헌들을 분석하여 각 문헌의 주제를 대표할 수 있는 단어나 단어구를 자동으로 추출하고, 이를 해당 문헌의 색인어로 부여하는 기법이다. 이러한 기법은 문헌을 구성하는 단어들을 일정한 기준에 의해 문헌을 대표하는 주제어와 그렇지 않은 비주제어로 구분하고, 주제어로 평가된 단어들로부터 색인어를 선정하는 방법에 기초하고 있다[15, 17, 25].

본 논문에서는 한국어 문헌에 대한 가장 유효한 색인어를 자동으로 결정하기 위하여, 먼저 각 문헌의 텍스트(text)를 대상으로 형태소 분석(Morphological Analysis)을 수행하여 색인어 후보가 되는 단순명사(simple noun)를 추출한다. 다음 단계로 추출된 단순명사를 이용하여 유한 오토마타(Finite Automata)를 구성한다. 그리고 구성된 유한 오토마타에 포함된 명사들에 대하여 Sparck Jones의 역문헌빈도(Inverse Document Frequency)와 Shannon의 정보이론을 이용한 중요도 계산공식을 사용하여 색인어 후보의 중요도를 계산한다. 만약 계산된 중요도가 일정한 기준 이상의 값을 가지면 이를 각 문헌의 주제를 대표하는 색인어로 선정한다. 이러한 단계를 거쳐 문헌집단 내의 모든 문헌에 대하여 색인어와 이의 중요도가 결정되면 적합성 검증(Relevance Verification) 단계를 거쳐 최종적인 색인어와 그 중요도를 결정한다. 적합성 검증은 각 문헌과 이 문헌에 부여된 색인어 후보들을 색인자에게 제시하여 색인어로 선정되었으나 실제로는 색인어로서 가치가 없는 용어들을 삭제하는 작업이다. 그리고 마지막 단계에서 각 문헌의 색인어로 구성된 유한 오토마타를 통합하여 검색용 유한 오토마타를 생성한다. 그리고 생성된 검색용 유한 오토마타를 검색과정에 이용한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 자동색인과 관련된 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 자동색인 기법에 대해 살펴본다. 그리고 4장에서는 한국정보처리학회와 한국정보과학회의 논문지에 게재된 논문을 대상으로 적용한 실험을 통해 본 논문에서 제안한 자동색인 기법의 성능을 검증한다. 마지막으로 5장에서 결론과 향후 연구과제를 기술한다.

## 2. 관련연구

Baudin 등[3]은 멀티미디어 문헌들을 대상으로 이 문헌들이 다루는 분야에 대한 도메인 모델(domain model)과 14개의 근사 검색 휴리스틱(proximity retrieval heuristic)을 이용하여 정보 이용자의 질의어로부터 개념 색인을 획득하는 기법을 제안하고 있다. 획득된 색인어의 적합률은 약 65%, 정확률은 약 50% 정도의 성능을 나타낸다.

Tzeras 등[18]은 베이스 추론망(Bayesian Inference Network)을 사용하여 색인어를 자동적으로 생성하는 기법을 제안하였다. 이 논문에서는 수작업으로 색인어를 부여한 문헌들로부터 초기에 자동적으로 생성된 색인어 사전과 주제 분야의 각 용어들을 서술자(descriptor)로 사상(mapping)할 수 있도록 하는 규칙을 사용한다. 그러나 이 논문에서 제안한 기법은 전통적인 통계적 기법에 기초한 자동색인 기법보다 뛰어난 성능

을 보이지는 못했다.

Gordon[11]은 유전자 알고리즘(Genetic Algorithm)을 이용하여 문헌에 색인어를 부여하는 기법을 제안하였다. 이 논문에서는 문헌에 후보 키워드들을 부여한 후, 유전적 교배(crossover)와 돌연변이(mutation)를 통해 해당 문헌을 가장 잘 나타낼 수 있는 키워드들을 선택하여 색인어로 부여하는 기법을 제안하였다. 그러나 이 논문에서는 실험 결과치를 제시하지 않았다.

Blosseville 등[4]은 자연어 처리 기법과 기호 학습(Symbolic Learning) 기법을 이용하여 문헌에 색인어를 부여하고 분류하는 기법을 제안하였다. 이 논문에서는 텍스트 문헌에 대하여 형태소 분석, 구문·분석, 의미 분석까지의 완전한 자연어 분석을 수행하고, 이 결과에 대해 차별 분석(discriminant analysis)을 수행하여 색인어를 부여한다. 그리고 텍스트가 아닌 문헌에 대해서는 기호 학습을 통해 색인어를 부여하였다.

그리고 한국어 문헌에 대해 언어 분석(형태소 분석, 구문 분석)을 수행한 후, 색인어를 추출하는 연구들에는 최기선 등의 연구[19], 김민정 등의 연구[20], 김판구 등의 연구[21], 서은경의 연구[22], 이현아 등의 연구[23], 임형묵 등의 연구[24] 등이 있다. [19]의 KAIS는 모든 동사들에 대해 격틀을 사전으로 구성한 후, 어휘가 필수격이면 색인어로 추출하고 선택격이면 추출하지 않는다. 이 연구에서는 적합률이 약 80%, 부적합률은 약 50%의 성능을 보이고 있다. [20]은 형태소적 언어 정보와 간단한 구문 분석 결과를 이용하여 복합어를 추출하는 기법을 제안하고 있으나, 실험 결과치는 제시하지 않았다. [21]은 형태소 분석, 불용어 제거, 부분 구문 분석을 수행한 결과가 이 논문에서 제안하는 복합어 구성 조건에 맞는다면 복합어를 구성한다. 또한 구성된 복합어를 여러 개의 개념으로 분해하는 기법을 사용하여 색인어를 선정한다. 이 연구에서는 평균 정확률이 약 87%, 부적합률은 약 15% 정도이다. [22]는 형태소 분석과 단어 가중 기법을 이용하여 단일어와 명사구를 동시에 선택하는 구문·통계적 기법을 제안하고 있으며 재현률은 약 60%, 정확률은 약 50% 정도이다. [23]은 원문 자료에 대하여 형태소 분석을 수행한 후, 명사 추출을 위한 규칙을 적용하여 명사를 추출한다. 이 연구에서는 약 82.3%의 색인어 정확률을 보이고 있다. [24]는 구문 분석 기법과 불용어 제거 기법, 스테밍(stemming) 기법, 시소러스(thesaurus) 등을 사용하여 색인어를 추출하는데, 29.5%의 재현률과 59%의 정확률을 보이고 있다.

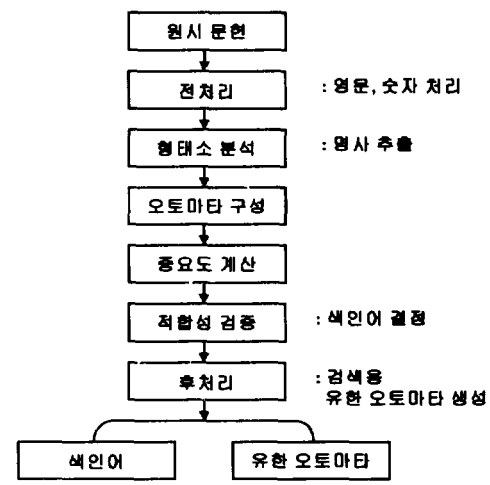
## 3. 자동색인기법

2장에서 살펴 본 바와 같이, 한국어 문헌에 대한 자동색인 기법들은 주로 형태소 분석(Morphological Analysis), 구문 분석(Syntax Analysis) 등의 언어학적 처리를 통해 명사들을 추출한다. 그리고 추출된 명사들을 불용어 리스트와 비교하여 불용어를 제거한 후, 색인어를 선정하는 과정을 거치게 된다. 그러나 이들 연구들은 언어학적 처리에 지나치게 의존하고 있기 때문에 품사를 분석하는 과정에서 애매모호함이 발생하게 되고, 아울러 복합명사(Complex Noun)를 처리하기 위한 별도의 처리 방법이 요구된다. 그리고 불용어 처리에 사용되는 불용어 리스트가 대상 문헌의 주제 분야별로 따로 구축되어야 하며, 불용어 리스트의 크기가 방대해지는 문제점이 발생하게 된다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 단순명사 추출에만 언어학적 처리를 사용하고, 색인어 선정을 위한 주요 처리들은 단순 문자열(Simple String) 처리 기법과 통계적 기법을 사용하는 자동색인 기법을 제안한다.

### 3.1 색인 시스템의 구성

본 논문의 자동색인 시스템의 구성은 (그림 1)과 같다.




(그림 1) 자동색인 시스템의 구성
(Fig. 1) Configuration of Automatic Indexing System

본 논문에서 제안하는 자동색인 기법은 복합명사에 대한 처리와 품사의 애매모호함에 대한 엄격한 분석을 수행하지 않는 간단한 형태의 형태소 분석을 수행하여 색인어 후보가 되는 단순명사들을 추출한다. 그런 후에 이들 단순명사들을 구성하는 각각의 음절들에 의해 상태가 전이되는 유한 오토마타(Finite Automata)를 구성한다. 생성된 유한 오토마타에는 각 명사들의 단어빈도(Term Frequency)와 문헌빈도(Document Frequency)가 포함되어 있다. 생성된 유한 오토마타와 단어빈도를 이용해서 색인어 후보인 각 명사(단순명사)들의 중요도를 계산하여 색인어를 선정한다. 이와 같은 방법으로 색인어의 선정이 끝나면 각 문헌의 유한 오토마타를 모두 통합하여 검색용 유한 오토마타를 생성하는 후처리 과정을 수행한다. 이 검색용 유한 오토마타는 사용자가 입력한 질의어와 각 문헌의 색인어를 비교하여 적합문헌을 찾아내는 검색과정과, 질의어로 구성된 복합명사에 대한 적합문헌을 찾아내는 재검색과정에서 이용된다.

다음 각 절에서 이러한 과정에 대해 자세히 설명한다.

### 3.2 전처리(Preprocessing)와 형태소 분석

전처리 단계에서는 문헌에 대한 형태소 분석을 수행하기 전에, 문헌의 텍스트 중에서 형태소 분석의 대상이 되지 않는 영문자, 특수 문자, 숫자, 구두점 등을 삭제한다. 이때 영문자는 기본적으로 모두 삭제하나, 논문의 표제(title)에 나타나는 영어 단어와 요약문에 나타나는 영어 단어 중에서 독립적으로 존재하는(괄호의 밖에 존재하는) 단어는 색인어일 가능성이 높으므로 색인어로 지정한다. 색인어로 지정된 영어 단어들은 적합성 검증 단계에서 색인어로서의 중요도를 판단하여 색인어로서의 중요도가 낮으면 삭제하고, 중요도가 높으면 한글 단어로 번역한다.

전처리를 거친 문헌의 텍스트에 대해 형태소 분석을 수행하여 단순명사들을 추출한다. 이때 품사의 애매모호함에 대한 상세한 처리나 복합명사에 대한 처리는 형태소 분석에서 제외한다. 그 결과 명사일 가능성이 있는 단어들이 모두 추출된다.

이러한 전처리와 형태소 분석을 통하여 명사를 추출한 예는 〈표 1〉과 같다.

### 3.3 유한 오토마타의 구성

전처리와 형태소 분석이 끝나면 각 문헌의 텍스트를

〈표 1〉 명사의 추출
〈Table 1〉 Extraction of Noun

| | |
|---|---|
| 원 문 | 〈문헌 : is21010077〉<br>표제 : 멀티미디어 지식베이스 시스템을 위한 객체 지향 지식 표현 및 규칙 그룹화<br>요약 : 본 논문에서는 데이터베이스와 지식베이스, 그리고 하이퍼미디어를 적절히 통합하는 새로운 지식 표현 모델을 제안한다. 이 모델은 지식들과 객체 지향 개념으로 통합하는 잘 정의된 구조를 가지고 있기 때문에 다양한 형태의 지식들을 효과적으로 표현할 수 있다. 또한 본 논문은 규칙 기반의 멀티미디어 지식들을 객체 지향 개념으로 그룹화하는 방법도 제시한다. |
| 추출된 명사 | 멀티미디어, 지식베이스, 시스템, 객체, 지향, 지식, 표현, 규칙, 그룹화, 논문, 데이터베이스, 하이퍼미디어, 통합, 모델, 제안, 개념, 정의, 구조, 형태, 기반, 방법, 제시 |

음절 단위로 처리하면서 유한 오토마타를 생성한다. 형태소 분석에 의해 추출된 단순명사들과 현재 처리하고 있는 단어가 서로 일치하면, 단어의 각 음절을 입력으로 하여 상태가 전이되는 유한 오토마타를 구성한다. 그런 후 단어(현재 처리하고 있는 명사)의 단어빈도를 증가시킨다. 이때 두 개 이상의 단순명사가 연결되어 복합명사를 구성하면 각각의 단순명사에 대한 유한 오토마타를 구성하고, 이를 연결하여 전체 복합명사에 대한 유한 오토마타를 구성한다. 이러한 과정으로 유한 오토마타를 구성하는 알고리즘은 〔알고리즘 1〕과 같다.

〔알고리즘 1〕 유한 오토마타의 구성
**입력** : 1) NounList : 단순명사들의 리스트
2) DocText : 전처리가 끝난 문헌의 텍스트
**출력** : 1) FA : 유한 오토마타
2) TermFreqList : 각 명사들(단순명사, 복합명사)의 단어빈도(Term Frequency)

```
construct_FA(NounList, DocText,
             FA, TermFreqList)
begin
  for( DocText내의 각 명사들 )
  begin
    CurText = 현재 처리하고 있는 DocText의 단어
    for( NounList내의 각 명사들 )
    begin
```

```
        CurNoun = 현재 처리하고 있는
                  NounList의 명사
        if( CurText == CurNoun )
        then begin
          FA에 CurNoun을 추가
          TermFreqList[CurNoun]의 값을 1증가
          if( CurNoun이 복합명사의 구성요소 )
          then begin
            이미 구성되어 있는 복합명사의
              FA 부분에 CurNoun의 FA를 연결
            TermFreqList[복합명사]의 값을 1증가
          end.
        end.
      end.
    end.
end.
```

3.2절의 〈표 1〉에 제시된 명사들을 대상으로 [알고리즘 1]을 적용하여 구성한 유한 오토마타의 일부분이 (그림 2)에 제시되어 있다.

### 3.4 색인어 선정

이 절에서는 3.3절의 알고리즘에 의해 구성된 유한 오토마타와 명사(색인어 후보)들의 단어빈도를 이용하여 각 명사의 중요도를 계산하고, 계산 결과에 의해 색인어를 선정하는 과정에 대해 기술한다. 색인어는 중요도를 계산한 후, 색인자(indexer)에 의한 적합성 검증(Relevance Verification)을 거쳐 최종적으로 결정된다.

#### 1) 색인어의 중요도 계산

본 논문에서는 Sparck Jones가 제안한 역문헌빈도(Inverse Document Frequency)[15, 25]와 Shannon의 정보이론을 이용한 중요도 계산공식[15, 25]을 사용하여 각 색인어 후보의 중요도를 계산한다. 이때, 표제에 나타나는 명사들은 색인어일 가능성이 높으므로 추가적인 가중치를 부여한다.

(1) 역문헌빈도에 의한 색인어 후보의 중요도

역문헌빈도에 의한 중요도(TFIDF)는 어떤 용어가 문헌집단에 많이 나타나면서 해당 용어가 나타나는 문헌의 수가 적을수록 색인어로서의 가치가 크다는 가설에 근거하며, 이 가설은 Sparck Jones가 수행한 실험에 의해 증명되었다[15, 25].

먼저, 색인어 후보인 명사 $i$의 역문헌빈도 $IDF_i$를 계산하는 공식은 다음과 같다.

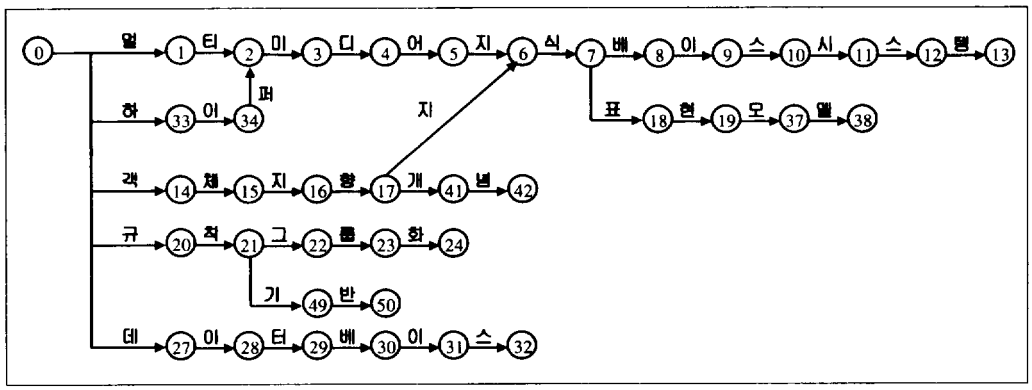
$$IDF_i = \log_2(n) - \log_2(DocFreq_i) + 1$$

$n$ : 문헌집단 내의 문헌의 수,
$DocFreq_i$ : 명사 $i$의 문헌빈도
(*Document Frequency*)

다음으로 문헌 $j$에서 명사 $i$의 중요도 $IDF_Weight_{ij}$를 계산하는 공식은 아래와 같다.

$$IDF\_Weight_{ij} = TermFreq_{ij} * IDF_i$$

$TermFreq_{ij}$ : 문헌 $j$에서의 명사 $i$의 출현빈도
(단어빈도, *Term Frequency*)




(그림 2) 구성된 유한 오토마타
(Fig. 2) Finite Automata Constructed

(2) 정보이론을 시용한 중요도

정보이론에 따르면 문헌의 색인어는 해당 문헌의 내용에 관한 불확실성을 감소시키는 효과를 가져온다. 즉, 색인어가 전달하는 정보량은 이 색인어에 의해 감소되는 불확실성의 크기만큼이고, 감소되는 불확실성의 크기가 클수록 색인어로서 가치가 크다[15, 25]. 결과적으로 문헌집단 내에서 나타나는 빈도가 높은 용어일수록 이 용어가 문헌의 내용에 대해 전달하는 정보량은 감소되고, 잡음(noise)의 크기는 커져 색인어로서의 가치가 적어진다[15, 25].

먼저, 색인어 후보인 명사 $i$가 전달하는 잡음 $Noise_i$의 크기를 계산하는 공식은 다음과 같다.

$$Noise_i = \sum_{j=1}^{n} \frac{TermFreq_{ij}}{CollectFreq_i} * \log_2 \frac{CollectFreq_i}{TermFreq_{ij}}$$

$n$ : 문헌집단 내의 문헌의 수,
$CollectFreq_i$ : 명사 $i$의 장서빈도 (*Collection Frequency*)

앞의 공식에서 알 수 있듯이, 문헌집단 내의 전체 문헌에서 명사 $i$가 고르게 분포되어 있으면 잡음이 최대가 되어 색인어로서 적합하지 않다.

다음으로 문헌 $j$에서 명사 $i$의 중요도 $Signal_Weight_{ij}$는 다음과 같은 공식에 의해 계산된다.

$$Signal\_Weight_{ij} = TermFreq_{ij} * (\log_2(CollectFreq_i) - Noise_i)$$

(3) 색인어 후보의 중요도

논문의 표제는 논문의 주제를 함축하고 있기 때문에 표제에 나타나는 명사는 색인어일 가능성이 높다. 그러므로 이러한 명사들은 앞에서 계산된 중요도에 가중치를 더하여 최종적인 중요도를 계산한다. 문헌 $j$에서 명사 $i$의 최종적인 중요도 $Weight_{ij}$는 다음과 같이 계산된다.

$$Weight_{ij} = (IDF\_Weight_{ij} + Signal\_Weight_{ij}) * Title\_Weight_i$$

$Title\_Weight_i$ : 표제에 나타나는 명사의 가중치
$\begin{cases} 2, & \text{명사 } i\text{가 표제에 나타난 경우} \\ 1, & \text{그렇지 않은 경우} \end{cases}$

이때, 색인어로서의 중요도는 각각의 단순명사들에 대해서만 계산되고, 복합명사들에 대해서는 중요도를 계산하지 않는다.

**2) 색인어 결정**

각 색인어 후보들의 중요도가 계산되면 중요도에 대한 임계치(threshold)를 적용하여 이 임계치에 미달하는 색인어 후보들은 삭제하고, 임계치를 초과하는 색인어 후보들만을 색인자에게 제시한다. 이러한 적합성 검증 단계를 거치면 최종적인 색인어가 선택된다. 만약 선택된 색인어가 복합명사를 구성하는 단순명사일 경우에는 이 단순명사를 포함하는 가장 긴 길이의 복합명사도 색인어로 선택한다. 결과적으로 단순명사와 이 단순명사를 포함하는 가장 긴 복합명사만이 색인어로 선택된다.

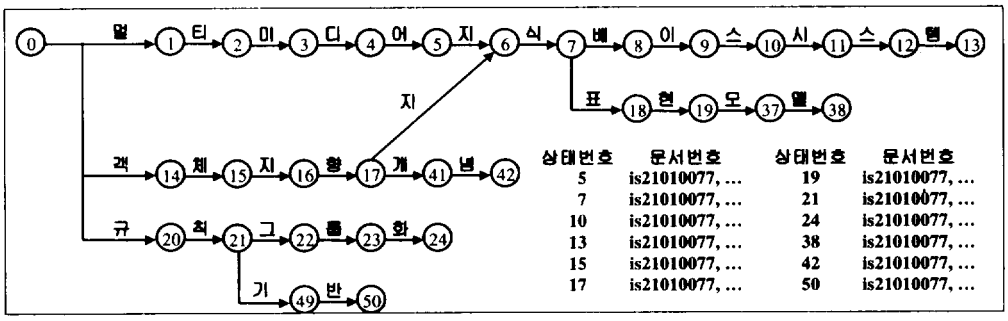
또한 색인자에 의한 적합성 검증 단계를 통해 전처리 단계에서 선정된 영어 단어들 중에서 색인어로서의 가치가 높은 영어 단어들은 한글 단어로 번역하여 색인어로 선정하고, 그렇지 않으면 삭제한다.

### 3.5 후처리

색인어의 선정이 끝난 후, 문헌 집단 내의 모든 문헌에 대하여 각 문헌별로 구성되어 있는 유한 오토마타에서 색인어로 선정된 부분만을 추출하여 이를 하나의 유한 오토마타로 통합한다. 그리고 통합된 유한 오토마타에 역파일(inverted file)의 개념을 적용하여 검색용 유한 오토마타를 생성한다. 검색용 유한 오토마타는 통합된 유한 오토마타에 존재하는 각 색인어의 끝에 해당 색인어를 포함하는 문헌들의 리스트를 연결하여 구성한다. 검색용 유한 오토마타는 검색 과정에서 이용하는데, 사용자가 입력한 질의어와 색인어를 음절 단위로 비교하게 되므로 해당 질의어가 색인어로 존재하지 않는다는 것을 보다 빠르게 알 수 있다. 또한 질의어와 색인어의 비교가 성공한 경우에는 색인어에 연결된 문헌 리스트를 이용하여 적합 문헌을 바로 검색할 수 있다.

그리고 복합 명사에 대한 구성 요소의 분할이나 결합을 통해 색인어를 결정하는 것과 같은 복잡한 과정을 거치지 않고도 검색용 유한 오토마타에서 직접 복합명사를 처리할 수 있다. 또한 복합명사를 구성하는 일부 단순명사만을 입력하여 전체 복합명사에 대한 적합문헌을 검색할 수 있다. 결과적으로 보다 빠르고 편리한 검색, 재검색 처리를 수행할 수 있다.

(그림 2)의 유한 오토마타를 이용하여 생성된 검색용 유한 오토마타의 예는 (그림 3)과 같다.




(그림 3) 생성된 검색용 유한 오토마타
(Fig. 3) Finite Automata for Retrieval Constructed

## 4. 실험 및 평가

본 논문에서 제안하는 자동색인 기법의 성능을 평가하기 위해 한국정보처리학회와 한국정보과학회의 논문지에 게재된 논문들을 대상으로 실험을 수행하였다. 색인어 후보(단순명사)는 각 논문의 표제(title)와 요약문에서 추출하였다. 실험 방법은 수작업으로 추출된 색인어와 본 논문에서 제안한 방법에 의해 추출된 색인어를 비교하여 색인어의 재현률(Recall Ratio)과 정확률(Precision Ratio)을 측정하는 방법을 사용하였다. 자동색인을 위한 프로그램은 Sun Workstation 상에서 C언어를 이용하여 구현하였으며 각 문헌은 미리 띄어쓰기와 맞춤법에 맞게 수정하였다.

먼저, 실험에 사용된 문헌(논문) 집단의 특성을 살펴보면 〈표 2〉와 같다.

〈표 2〉 실험 대상 문헌집단의 특징
〈Table 2〉 Characteristic of Document Collection for Testing

| | |
|---|---|
| 문헌의 수 | 1,000 개 |
| 문헌집단의 크기 | 897 KByte |
| 추출된 명사의 수 | 23,642 개 |

이러한 문헌집단에 대해 자동색인 기법을 적용한 결과는 〈표 3〉과 같다. 이때, 추출된 색인어의 재현률과 정확률은 다음 공식에 의해 계산하였다.

$$\text{색인어의 재현률} = \frac{\text{자동 추출된 적합 색인어의 수}}{\text{수작업으로 추출된 색인어의 수}}$$

$$\text{색인어의 정확률} = \frac{\text{자동 추출된 적합 색인어의 수}}{\text{자동 추출된 색인어의 총수}}$$

여기에서 색인어의 수는 단순명사의 수를 말하고, 만약 색인어가 복합명사이면 해당 복합명사를 구성하는 각각의 단순명사를 별개의 색인어로 간주하였다. 또한 적합 색인어는 제안한 자동색인 기법에 의해 추출된 색인어들 중에서 수작업으로 추출된 색인어와 일치하는 색인어를 말한다. 또한 〈표 3〉에서 "혼합공식"이라 함은 본 논문에서 사용한 중요도 계산공식을 사용한 경우를 말한다. 그리고 "역문헌빈도"는 역문헌빈도를 사용하여 색인어 후보의 중요도를 평가한 경우이고, "정보이론"은 정보이론을 이용하여 색인어 후보의 중요도를 평가한 경우이다.

〈표 3〉에서 알 수 있듯이 색인어 후보의 중요도를 계산할 때 역문헌빈도나 정보이론만을 이용하는 경우보다 본 논문의 중요도 계산공식을 사용한 경우가 다소 우수한 성능을 보이고 있다. 이는 본 논문에서 사용한 중요도 계산공식에 의해 색인어 선정의 정확성이 향상됨을 보여주고 있는 것이다. 이러한 결과를 언어학적 분석을 통해 색인어를 선정하는 기존의 연구들[19, 20, 21, 22, 23, 24]과 비교하여 보면 그 성능이 결코 뒤떨지지 않는다는 것을 알 수 있다.

더구나 본 논문에서 제안한 자동색인 기법은 색인어 선정 과정에 유한 오토마타를 사용함으로써 기존의 연구들에서 문제점으로 제기되고 있는 복합명사와 불용어에 대한 처리를 하지 않는다는 점과 생성된 유한 오토마타를 검색과정에서 다시 사용할 수 있다는 점에서 기존의 연구들 보다 우수하다고 할 수 있다. 물론 색인어 선정의 대상 범위를 표제와 요약문만이 아닌 문헌의 전

문(full text)까지 포함할 경우에는 정확률이 다소 감소될 수도 있다.

〈표 3〉 색인기법에 대한 성능평가 결과
〈Table 3〉 Result of Evaluation for Automatic Indexing Technique proposed

| 수작업으로 추출된 색인어의 수 | | 8,317 개 |
|---|---|---|
| 자동 추출된 색인어의 수 | 혼합공식 | 7,429 개 |
| | 역문헌빈도 | 7,334 개 |
| | 정보이론 | 7,308 개 |
| 자동 추출된 적합 색인어의 수 | 혼합공식 | 4,513 개 |
| | 역문헌빈도 | 4,392 개 |
| | 정보이론 | 4,211 개 |
| 색인어의 재현률 | 혼합공식 | 54.26% |
| | 역문헌빈도 | 52.81% |
| | 정보이론 | 50.63% |
| 색인어의 정확률 | 혼합공식 | 60.75% |
| | 역문헌빈도 | 59.89% |
| | 정보이론 | 57.62% |

## 5. 결 론

기존의 한국어 문헌에 대한 자동색인 기법들은 형태소 분석이나 구문 분석과 같은 언어학적 처리를 통해 명사들을 추출하고, 이들 중에서 불용어 리스트에 존재하지 않는 모든 명사들을 색인어로 선정한다. 그러나 이러한 기법들은 언어학적 처리에 의존하고 있기 때문에 품사를 분석하는 과정에서 애매모호함이 발생하며, 단순명사들로 구성되는 복합명사를 처리하기 위한 별도의 처리 방법이 요구된다. 또한 불용어 처리에 사용되는 불용어 리스트가 대상 문헌의 주제 분야별로 따로 구축되어야 하고, 그 크기가 방대해지는 문제점이 발생하게 된다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 단순명사의 추출에만 언어학적 처리를 사용하며, 색인어 선정을 위한 주요 처리에는 단순 문자열(Simple String) 처리 기법과 통계적 기법을 사용하는 자동색인 기법을 제안하였다. 본 논문에서 제안한 자동색인 기법은 복합명사에 대한 처리와 품사의 애매모호함에 대한 엄격한 분석을 수행하지 않는 단순한 형태의 형태소 분석을 통하여 색인어 후보가 되는 단순명사들을 추출하였다. 그리고 추출된 단순명사들을 구성하는 각 음절들에 의해 상태가 전이되는 유한 오토마타(Finite Automata)를 구성하였다. 생성된 유한 오토마타는 각 명사들의 단어빈도(Term Frequency)를 포함하고 있으며, 이 유한 오토마타와 단어빈도를 이용하여 각 명사들의 중요도를 계산하였다. 또한 색인어의 선정이 끝나면 각 문헌의 색인어에 대한 유한 오토마타를 모두 통합하고, 여기에 역과일의 개념을 적용한 검색용 유한 오토마타를 생성하였다. 그리고 이를 사용자가 입력한 질의어와 각 문헌의 색인어를 비교하여 적합문헌을 찾아내는 검색과정에 이용하였다.

결과적으로 본 논문에서 제안한 자동색인 기법을 사용하게 되면, 언어학적 처리에 의존함으로써 발생하는 품사의 애매모호함과 복합명사의 처리에 따른 부담(overhead)을 줄일 수 있어 색인작업을 보다 빠르게 수행할 수 있었다. 또한 자동색인 과정을 통해 선정된 색인어들과 수작업으로 선정한 색인어들을 비교한 실험결과, 본 논문에서 제안한 자동색인 기법이 54.26%의 재현률과 60.75%의 정확률을 보였으며 이러한 결과는 기존의 자동색인 기법들보다 우수하다고 할 수 있다.

그리고 본 논문에서 제안한 자동색인 기법은 유한 오토마타를 사용함으로써 기존의 자동색인 기법들에서 문제점으로 제기되고 있는 복합명사와 불용어에 대한 처리를 하지 않을 뿐만 아니라 생성된 유한 오토마타를 검색 과정에서 다시 사용하고 있다.

앞으로 색인과정에서 사용되는 유한 오토마타와 검색용 유한 오토마타를 구성하는 과정에서 요구되는 대량의 저장 공간을 줄이기 위한 연구와 생성된 검색용 유한 오토마타를 효과적으로 이용하는 검색기법에 대한 연구를 계속할 것이다. 더불어 논문의 표제와 요약만이 아닌 전문(full text)을 대상으로 하는 실험을 수행할 것이다.

## 참 고 문 헌

[1] Alfred V. Aho and Margaret J. Corasick, "Efficient String Matching: An Aid to Bibliographic Search," Communication of the ACM, ACM, Vol.18, No.6, pp.333-340, 1975.

[2] Junichi Aoe, Yoneo Yamamoto, and Ryosaku Shimada, "A Method for Improving String Pattern Matching Machines," Computer Algo-

rithms String Pattern Matching Strategies, IEEE Computer Society Press, pp.86-90, 1994.

[3] Catherine Baudin, Smadar Kedar, Jody Gevins Underwood, and Vinod Baya, "Question -based Acquisition of Conceptual Indices for Multimedia Design Documentation," Proc. of the 11th National Conf. on Artificial Intelligence, AAAI, pp.452-458, 1993.

[4] M. J. Blosseville, G. Hébrail, M. G. Monteil, and N. Pénot, "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques used Together," Proc. of the 15th Annual Int'l ACM/SIGIR Conf. on Research & Development in IR, ACM SIGIR, pp.51-58, 1992.

[5] Jonathan D. Cohen, "Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting," Journal of the American Society for Information Science, Vol.46, No.3, pp.162-174, 1995.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauerand, and R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, Vol.41, No.6, pp.391-407, 1990.

[7] Martin Dillon and Ann S. Gray, "FASIT: A Fully Automatic Syntatically Based Indexing System," Journal of the American Society for Information Science, Vol.34, pp.99-108, 1983.

[8] Jang-Jong Fan and Keh-Yih Su, "An Efficient Algorithm for Matching Multiple Patterns," Computer Algorithms String Pattern Matching Strategies, IEEE Computer Society Press,, pp.91-103, 1994.

[9] N. Fuhr, S. Hartmann, G. Lustig, M. Schwantner, K. Tzeras, and G. Knorz, "AIR/X- a Rule-Based Multistage Indexing System for Large Subject Fields," Proc. of the 8th National Conf. on Artificial Intelligence, AAAI, pp.789-795, 1990.

[10] Hideo Fujii and W. B. Croft, "A Comparison of Indexing Techniques for Japanese Text Retrieval," Proc. of the 16th Annual Int'l ACM/SIGIR Conf. on Research & Development in IR, ACM SIGIR, pp.237-246, 1993.

[11] M. Gordon, "Probabilistic and genetic algorithms for document retrieval," Communication of the ACM, Vol.31, pp.1208-1218, 1988.

[12] Yasushi Ogawa, Ayako Bessho, and Masako Hirose, "Simple Word Strings as Compound Keywords: An Indexing and Ranking Method for Japanese Texts," Proc. of the 16th Annual Int'l ACM/SIGIR Conf. on Research & Development in IR, ACM SIGIR, pp.227-236, 1993.

[13] Richard Osgood and Ray Bareiss, "Automated Index Generation for Constructing Large-scale Conversational Hypermedia Systems," Proc. of the 9th National Conf. on Artificial Intelligence, AAAI, pp.309-314, 1991.

[14] Yasushi Ogawa, and Masajirou Iwasaki, "A New Character-based Indexing Method using Frequency Data for Japanese Documents," Proc. of the 18th Annual Int'l ACM/SIGIR Conf. on Research & Development in IR, ACM SIGIR, pp.121-129, 1995.

[15] Gerard Salton and Michael J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill International Editions, 1987.

[16] Gerard Salton and Chris Buckley, "A Comparison between Statistically and Syntatically generated Term Phrase," Technical Report 89-1027, Cornell University, 1989.

[17] Gerard Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," Addition Wesley, 1989.

[18] Kostas Tzeras and Stephan Hartmann, "Automated Indexing Based on Bayesian Inference Networks," Proc. of the 16th Annual Int'l ACM/SIGIR Conf. on Research & Development in IR, ACM SIGIR, pp.22-34, 1993.

[19] Keysun Choi, Young S. Han, "Syntatic Analysis Based Automatic Indexing for Korean Texts," Proc. of the Korea-US Bilateral Workshop on Computers, Artificial Intelligence and Cognitive Science, pp.199-206, 1991.

[20] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법", 한국정보과학회 가을 학술발표논문집, 제 19권, 제 2호, pp.1005-1008, 1992.

[21] 김판구, 조유근, "상호 정보에 기반한 한국어 텍스트의 복합어 자동색인", 한국정보과학회 논문지, 제 21권, 제 7호, pp.1333-1340, 1994.

[22] 서은경, "구문·통계적 기법을 이용한 한국어 자동색인에 관한 연구", 정보관리학회지, 제 10권, 제 1호, pp.97-124, 1993.

[23] 이현아, 홍남희, 이종혁, 이근배, "한국어 형태소 구조규칙에 기반한 색인 시스템의 구현", 한국정보과학회 봄 학술발표논문집, 제 22권, 제 1호, pp.933-936, 1995.

[24] 임형묵, 정상철, 신동욱, 김형근, 최기선, "시소러스를 기반으로 하는 자동색인 시스템에 관한 연구", 한국정보과학회 봄 학술발표논문집, 제 21권, 제 1호, pp.173-176, 1994.

[25] 정영미, 정보검색론, 구미무역(주), 1993.

**유 춘 식**

1991년 8월 전북대학교 전산통계학과 졸업(이학사)

1994년 전북대학교 대학원 전산통계학과(이학석사)

1994년~현재 전북대학교 대학원 전산통계학과 박사과정

관심분야 : 디지털 도서관, 정보검색, 자동색인, 분산색인, 인공지능, 멀티에이전트 시스템 등

**우 선 미**

1995년 서남대학교 전자계산학과 졸업(이학사)

1997년 전북대학교 대학원 전산통계학과(이학석사)

1997년~현재 전북대학교 대학원 전산통계학과 박사과정

관심분야 : 디지털 도서관, 정보검색, 인공지능, 적응형 사용자 인터페이스 등

**유 철 중**

1982년 전북대학교 전산통계학과 졸업(이학사)

1985년 전남대학교 대학원 계산통계학과(이학석사)

1994년 전북대학교 대학원 전자계산학과(이학박사)

1982년~1985년 전북대학교 전자계산소 조교

1985년~1996년 기전여자전문대학 전자계산과 근무

1997년~현재 전북대학교 자연과학대학 컴퓨터과학과 전임강사

관심분야 : 디지털도서관, Multilmedia, Distributed Object Computing Environment, OOSE, HCI, Cognitive Science 등

**이 종 득**

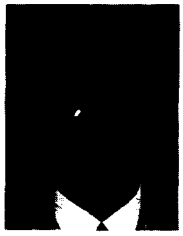
1983년 전북대학교 전산통계학과 졸업(이학사)

1989년 전북대학교 전산통계학과 (이학석사)

1998년 전북대학교 전산통계학과 (이학박사)

1992년~현재 서남대학교 전자계산학과 조교수

관심분야 : 디지털 도서관, 정보검색, 인공지능, 개념 클러스터링 등



## 권 오 봉

1980년 고려대학교 전기공학과 졸업(공학사)
1983년 고려대학교 전기공학과(공학석사)
1992년 일본 구주대학교 정보공학과(공학박사)
1992년~1993년 일본 구주대학교 정보공학과 조교수
1993년~현재 전북대학교 컴퓨터과학과 조교수
관심분야 : 디지털 도서관, 정보검색, 컴퓨터 그래픽스, 병렬처리 등

## 김 용 성

1978년 고려대학교 수학과 졸업(이학사)
1984년 광운대학교 전산학과(이학석사)
1992년 광운대학교 전산학과(이학박사)
1985년~현재 전북대학교 컴퓨터과학과 교수
1996년~1998년 1월 한국학술진흥재단 전문위원
관심분야 : 디지털 도서관, 정보검색, 인공지능, 인터넷 기반 정보검색, 멀티미디어 시스템, 등