SGML 한글문서의 논리적 구조에 근거한 색인기법에 관한 연구*

A Study of Automatic Indexing Technique based on Logical Structure of SGML Hangul Document

유석종(Seok-Jong Yu)** · 고영곤(Young-Kon Ko)** · 최윤철(Yoon-Chul Chay)**

□ 목 차 □

초 록

기존 색인 시스템은 전자문서에 대하여 전문색인(full-text indexing) 방법만을 지원하며, 문서의 논리적 구조를 검색 방법으로 적절하게 활용하지 못하고 있다. 대부분의 전자문서는 특정 시스템에 의존적인 형식으로 되어 있으며, 문서의 물리적 형태만을 나타내고 논리적 구조에 대한 정보는 포함하고 있지 않다. 이에 반해 1986년에 ISO에서 문서교환에 대한 표준방식으로 제정한 SGML(Standard Genralized Markup Language)은 문서의 논리적 구조에 대한 정보를 포함하고 있다.

본 논문에서는 기존의 전문색인 시스템의 단점을 보완하고 표준문서형식을 사용하기 위해 SGML 문서에서의 색인 시스템을 설계 구현하고자 한다. 기존 색인 시스템에서는 문서 전체에 대하여 색인이 이루어지는데 비하여 본 시스템에서는 SGML 문서의 구성요소인 엘리먼트에 기반하여 색인 영역을 지정할 수 있게 하였다. 따라서 문서의 논리적 구조를 반영한 다양한 검색기법에 응용될 수 있다. 또한 본 시스템에서는 SGML 한글문서에 대하여 자동색인이 가능하다.

ABSTRACT

Conventional indexing sytstems support only full-text indexing method for electronic documents and do not use logical structure of documents in retrieval. Most electronic documents are in different formats depending on various systems. Also, they only indicate physical style of the document without considering any logical structure. Thus, in the effort to standardize the exchange of documents, ISO developed SGML(Stadard Generalized Markup Language) which contains information about logical structure of the documents.

In this paper, to resolve the disadvantages of full-text indexing method and to use standard document format, indexing system for SGML document is designed and implemented.

In this system, user can assign indexing domain on elements, thus the logical structure of document is reflected in retrieving information. Various retrieval methods can be implemented by using the structural information of the document. In addition, automatic indexing for SGML Hangul document is supported in this system.

* 이 연구는 '92년도 한국과학재단 연구비지원에 의한 결과임(과제번호:9211-1100-012-2)

** 연세대학교 컴퓨터과학과 멀티미디어/그래픽스 연구실

■ 논문접수일 : 1995년 10월 21일

1. 서 론

1.1 연구 배경

사회가 복잡해짐에 따라 각종 정보가 기하급수적으로 증가하고 있다. 멀티미디어 시대를 맞이하여 비디오나 이미지, 사운드와 같은 새로운 형태로 표현되는 정보가 늘고 있다. 비디오나 이미지 같은 멀티미디어가 인간의 새로운 욕구를 충족시켜 주고 있지만, 많은 양의 정보를 표현하는 데 가장 효과적인 매체는 역시 전통적인 텍스트이며 앞으로도 텍스트는 가장 중요한 정보표현 방식으로 남을 것이다. 텍스트로 표현된 수많은 문서정보를 관리하고, 원하는 정보를 신속하게 얻기 위하여 개발된 정보검색시스템은 정보화사회의 필수적인 요소이다(8). 컴퓨터를 이용한 정보검색시스템은 방대한 양의 자료관리에 신속할 뿐만 아니라 일관성도 유지할 수 있다(4). 정보검색시스템에서 사용하는 많은 전자문서들은 문서의 논리적인 구조에 대한 정보를 갖고 있지 않고 모양새에 대한 정보만 갖추고 있다(6). 정보검색을 위한 색인어 추출에 있어서 기존에는 문서구조에 대한 고려없이 본문전체를 대상으로 색인하는 단순한 방법이 사용되었다. 이에 반하여 여러분야에서 표준문서방식으로 널리 사용되고 있는 SGML 문서의 색인기법은 문서의 논리적 구조를 색인과정에 반영하여 융통성 있는 정보검색을 가능하게 한다.

1.2 연구 목적

대부분의 기존 색인시스템에서 사용하는 전자문서는 본문내에서의 특정범위를 지정할 수 없으며 본문 전체에 대하여 색인과정이 수행된다. 기존 색인시스템에서 생성된 색인어화일을 통한 검색은 문서구조에 대한 고려를 할 수 없으며 다양하고 효과적인 검색기법을 적용하기 힘들다.

이에 반하여 문서의 논리적 구조를 내포하고 있는 SGML문서에서는 문서 구조에 기반하여 특정한 색인범위를 지정할 수 있으며, 또한 검색시에도 문서의 논리적 구조 정보를 이용하여 다양한 검색이 가능하다(6).

본 논문에서는 전자문서의 표준형식으로 널리 사용되고 있으며 문서 구조 정보를 포함하고 있는 SGML 문서에서 논리적 구조에 기반한 색인기법을 제시하고 효과적인 정보검색시스템에 응용될 수 있는 색인시스템을 설계 구현하고자 한다.

본 시스템에서는 기존의 문서전체를 색인하는 방법에 비하여 SGML 문서의 구성요소에 기반하여 색인범위를 지정하여 개별적이고 부분적인 색인이 가능하며 색인효율을 높일 수 있다. 또한 문서의 논리적 구조 정보를 색인어 화일에 유지함으로써 다양한 검색 방법을 적용할 수 있게 하였다. 대부분의 색인시스템은 자체 시스템에 의존적인 문서형식을 지원하기 때문에 시스템 간에 문서 교환이 힘들다는 단점이 있다. 표준문서 형식으로 대두되는 SGML 문서형식을 채용함으로써 향후 SGML 문서를 지원하는 시스템들과의 자료의 공유가 가능하다는 장점이 있다(6).

본 시스템은 특히 방대한 정보가 네트웍으로 구성되어 있고 다양한 검색도구가 필수적인 하이퍼미디어(Hypermedia) 시스템에서 검색 모듈의 하부 시스템으로 응용될 수 있다(11).

1.3 관련 연구

본 시스템은 효과적인 색인을 위하여 SGML 문서에 관련된 부분과 검색의 기초 자료인 색인어를 추출하는 부분으로 구성된다. 기존의 SGML 문서 상에서 검색하기 위하여 스트링 탐색(string search)와 같은 단순한 기법만이 주로 사용되었으며, 색인을 통한 본격적인 형태의 검색에 대한 연구는 그다지 이루어지지 않은 것으로 알려져 있다.

기존 색인시스템에서 사용하는 전자문서는 문서의 물리적 형태를 표현하기 위하여 절차적 마크업(procedural markup)을 사용한다는 것이 특징이다. 절차적 마크업은 글자체나 글자크기 등을 기술하기 위하여 텍스트에 부가되는 코드를 말한다. 절차적 마크업으로 구성된 문서는 문서의 물리적 형태 정보만을 포함하며, 해당 시스템마다 사용하는 마크업형식이 달라 문서교환이 불가능하다(2,6). 기존의 전자문서에는 보통 문서구조정보가 포함되어 있지 않기 때문에 이러한 문서형식을 사용하는 기존 시스템에서는 문서 구조를 고려한 색인은 기대할 수 없다. 이에 반하여 SGML 문서에서는 일반적 마크업(generalized markup)을 사용하여 문서의 구조 및 속성을 표현하고 SGML 문서에 포함된 문서구조정보를 색인 시에 활용한다.

- 절차적 마크업(troff로 마크업된 문서의 예)

  bp: 새로운 페이지의 시작
  ps 12: 글자크기를 12로 지정
  ft1: 글자꼴을 이탤릭으로
  This is a example of Procedural Markup

- 일반적인 마크업(SGML 실제문서부의 예)

  <title>
  Generalized Markeup Example
  </title>

영어권에서 사용하는 색인기법은 언어의 특성상 불용어만을 제거한 후에도 적절한 색인어 추출이 가능하다(8). 다음은 영문 색인 시스템의 사용하는 불용어 예이다.

- ORBIT 시스템: 8개의 불용어 사용
- van Rijsbergen(1975): 250개 불용어 사용
- Brown corpus(1982): 영문학 분야에서 1,014,000단어를 수집하여 이중 425개의 불용어를 선택(9)

본 시스템에서는 Brown corrpus에서 선택한 불용어를 영문색인에 사용하였다. 형태소 분석을 통한 불용어 제거 기법은 대개의 한글색인시스템에서 사용되어 왔던 방법으로 한글 색인에 적당하고 구현이 쉽다는 장점이 있다. 하지만 한글은 교착어적인 특성으로 인해 완벽한 형태소 분석이 어렵고, 영어의 경우보다 규칙에 예외가 많기 때문에 한글자동색인은 쉽지 않다. 또한 불용어가 문서의 주제마다 다르기 때문에 여러 분야에 공통적으로 적용 가능한 일반적인 불용어 목록을 구성하는 것이 중요하다(4).

2. SGML 개요

사회가 복잡해지면서 급증하는 수많은 정보를 다루기 위해 전자문서상에서 컴퓨터를 통한 관리가 필수적이다. 또한 정보의 공유를 위해

시스템 간의 전자문서의 상호교환 역시 중요해지고 있다. 이에 ISO 8879-1986에서는 서로 다른 시스템 간의 표준문서교환방식으로 SGML(Standard Generalized Markup Language)을 제정하였다.

2.1 SGML 이용 현황

정보구조화 언어인 SGML국제 표준은 각국의 주요 정부기관과 기업체에서 급격한 속도로 수용되고 있고, 각종 시스템 개발에 이용되고 있다(7). 다음은 SGML이 여러 분야에서 이용되고 있는 사례이다.

- 세계정보통신망인 Internet WWW(World Wide Web)상에 SGML의 Subset인 HTML(HyperText Mark-Up Language) 문서가 이용되고 있으며, 하이퍼미디어 시스템의 정보공간을 표현하는 스킴(scheme)으로 SGML과 그 응용인 HyTime(Hypermedia/Time-based Document Structuring)을 채택한 시스템 개발이 활성화되고 있다.
- 미 국방성(DOD: United States Department of Defense)에서는 1988년부터 모든 제출 자료를 SGML로 작성된 형태로만 접수하고 있으며, 미 출판협회(AAP: American Association of Publishers)에서도 표준으로 채택하고 있다.
- 최근 동양권에서도 정보 구조화 및 지식화에 관심이 높아 일본 후지쯔 등의 업체에서 SGML제품들을 개발하고 있으며, 싱가포르에서는 자국의 초고속 정보통신망에서 SGML(Standard for Document and Informantion Management, SGML Asia-Pacific, The Regent Hotel, Singapore, 10~12 October 1994)를 개최하는 등 이 분야에 대한 관심을 고조시키고 있다(7).

2.2 SGML 구성

SGML은 시스템과 응용에 독립적인 문서의 구조 정보를 위해 마크업(markup)을 사용한다. 대개의 전자문서는 내용 외에도 레이아웃(layout), 글자체 등의 정보를 필요로 한다. 이런 레이아웃이나 글자체 등과 같은 추가적인 정보를 문서 내에 삽입하는 것을 마크업이라고 한다. SGML에서는 태그(tag)와 같은 마크업을 이용하여 문서구조 정보를 나타낸다.

SGML 문서는 선언부와 문서타입정의부, 실제문서부로 구성된다(1).

2.2.1 SGML 선언부

선언부는 다른 시스템으로 문서가 전송되었을 때 호환성을 위하여 필요한 부분이다. 선언부에는 SGML문서에서 이용하는 문자집합(character set)과 이 문서에서 특수기능을 하는 문자에 대한 설명, 그리고 이 문서에서 사용되는 기능 등에 대한 정보가 포함되어 있다.

2.2.2 문서타입정의부(Document Type Definition: DTD)

문서타입정의부는 다음의 세가지 구성요소를 이용하여 작성하고자 하는 문서의 논리적 구조를 미리 정의한다. 정의 방법은 프로그래밍언어를 정의하기 위하여 BNF-Notation을 이용하는

것과 유사하다. 각 구성요소는 아래와 같다.

- 엘리먼트(Element): 문서의 논리적 단위를 나타내며, 태그를 통하여 문서의 실제 내용 부분을 마크업한다. 한 엘리먼트는 다른 엘리먼트나 실제 문서내용을 포함하는 방식으로 구성되며, 전체 문서는 이를 통하여 엘리먼트들을 노드(node)로 가지는 일종의 트리(tree)로 구성된다.
- 엔티티(Entity): 하나의 단위로서 참조되는 문자의 집합으로서의, 일종의 매크로(macro)와 유사하며 엔티티는 크게 내부 엔티티(internal entity)와 외부 엔티티(external entity)로 나뉜다. 내부 엔티티는 문서타입정의부에 참조할 내용이 선언되며, 외부 엔티티를 이용함으로써 그림화일 같은 비텍스트 부분도 문서에 포함이 가능하다.
- 속성(Attribute): 엘리먼트의 시작태그를 수식하는 일종의 매개변수(parameter)로서 엘리먼트에 부가적인 정보를 담고 있다.

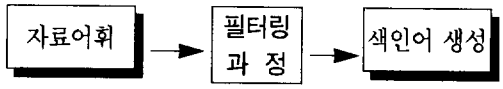
2.2.3 실제문서부(Document Instance: DI)

문서타입정의부에 정의된 문법에 맞게 작성된 실제 내용을 포함하고 있는 문서이다.

3. 색인 개요

3.1 색 인

색인은 개개의 정보 자료의 특성을 표현하는 데이타요소를 뽑아 각 정보 자료의 내용을 대표하도록 한 것으로 색인 작업을 통해 만들어진다. 색인은 정보검색 시 방대한 양의 정보로부터 이용자가 원하는 정보만을 걸러 내어 주는 여과기의 구실을 한다. 색인어는 정보검색 질의어(query language)와의 비교를 통해 정보검색의 기초 자료로 이용된다(4).



<색인과정 개요>

3.2 색인의 유형

색인의 유형에는 크게 정보 자료의 주제를 나타내는 요소를 색인어로 선택하는 주제색인과 저자명, 표제, 기관명, 출판년, 프로젝트명 등 주제와는 직접적으로 관계없는 요소를 선택하는 비주제색인방식이 있다. 다음은 주제색인 기법의 종류이다.

- 용어추출색인과 용어부여색인: 색인어를 정보 자료 자체로부터 추출하는 경우는 용어추출색인이 되고, 색인자가 정보 자료의 내용을 분석한 후 적절한 색인어를 부여하는 경우는 용어부여색인이 된다. 본 연구는 용어추출색인에 속한다.
- 자연언어색인과 통제언어색인: 정보 자료 속에 나타나 있는 형태 그대로의 용어를 색인어로 채택하는 방식으로 색인어 선택시 통제를 가하지 않는다. 따라서 시소러스(thesaurus, 유사어)나 같은 어간을 갖는 용어 또는 동음이의어가 그대로 남게 되어 검색

시효율을 떨어뜨릴 수 있다. 통제언어색인은 색인어의 선택을 돕기 위해 통제어휘집이 사용되며, 어형의 조절을 위한 규칙이 사용된다(4). 본 연구는 자연언어색인기법에 속한다.

3.3. 자동색인기법

자동색인의 궁극적인 목적은 인간이 작성한 색인과 같은 색인을 만드는 것이며, 이 작업을 컴퓨터를 사용하여 기계적으로 처리하는 것이다. 컴퓨터를 이용한 자동색인은 방대한 양의 자료에 효과적이며 신속하고 일관성 있는 색인 작업을 할 수 있다는 장점이 있다. 자동색인은 색인어를 선정하는 기준에 따라 다음과 같은 기법이 있다.

- 형태소 분석에 의한 색인: 문헌을 이루고 있는 각 문장에 대한 형태소(언어에서 의미를 가진 최소단위) 분석결과로부터 주제를 나타내는 단어나 구를 식별해내는 방법이다. 각 품사별 사전 또는 조사, 어미 사전 등을 이용하여 한 문장의 형태소를 분석한 후, 의미있는 단어를 색인어로 채택하는 기법이다.
- 구문 분석에 의한 색인: 문헌을 이루고 있는 각 문장에 대해 구두점이나 전치사, 접속사, 조사 등을 단서로 하여 문장을 문법적으로 분석하고, 전치사구나 명사구 등과 같이 특정 기능을 하는 단어나 구를 찾아낸 다음, 이 가운데 빈번하게 나타나는 단일어나 복합어를 색인어로 채택하는 기법이다.
- 통계적 방법에 의한 색인: 통계적 방법에 의한 자동색인에서는 색인어 선정을 위한 기준으로 주어진 문헌에서 특정 단어가 사용된 빈도수 정보가 사용된다. 단어의 사용 빈도수가 지나치게 높거나 낮은 단어는 주제어에서 제외된다(3,4).

3.4 한글자동색인

영문자료색인의 경우는 단순히 시스템에 미리 정의된 불용어만을 제거하고 나머지를 색인어로 채택하는 경우에도 비교적 좋은 결과를 가져오지만, 한글의 경우는 영어와 같은 굴절어와 달리 교착어로서 여러 개의 형태소가 결합하여 하나의 단어를 이루는 경우가 흔하고, 특히 용언의 활용형에 어미가 활발하게 결합하여 단어를 구성하기 때문에, 색인어로 부적합한 용언을 접속되는 어미 정보만 가지고 찾아내는 것이 쉽지 않다. 따라서 효과적인 불용어 사전 구성으로 자료 어휘의 양을 줄이는 것이 중요하다(3,8).

본 연구에서는 한글과 영문이 혼용된 자료 화일에 대하여 불용어 제거 기법을 사용하여 색인어를 추출한다.

4. SGML문서의 색인기법

4.1 정보검색 유형의 비교

컴퓨터에 의한 초기단계의 정보검색은 전문에 걸쳐 스트링 탐색(string search) 또는 스

<정보검색 유형의 비교>

트링 매칭(string matching)에 의하여 이루어졌다. 스트링 탐색은 본문에 나타나는 모든 텍스트 자료와 한 문자(character) 단위로 비교가 이루어지므로 자료의 양이 커질수록 효율이 급격히 떨어지게 된다(10).

이를 보완한 본격적인 정보검색단계에서는 문서자료의 내용을 대표하는 주제어들을 자동색인에 의해 추출하여 검색에 사용하는 전문검색시스템(full-text retrieval system)이다. 색인어를 통한 검색은 스트링탐색 단계보다는 효율적인 검색방식이지만, 문서 구조를 활용하지 못하고 다양한 검색기법의 적용이 힘들다.

본 논문의 연구 방향인 구조적 색인기법은 여러 분야에서 표준문서방식으로 사용되고 있는 SGML 문서를 통하여 문서의 논리적 구조에 기반한 색인어를 추출하고 이를 활용한 다양한 검색방법을 가능하게 한다.

4.2 기존의 색인시스템의 문제점

기존의 색인시스템은 문서구조에 대한 고려없이 문서 전체에 걸쳐 전문색인을 수행한다. 따라서 다음과 같은 문제점을 안고 있다.

- 사용자가 문서의 특정부분을 지정할수 없고 문서 전체에 대하여 색인이 이루어지므로 비효율적이다.
- 문서내에서 제목이나 본문과 같이 성격이 다른 부분에 대하여 구별없이 일률적으로 색인작업이 이루어지므로 색인어 자료에 문서의 논리적 구조정보를 유지할 수 없다. 따라서 문서의 논리적 구조를 검색에 적용할 수 없다.
- 기존 시스템에서 추출된 색인어에 의한 검색 방식은 문서 전체에 대하여 검색이 수행되므로 효과적인 검색이 어렵다.

4.3 논리적 문서구조에 기반한 색인시스템

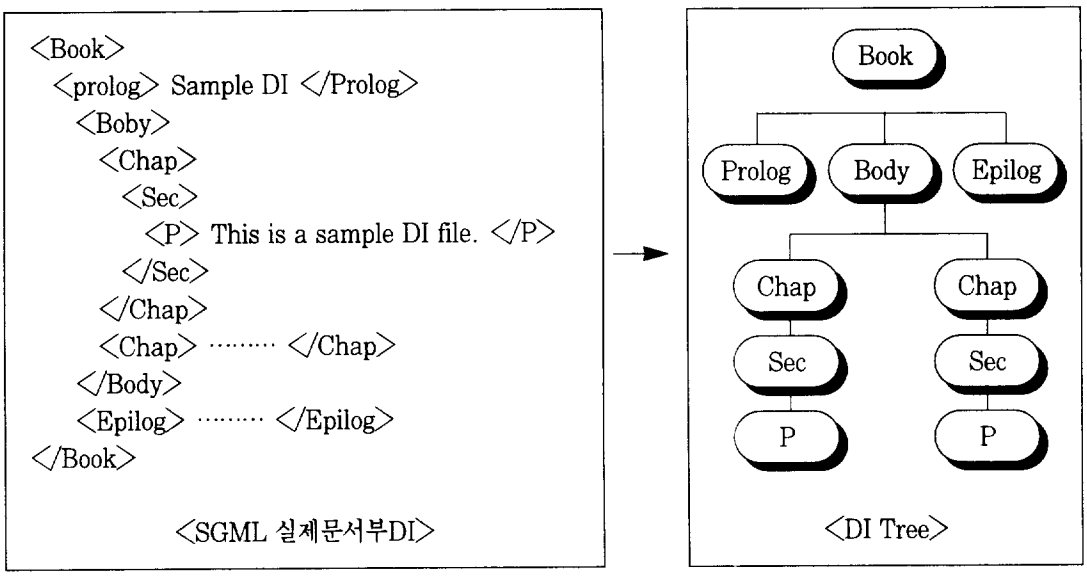
논리적인 구조정보를 가지고 있는 SGML 문서에서의 색인시스템은 문서 구성 요소에 대하여 선택적이고 부분적인 색인이 가능하다. 또한 문서구조정보가 색인어화일에 유지되므로, 단순한 전문검색에 비하여 다양한 검색이 가능하다.

다음은 문서의 논리적 구조에 기반한 색인방식의 잇점이다.

- 색인할 영역을 SGML 문서의 구성요소인 엘리먼트 단위로 지정할 수 있다. 문서의 특정 부분에 대하여 색인이 가능하므로 효율적이다.
- 문서내의 논리적인 구조, 즉 제목, 소제목, 본문, 주석 등과 같이 특성이 다른 부분에 대하여 선택적인 색인이 가능하다.
- 검색시에도 색인의 장점이 적용된다. 즉 문서의 논리적 구조에 근거한 가중치검색과 같은 다양한 검색이 가능하다(6).

4.3.1 색인과정 개요

SGML 문서에서 색인은 다음의 순서로 이루어진다. SGML 파서(parser)를 통하여 주어진 SGML 문서의 DTD 화일과 DI 화일을 파싱(parsing)을 한다. 파싱과정으로부터 얻어진 문서의 논리적 구조 정보를 이용하여 DI트리(tree)를 생성한다. DI트리를 사용하여 색인하려는 범위를 지정한 후 한글자동색인 모듈을



<그림 1> DI와 DI 트리

통하여 색인어를 추출한다. 본 시스템에서 색인영역 지정은 SGML 문서의 기본 구성요소인 엘리먼트 단위로 이루어지며, 한글자동색인 작업은 형태소 분석방법에 의한 불용어제거 기법이 사용되었다.

4.3.2 SGML 문서 파싱

SGML 파서는 SGML 문법에 맞게 DTD 화일이 작성되었는가와 SGML 선언부에 선언된 문자집합으로 구성되었는지를 확인한다. 또한 DI 화일이 DTD에 정의된 문서형식에 맞게 작성되었는지를 확인한다.

4.3.3 DI 트리 생성

DI 트리는 본 연구의 핵심자료구조로써 SGML 문서를 SGML 파서로 파싱하고, APT (Application Programming Interface)를 통하여 얻은 파싱 정보를 이용하여 구성된다.

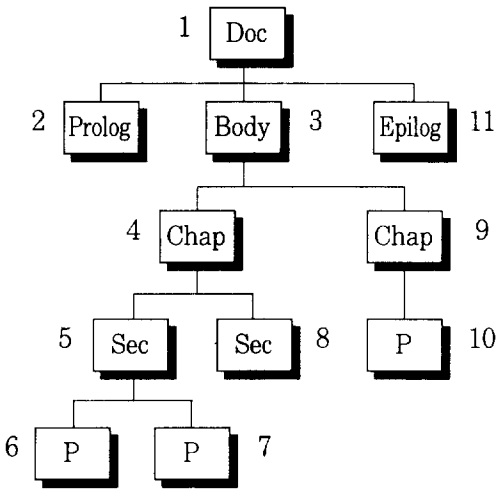
DI트리에서 하나의 노드는 SGML 문서의 엘리먼트에 대응된다. 각 노드는 해당 엘리먼트의 속성, 실제 데이타 등의 정보를 가지고 있다. <그림 1>은 DI 문서와 DI 트리가 대응되는 것을 보여주고 있다.

• DI 트리의 노드번호

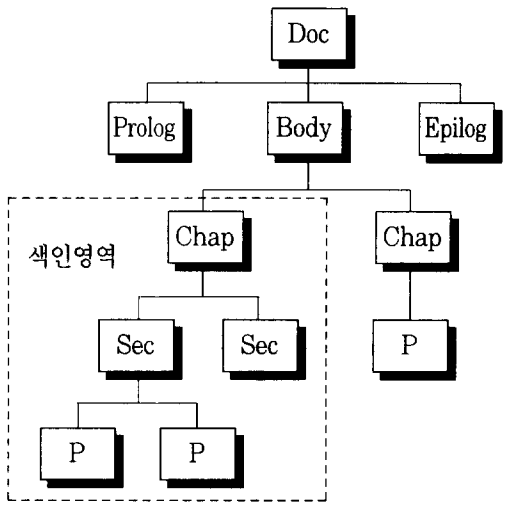
DI 트리내에서 하나의 엘리먼트를 탐색하기 위해서 각 노드의 고유 번호가 필요하다. 노드번호를 부여하기 위해서 <그림 2>와 같이 루트(root)노드로부터 깊이 우선 탐색(Depth First Search)식으로 일련번호를 부여한다.

4.3.4 색인 영역 지정

색인 영역의 지정은 SGML 문서의 구성요



<그림 2> DI 트리에서 노드 번호



<그림 3> 색인영역 지정(chap 선택시)

소인 엘리먼트 단위로 이루어진다. 색인을 원하는 엘리먼트를 선택하면 선택된 엘리먼트가 포함하는 모든 엘리먼트(sub-element)와 각 엘리먼트 내의 모든 자료들이 포함된다. 예를 들어 루트 엘리먼트를 선택하면 문서내의 모든 자료가 선택된다. 같은 이름의 엘리먼트가 문서중에 여러번 나타날수 있으므로, DI 트리에서 각 엘리먼트의 위치를 루트까지의 전체 경로(path)로 유지한다. <그림 3>에서 "Body" 노드 아래 "Chap" 노드를 선택하면 하부 엘리먼트가 모두 색인영역으로 지정되는 것을 보여주고 있다.

4.4 한글자동색인

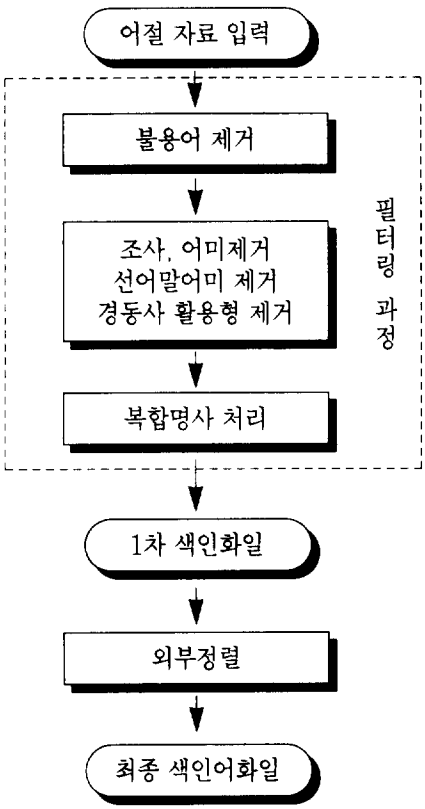
본 연구에서는 형태소 분석에 의한 불용어 제거 기법을 사용하였다. 불용어 이외의 모든 용어를 색인어로 선택하는 불용어 제거 기법은 통계적인 기법보다는 효율이 좋으나 불용어 리스트가 주제분야에 따라 다르므로 부적절한 색인어가 선택될수 있다(4). 색인과정은 <그림 4>와 같다.

- 어절 분리: 어절 분리기호를 기준으로 자료화일에서 어절을 분리한다.
- 불용어(stopword) 제거: 불용어 제거는 색인작성시 어휘자료의 양을 줄여주는 단계로 효율적인 불용어사전의 구성은 매우 중요하다. 다음에 해당하는 어절은 불용어로 간주하여 제거된다(5).

① 복자음 받침으로 시작되는 어절: 복자음 받침의 음절은 주로 한글에만 존재하고 외국어 표기에도 사용될 확률이 적다(예: 않는, 많이, 읽고 등).

② 중요도가 낮은 일반 어절: 많이 사용하지만 주제어로서 중요도가 떨어지는 명사, 대명사와 동사, 형용사 등의 용언, 부사, 감탄사 등은 제거한다(예: 설명, 시작, 천천히, 각각 등).

- 조사, 어미 제거:형용사나 동사 등의 용언에 접속된 어미와 체언에 접속된 조사를 제거한



<그림 4> 색인 과정

다. 그리고 '을'과 '는'과 같이 조사로서 사용되기도 하고 어미로서 사용되는 형태소는 제거한 후 앞의 어절이 체언이면 색인어로 등록하고, 용언이면 이 어절을 제거한다. 본 시스템의 자동색인모듈에서는 용언과 체언을 구별하지 않는다.

- 선어말어미, 경동사 활용형 제거: '었', '겠'과 같은 선어말어미가 접속되어 있으면 이를 제거하고, 선어말어미 앞에 '하', '되'와 같은 경동사나 이의 활용형이 있으면 이를 제거한다.
- 단일명사 및 복합명사 처리: 조사나 어미가 접속되지 않은 단일명사는 그 자체도 색인어로서 가치가 있으며, 그 어절 다음에 출현하는 어절과 접속되어 의미를 갖는 경우가 있다(5). 예를 들어 '정보 검색은' 또는 '정보검색은'과 같이 두 어절이 연결될 수도 있고 분리될 수도 있기 때문에 효율적인 검색을 위하여 '정보', '검색', 그리고 '정보검색' 세가지 경우를 모두 색인어로 등록시킨다. 위와 같이 연속되는 단일명사를 처리하기 위하여, 본 연구에서는 다음과 같은 알고리즘을 적용한다. 먼저 단일명사가 출현한 경우 그 다음에 나오는 어절은 다음의 4가지 경우가 있다.

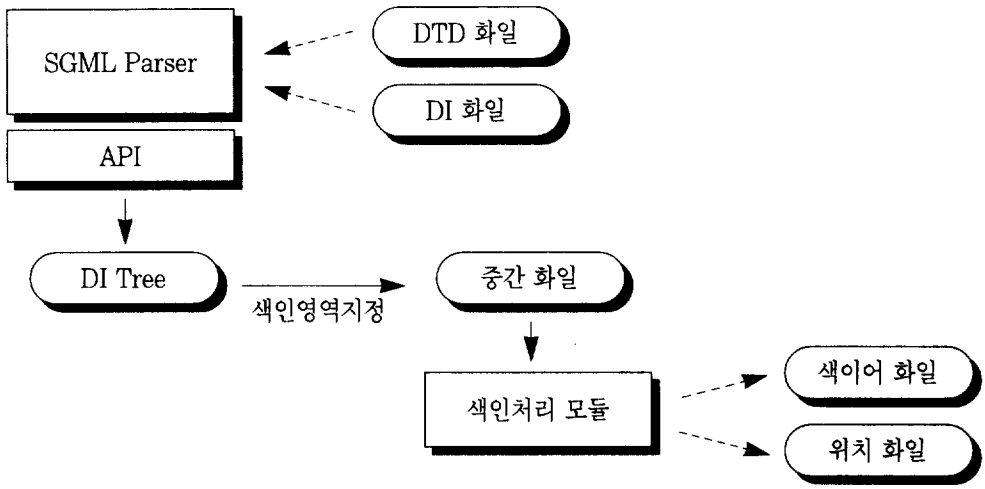
① 불용어인 경우(컴퓨터 분야는)
② 어미 접속 용언인 경우
③ 조사 접속 체언인 경우(정보 검색은)
④ 단일 명사인 경우(정보 검색 시스템은)

①,②번의 경우는 현재 단일명사만을 색인어로 등록하고 더이상 고려하지 않는다.

③번과 같이 조사가 접속된 명사가 다음에 나오는 경우는 앞의 단일 명사('정보')와 조사를 제거한 어간부분('검색')과 이들을 접속한 '정보검색' 모두를 색인어로 등록한다.

④번의 경우도 ③번과 마찬가지로 출현한 단일명사 각각과 서로 접속된 형태 모두를 등록한다(색인어: '정보', '검색', '시스템', '정보검색', '검색시스템', '정보검색시스템').

- 외부 정렬(external sort): 색인모듈에서 생성된 결과화일을 내부정렬인 퀵정렬(quick sort)와 외부정렬인 합병정렬(merge sort)를 통하여 정렬한다.
- 최종 색인어화일 및 위치화일 생성: 색인어와 색인어 위치정보를 서로 분리하여 최종색인어화일과 위치화일을 생성한다.



<그림 5> 시스템의 전반적인 처리과정

5. 색인시스템의 설계 및 구현

5.1 시스템 구성

본 시스템은 Windows 3.1 환경에서 Microsoft이 Visual C^{++} Ver 1.5.5를 사용하여 구현하였으며, SGML 파서는 기존의 ArcSGML에서 한글처리 등이 가능하도록 수정하여 사용하였다. <그림 5>은 시스템의 전반적인 처리과정이다.

<표 1>은 본 시스템에서 사용하는 사전의 종류이며, <표 2>에 본 시스템의 주요 메뉴를 설명하였다.

<표 1> 자동색인에 사용되는 사전

사전 종류	사 전 내 용
불용어 사전	한글불용어 사전
	영어불용어 사전
어미 사전	용언에 접속되는 어미 사전
조사 겸 어미 사전	조사로도 쓰이고 어미로도 사용되는 형태소
조사 사전	체언에 접속되는 조사 사전
선어말어미 사전	'겠', '었' 등의 선어말어미 사전
경동사활용어 사전	'하', '되'와 같은 경동사와 그 활용어 사전

<표 2> 메뉴기능 설명

메 뉴	메 뉴 기 능
SGML	색인범위에 포함될 엘리먼트를 지정한다.
Indexing	생성된 중간화일에 대하여 색일을 수행한다.
Search	생성된 색인어화일을 이용하여 검색한다.

5.2 색인 범위선택 대화상자

색인 범위 선택은 두가지 방식으로 지원된

다. DI 트리내에서 같은 이름이 엘리먼트가 1개 이상 출현할 수 있기 때문에, 트리내에서 특정 위치에 있는 하나의 엘리먼트를 지정하는 단일선택방식과 트리내에 출현한 특정이름의 엘리먼트 모두를 색인범위로 선택하는 복수선택방식이 있다.

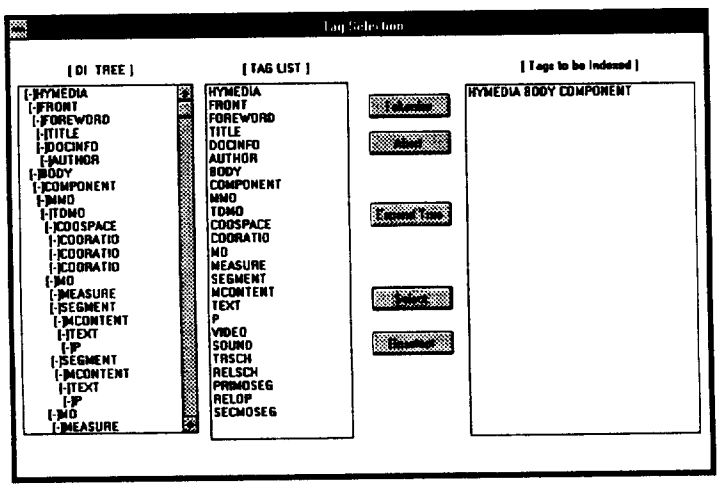
- 단일 선택방식: 색인범위선택 대화상자에 왼쪽에 있는 DI 트리 상에서 색인을 원하는 엘리먼트를 선택하면, 선택된 엘리먼트만이 색인영역지정 상자에 포함된다. 실제 색인 모듈에서는 그 하부 엘리먼트까지 포함된다.
- 복수 선택방식: 태그리스트 상자는 DI 트리에 출현하는 모든 엘리먼트가 수록되어 있다. 태그리스트 상자에서 엘리먼트를 선택하면 DI 트리상에서 선택된 엘리먼트와 같은 이름의 모든 엘리먼트들이 색인영역으로 선택된다.

<그림 6>의 가장 오른쪽에 있는 리스트 상자에는 색인작업을 수행할 SGML 문서의 DI 트리가 표시된다. 중간에 리스트 상자에는 해당 문서의 모든 엘리먼트 이름이 나타난다. 가장 왼쪽의 리스트 상자에는 색인 범위로 선택된 엘리먼트의 경로가 표시된다.

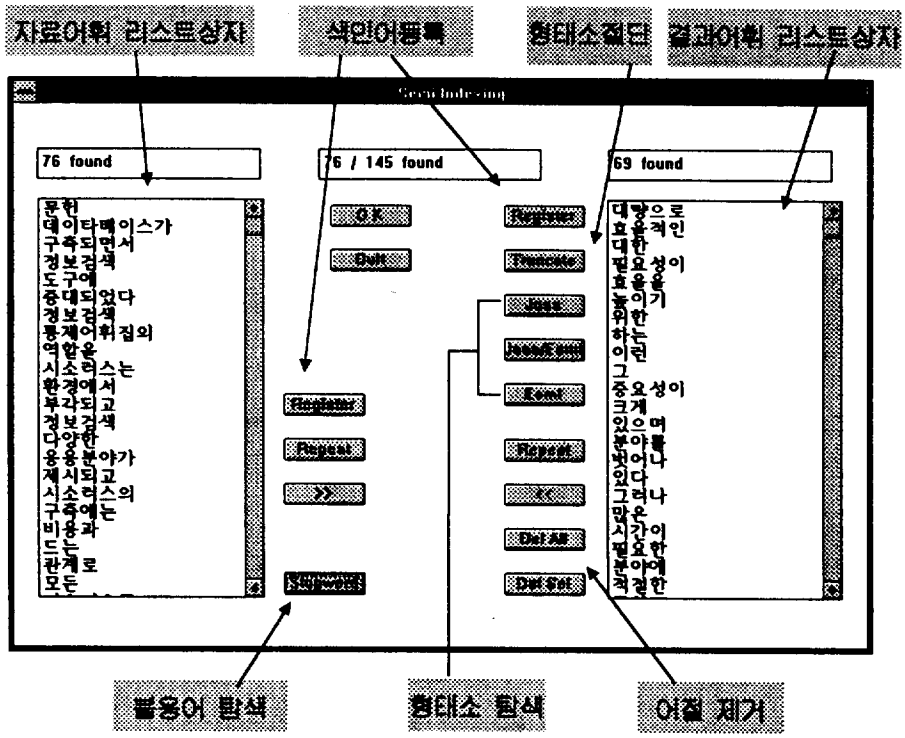
5.3 자동색인과 반자동색인

생성된 중간화일에 대하여 색인과정을 적용한다. 색인과정은 사용자의 상호작용(interaction)이 필요없는 자동색인방식(Automatic indexing)과 사용자가 색인과정의 각 단계별로 처리결과를 확인할수 있는 반자동색인방식(Semi-automatic indexing) 중에 선택할 수 있다. 색인과정이 끝나게 되면 정렬되지 않은 1차 색인화일이 생성된다.

- 자동색인: 자동색인은 자료의 양이 비교적



<그림 6> 색인범위 선택 대화상자



<그림 7> 반자동색인 수행 화면

많고 정확도가 크게 요구되지 않는 경우에 사용된다. 색인범위를 지정하여 중간화일을 생성하면 사용자와 시스템 간에 상호작용이 전혀 없이 색인어화일을 생성한다.

- 반자동색인: 반자동색인은 색인소요시간보다는 색인결과의 정확도가 중요한 경우에 사용된다. 색인과정의 각 단계별 처리결과를 사용자가 확인하여 처리결과를 사용자가 변경할수 있다. <그림 7>은 반자동색인 수행과정중 불용어처리결과이다. 왼쪽 리스트 상자의 단어들은 자료 어휘이고 오른쪽 리스트 상자는 필터링을 통하여 얻은 결과이다. 중간에 놓인 명령 버튼에 의하여 과정별로 사용자가 명령을 주고 오른쪽 리스트 상자에 나타나는 처리 결과를 확인하여 잘못된 결과를 복구할수 있다.

5.4 자료 구조

- DI 트리 자료구조

DI 트리는 실제문서부(DI)의 정보를 트리 구조로 저장하고 있다. DI 트리는 자식노드(child node)의 갯수의 제한이 없는 일반적인 트리(general tree)로 구성되어 있다. <그림 8>는 DI 트리의 각 노드들 간의 관계를 나타낸 것이다.

다음의 DI 트리의 자료구조이다.

```
class CTreeNode                    //DI 트리노드
{ public: int NodeNo;              //노드 번호
          CStrring EleName;        //엘리먼트 이름
          long DataStartPos;       //데이타의 시작위치
          long DataEtartPos;       //데이타의 끝위치
          CString TagAtt;          //속성 정보
          CTreeNode * pre;         //이전 노드에 대한 포인터
          CTreeNode * next         //이후 노드에 대한 포인터
          CTreeNode * parent;      //부모 노드에 대한 포인터
          CTreeNode * child;       //아들 노드에 대한 포인터
          int ChildNum;            //아들 노드의 갯수
          BOOL Select;             //색인 선택 여부
};
```

<DI 트리의 자료구조>

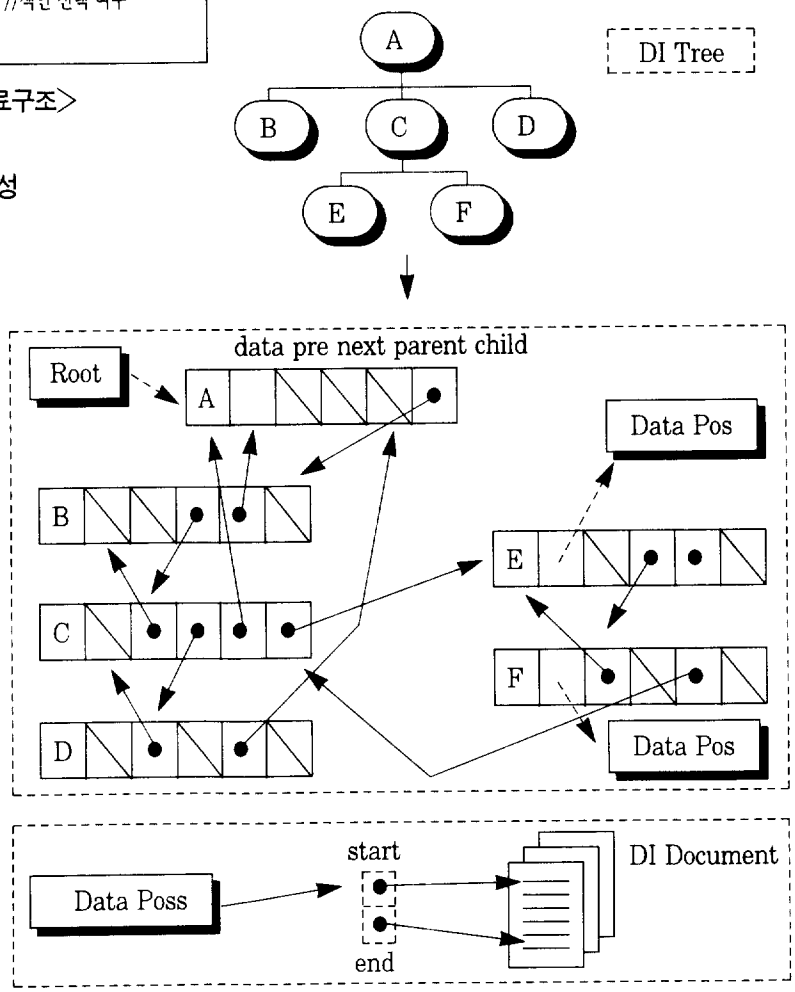
5.5 정렬 및 색인어화일 생성

5.5.1 정 렬

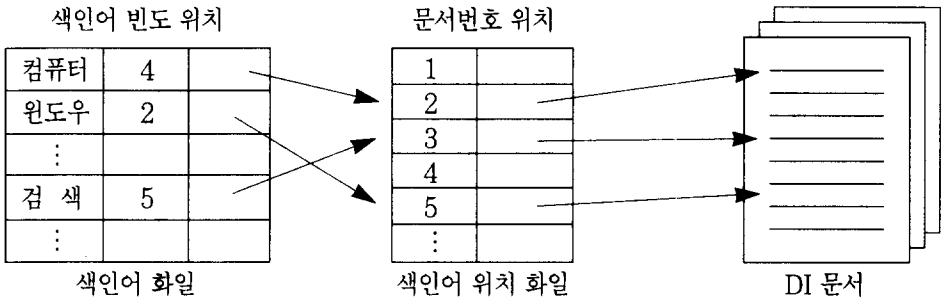
색인모듈에 의해 생성된 정렬되지 않은 1차색인어화일을 정렬하고 중복된 단어를 제외한다. 자료 어휘의 양이 메모리가 허용하는 한계보다 클 수 있으므로 정렬하려는 자료 화일을 메모리가 허용하는 크기로 나누어 각각에 대하여 내부 정렬인 퀵정렬을 사용하여 정렬하고, 각각의 정렬된 런(run) 화일을 외부 정렬인 합병정렬을 통해 하나의 화일로 만든다. 정렬을 마친 후 중복어를 제거하여 최종 색인어 화일과 색인어위치 화일을 생성한다.

5.2.2 색인어 화일 구성

최종 색인 결과는 도치색인어 화일(Inverted Index file)과 색인어 위치 화일로 <그림 9>와 같이 구성된다. 도치화일은 추출된 색인어를 정렬하여 화일탐색시 사용할 키 필드(key



<그림 8> DI 트리 노드의 연결관계



<그림 9> 색인어 화일 구성

field)로 하여 구성한 화일로 탐색이 빠르고, 불리안 검색에 적합한 화일조직이다(4,9).

6. 검색 예 및 시스템 평가

6.1 검색 예

<그림 10>은 본 시스템에서 생성된 색인어화일과 위치화일을 가지고 실제 검색을 수행하는 예이다. 색인어 화일에 대해 빠른 탐색을 위해 이진 탐색(binary search)기법을 이용하여 색인어화일에 접근한다.

문서 내에서 검색이 수행될 범위를 엘리먼트 단위로 지정함으로 검색 효율을 향상시킬 수 있고, 제목, 본문, 주석 등 문서의 구성요소에 대하여 개별적인 검색이 가능하다.

즉 제목에 해당하는 엘리먼트부분을 검색영역으로 지정함으로써 전체를 검색하여야 하는 기존의 방식에 비하여 효율적인 검색이 가능하다.

6.2 시스템 평가

본문전체를 색인하는 기존 시스템에 비하여 본 시스템에서 엘리먼트 단위로 색인영역을 지정함으로써 효과적인 색인이 가능하였다. 또한 SGML 문서의 구조를 색인 화일에 유지함으로써 검색시에도 색인과정의 장점을 그대로 적용할 수 있을 것이다.

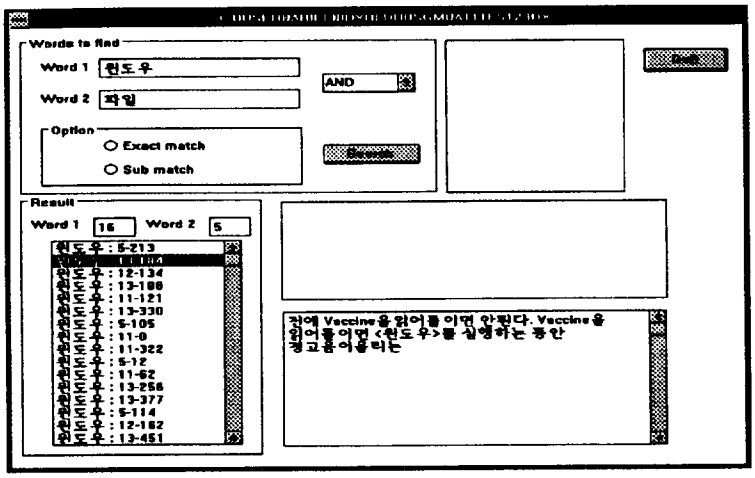
본 연구에서 문서의 구성요소 단위로 색인범위를 검색 예에서 엘리먼트 단위로 검색영역을 선택하여 검색을 수행하는 모듈은 구현하지 않았다.

아직 본 시스템은 한국어 색인에 있어서 미약하며, 큰 자료화일에 대하여 색인소요시간이 많이 걸린다는 단점이 있으며, 시스템의 성능을 향상하기 위해서는 효과적인 불용어 사전의 구성과 자료 사전과의 비교 알고리즘, 화일처리부분에서 개선이 있어야 할 것이다.

6.3 본 시스템의 응용

본 색인시스템은 문서의 논리적 구조에 기반한 색인과 검색 시의 장점을 제시하는 것으로써, 본 시스템의 자동색인모듈로부터 응용시스템으로의 API의 구축을 통해 SGML 문서형식을 지원하는 모든 검색시스템에 응용될 수 있다.

특히 하이퍼미디어 시스템에서는 방대한 정



<그림 10> 검색 예

보가 네트웍으로 구성되어 있기 때문에 사용자가 정보 공간을 편리하게 네비게이션(navigation)하기 위한 다양한 형태의 검색 메타포(metaphor)가 지원되어야 한다(11). 하이퍼미디어 시스템의 정보공간의 단위들이 SGML 문서형식으로 적절히 기술되어 있다면, 기존 하이퍼미디어 시스템에서 사용된 질의어(query language)에 의한 전문 검색(full-text retrieval) 메타포에 비해 정보공간단위별로 개별적인 검색을 통해 효율을 높일 수 있을 것이다. 예를 들어 성경(Bible)에 대한 하이퍼미디어 시스템을 구축하는 경우, 성경에 등장하는 수많은 인물들에 관한 내용을 각각 SGML 문서의 엘리먼트 단위로 대응하여 구성한다면, 특정 인물에 관련된 정보만을 부분적으로 검색하는 것도 가능할 것이다.

또한 비디오나 사운드, 이미지와 같은 멀티미디어 정보를 포함하고 있는 SGML 문서의 경우 특정 미디어에 대해서만 적용할수 있으며, 어떤 제품에 대한 소개와 사용 설명서를 다개국어(multilingual)로 지원하는 시스템에서도 각 언어별로 검색이 가능할 것이다(2).

7. 결론 및 향후 연구 방향

문서 구조에 대한 고려없이 본문 전체에 대하여 색인 작업이 이루어지는 기존의 색인 시스템에서는 선택적이고 부분적인 색인을 할 수 없고 문서의 논리적 구조를 반영한 검색이 불가능하였다.

본 색인시스템에서는 표준문서교환방식인 SGML 문서형식을 도입하여 문서의 논리적인 구조를 기반한 색인이 가능하며, 검색시에도 검색 영역을 조절하고 문서 구조를 검색에 반영할 수 있다.

SGML 문서에서는 실제문서부(DI) 안에 다른 문서들이 링크(link)로 연결되어 계속 확장이 가능하다(2). 본 시스템에서는 아직 링크되어 있는 문서 정보에 대한 고려는 없으며, 향후

연계된 연구에서는 문서 내에 링크된 문서를 포함한 색인처리가 가능해야 할 것이다.

본 연구에서 색인처리를 위해 사용한 형태소 분석을 통한 불용어 제거 기법은 한국어 색인에 적당하고 구현이 용이하다는 장점이 있는 반면 불용어가 문서의 주제마다 다르므로 여러 주제의 문서에 공통적으로 적용할 수 있는 일반적인 불용어 목록의 구성이 적절한 색인어의 추출을 좌우한다. 본 시스템 역시 자연언어처리의 한계를 극복하지 못하였으며 모든 필터링 과정을 마친 후에도 용언의 일부와 같이 색인어로 불필요한 단어가 색인어에 포함되는 단점이 있다. 본 시스템은 적절한 불용어 사전과 어휘 사전의 보강이 필요하며 한글자동색인 모듈에서 사용하는 알고리즘의 개선을 통해 색인 효율을 높여야 할 것이다.

참 고 문 헌

1. ISO 8879, *Information Processing-SGML*, 1987.
2. Martin Bryan, *An Authors guide to the Standard Generalized Markup Language*, Addison-Wesley, 1988.
4. 김영택, "자연언어처리", 교학사, 1994.
5. 정영미, 정보검색론-검색판, 구미무역(주) 출판부, 1993.
6. 이재윤, '동적 시소러스의 구축에 관한 실험적 연구', 연세대학교 석사논문, 1993.
7. 고연곤 · 이택경 · 박태진 · 최윤철, '하이퍼미디어와 정보검색', 정보과학회지, 제13권 제1호, 1995.
8. 전상훈, '색인어 생성기를 이용한 본문 검색기의 구현', 마이크로 소프트웨어 제136호, 정보시대, 1994.
9. William B. Frakes and Ricardo Baeza-Yates, *Information Retrieval(Data Structures & Algorithms)*, Prentice Hall, 1992.
10. Gerard Salton, *Automatic Text Processing*, Addison-Wesley, 1989.
11. Jakob Nielsen, *Hypertext & Hypermedia*, Academic Press, 1990.