IRS 기술을 이용한 전문(Full-text) 검색시스템

(주)한국파일링
김옥란 부장

IRS 기술을 이용한 전문(Full-text) 검색시스템

(주)한국파일링
김옥란

IRS 기술을 이용한 전문(Full-text) 검색시스템

I. 서론
II. 전문 데이타베이스 구축 방법론
III. IRS기술과 전문 검색시스템
IV. 결론

I. 서론

- 디지탈 도서관을 효율적으로 운영하기 위해서는 다양한 유형(텍스트, 이미지, 그래픽, 오디오, 비디오 등)의 매체에 수록된 정보를 수용할 수 있는 전문 데이타베이스와 이를 관리하기 위한 시스템이 필요

II. 전문 데이타베이스 구축 방법론

1. 이미지 기반 시스템
2. 아스키 기반 시스템
3. 마크업 기반 시스템

1. 이미지 기반 시스템 (Image-based System)

- 문헌들은 온라인으로 연결된 광학 디스크나 마그네틱 디스크에 전자적으로 저장
  (검색은 문헌에 미리 부여된 색인어를 통하여 이루어짐)

1. 이미지 기반 시스템 (Image-based System)

- 장점
  - 키워드 검색을 하므로 검색 속도가 빠름
  - 문헌이 전자적 파일에 저장되므로 물리적 위험으로부터 보호,
    물리적인 보관 공간 절약

1. 이미지 기반 시스템 (Image-based System)

- 단점
  - 이용자가 문헌 검색시 시스템에서 표현한 색인어와 일치하지 않으면 검색되지 않음
  - 이미지로 표현된 문헌의 삽입,삭제 등 재조작이 어렵고 응용력이 낮음

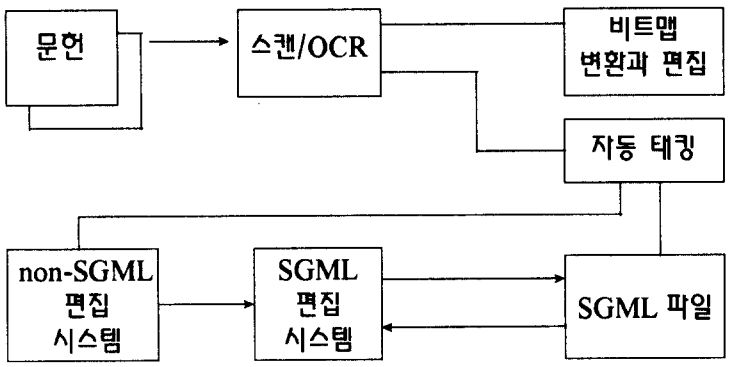
2. 아스키 기반 시스템 (ASCII-based System)

- 효율적인 전문검색을 위하여 원문을 컴퓨터가 처리할 수 있는 아스키 코드로 표현
- 장점
  - 원문에서 사용된 단어들은 사용하는 전문검색과 주제검색 가능
  - 기존의 색인보다 효율적인 색인 생성으로 검색속도가 빠름

2. 아스키 기반 시스템 (ASCII-based System)

- 장점
  - 데이터의 저장공간 절약
  - 문헌 재사용이 가능, 다양한 매체를 입력할 수 있어 응용성이 뛰어남
- 단점
  - 원문을 아스키 코드로 자동변환 시키기 어려움
  - 문헌 구조에 기반한 검색을 지원하지 않음

3. 마크업 기반 시스템 (Markup-based System)



SGML(기술적 마크업 언어의 국제표준)을 이용한 구축과정

3. 마크업 기반 시스템 (Markup-based System)

- 장점
  - 특정시스템과 하드웨어, 언어, 응용환경에 제한받지 않고 독립적
  - 다양한 유형의 정보를 다루고, 다양한 환경에서 정보를 제공하는 디지탈 도서관의 요구사항에 적합
  - 다양한 분야에서 응용가능하고, 광범위하게 유통될 수 있음 : 텍스트의 인코딩 언어로 SGML을 사용하는 경우, 아무런 장애없이 서로 교환가능

3. 마크업 기반 시스템 (Markup-based System)

- 단점
  - 마크업을 이해하기 어려움
  - 문헌을 마크업 하는데 드는 비용 부담
  - 너무 융통성이 크므로 환경에 따라 DTD (문서유형정의)의 수정이 필요

III. IRS 기술과 전문 검색시스템

1. IRS 란?
2. 전문 검색시스템의 개요
   (1) 자연어 검색시스템
   (2) 검색효율을 높이기 위한 방안
      1) 시소러스의 활용
      2) 검색기법
      3) 자동색인 시스템의 지원

1. IRS(Information Retrieval System) 란 ?

- 원론적인 정의 ; 정보 이용자가 필요로 하는 정보를 수집하여 내용을 분석한 후, 색인작업 등을 거쳐 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템

- 요즈음 전통적인 RDBMS의 여러가지 한계를 보완한 문헌정보 관리를 위한 전문 데이타베이스 검색엔진

2. 전문 검색시스템의 개요 (Full-text Retrieval System)

- 전문에 출현하는 불용어를 제외한 모든 용어에 대해 자연어 형식으로 검색할 수 있는 데이타베이스로서 신속하고 효율적인 검색을 위해 도치파일(Inverted File) 구조를 사용

(1) 자연어 검색시스템

- 장점
  - 별도의 색인작성을 위한 위한 경비 불필요
  - 새로운 정보의 접근이 용이
  - 색인 작성자의 주관에 따라 같은 주제가 여러가지 다른 용어로 표현될 수 있는 것 방지

(1) 자연어 검색시스템

- 단점

  탐색이 자유롭고 편리한 대신 정확율이 떨어질 수 있다

  - 같은 개념이 여러개의 다른 용어에 의해 표현될 가능성
  - 심층 정보 및 포괄적인 개념의 정보 검색이 어렵다

(2) 검색효율을 높이기 위한 방안

1) 시소러스의 활용

- 용어의 부정확성이나 용어의 변화 및 다양성에 대처하게 되고, 본문중의 용어와 검색자의 용어를 조정하여 검색의 정확율을 향상
- 자연어와 시소러스를 활용하여 상호보완적으로 검색할 수 있도록 지원되는 것이 가장 바람직
- 대부분의 상용 IRS 엔진에서 지원되고 있으나 관련 용어의 표준화에 효율을 기할 수 있는 표준제품에 대한 고려가 중요

2) 검색기법

- 불리안 논리 검색
  - 불리안 논리 연산을 통해 모든 조건을 만족하는 진리값 1 인 문헌만을 검색
  - 검색 메카니즘이 단순하면서도 높은 검색성능을 보여주므로 가장 널리 이용

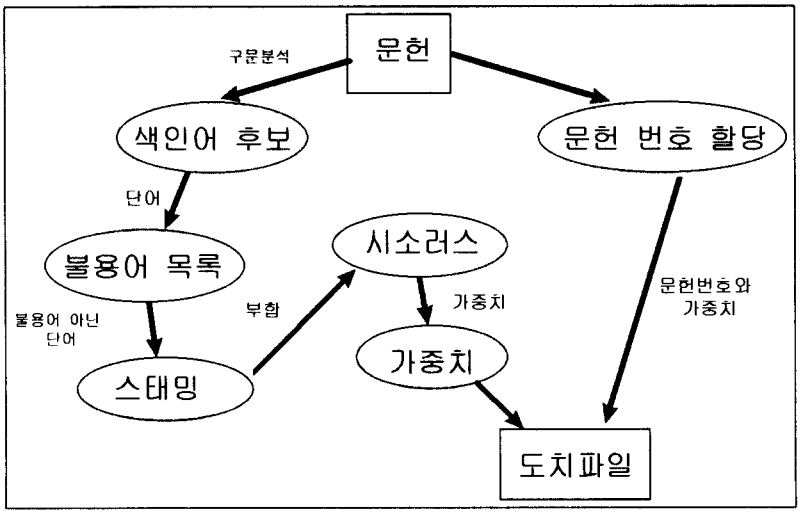
2) 검색기법

- 퍼지검색
  - 문헌의 색인어와 질문의 탐색어에 용어의 상대적 중요성을 나타내는 0 과 1 사이의 가중치를 부여함으로써 불리안 검색을 확장 또는 일반화하기 위한 검색기법

2) 검색기법

- **확률검색**
  - 어떤 문서의 주어진 질의에 대한 적절성 (Relevance)은 그 문서가 사용자에게 적절한 조건확률로서 표현되어야 한다는 것, 데이타베이스의 통계적인 특성과 표본 (Sampling)을 통해 추정

3) 자동색인 시스템의 지원



IV. 결론

- 디지탈도서관에서의 정보는 디지탈로 인코딩된 텍스트, 오디오, 비디오, 그래픽 등 다양한 유형의 정보가 통합된 멀티미디어 전문 데이타베이스에 기록된다.
- 따라서 디지탈도서관을 효율적으로 운영하기 위해서는
  - 첫째, 다양한 형식으로 표현된 정보를 수용할 수 있는 전문 데이타베이스와 이를 관리하기 위한 시스템이 필요

IV. 결론

- 둘째, 시간과 장소를 초월하여 이용자가 원하는 정보를 가장 효과적으로 검색할 수 있게 하는 이용자 우호적인 인터페이스
- 셋째, 기존 시스템과의 연동(Interoperability)
- 넷째, 서로 이질적인 특성을 갖는 시스템 상호간의 원활한 정보 유통체제 구축
  : SGML 및 Z39.50을 지원하는 검색엔진이 필요

IRS(Information Retrieval System) 기술을 이용한 전문(Full-text) 검색시스템의 자연어 검색기법에 관한 연구

(주)한국파일링
부장 김 옥 란

목 차

I. 서론

기존의 대부분의 기관에서는 데이타의 관리 및 검색을 위해 RDBMS (Relational Data Base Maagement System)를 채택하여 활용해 왔다. 그러나 최근에 정형화 되지 않은 데이타의 비중과 그에 대한 관리 및 활용의 중요성이 대두되어짐에 따라 많은 기관들이 IRS(Information Retrieval System)에 대한 높은 관심을 갖고, 도입을 위해 제반사항에 대한 검토를 추진하고 있다.

따라서 본고에서는 IRS에 대한 기본개념 및 전문(Full-text) 데이타베이스 시스템의 검색어 형식인 자연어 검색기법과 관련된 주요내용을 살펴보았다. 또한 유관 도서관들간의 상호 협력체제하에 단위 도서관의 소장자료를 공동활용(Inter Library Loan)하기 위한 기초방안을 제시하였고, 인터넷을 통한 검색구현 절차를 소개하였다.

II. IRS(Information Retrieval System)기술과 전문(Full-text) 검색시스템

1. IRS 란?

정보검색의 원론적인 측면에서 내려진 IRS의 정의는 "정보 이용자가 필요로 하는 정보를 수집하여 내용을 분석한 후, 색인작업등을 거쳐 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템" 으로 요약된다. 보통은 대상이 문헌이므로 문헌 정보 검색시스템이라고도 한다[정영미 93].

기존의 문헌정보 검색시스템은 주로 서지정보나 안내정보와 같이 정보자료에 대한 2차적인 정보를 검색대상으로 하는 참조정보 검색시스템(Reference Retrieval System)이 주류를 이루어 왔다. 그러나 최근에는 참조 정보뿐만 아니라, 자료의 내용을 분석 및 가공. 저장하여 특정한 개념을 포함하는 자료를 검색할 수 있는 시스템의 요구가 증가하고 있으므로 본문에 나타난 개념을 검색할 수 있도록 기능이 제공되는 전문 검색시스템(Ful-text Retrieval System)과 멀티미디어 정보검색에 대한 관심이 높아지고 있다[김영택 94].

따라서 요즈음 통용되고 있는 IRS의 개념은 전통적인 RDBMS의 여러가지 한계를 보완한 문헌정보관리를 위한 전문 데이타베이스 검색엔진으로 요약된다.

2. 전문검색시스템(Full-text Retrieval System)

가. 출현배경

전문 데이타베이스는 양적으로 점차 증가하고 있는데 Gale Directory of Database(Gale Research, 1994)에 의하면, 1985 년에 등록된 전문 데이타베이스의 수는 총 1926 개의 문자로 이루어진 데이타베이스중에서 28%인 535 개였으나, 1990 년에는 총 4212 개 중에서 42%인 1786 개로 증가하였고, 1993 년에는 총 6652 개 중에서 48%인 3155 개로 증가하였다고 한다.

이처럼 전문 데이타베이스가 증가하게 된 대표적인 이유으로는
첫째, 입력비용 및 저장비용이 감소되고
둘째, 많은 신문사들이 신문제작을 전산화하였으며
셋째, 전문 검색 소프트웨어 팩키지가 발전하고
넷째, 전반적인 인쇄매체 정보산업이 컴퓨터 지향으로 발전하고 있으며
다섯째, 온라인으로 1 차 정보원인 전문을 이용하려는 이용자의 요구가 증가하고 있으며

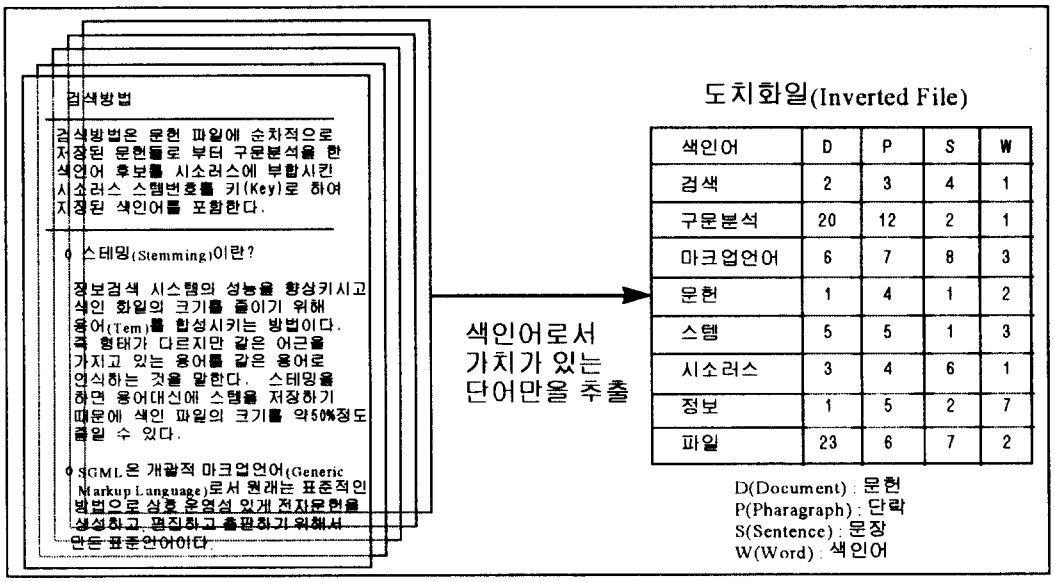
또한 생산자(색인 작성자)가 색인작업에서 초래할 수 있는 비일관성의 문제와 경비문제를 해결하기 위하여 전문 데이타베이스의 생산.개발에 주력을 가하게 되었기 때문인 것으로 본다.

나. 시스템 개요

문헌의 전문 전체가 수록되어 있기 때문에 전문에 출현하는 불용어를 제외한 모든 용어에 대해 자연어 형식으로 검색할 수 있는 데이타베이스로서 신속하고 효율적인 검색을 위해 도치파일(Inverted File)구조를 사용하고 있다.

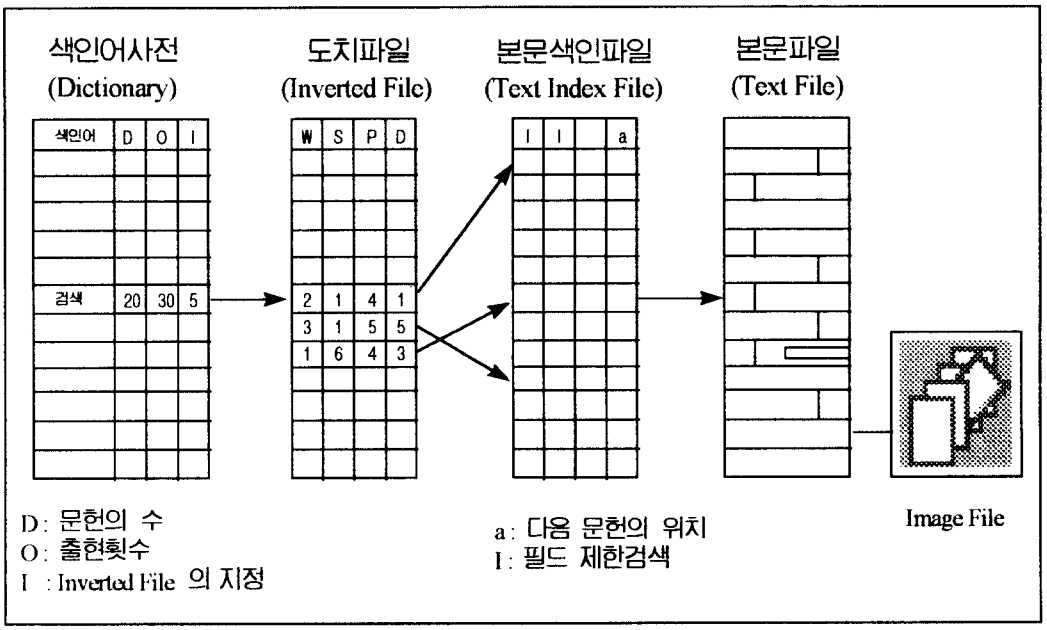
도치파일은 대부분의 상용 정보검색 시스템에서 적용하고 있는 정보 저장 방법으로서 재래식 순차파일과 체인파일의 단점을 보완하기 위해 고안된 일종의 변형된 체인파일이라 할 수 있다. 따라서 문헌파일과 디렉토리인 인버티드-색인파일로 구성되며, 색인어를 디렉토리로 하여 그 색인어를 포함하는 문헌번호를 색인어에 연결시켜 파일을 구성한다[Gerard Salton 75].

아래의 그림은 색인어로서 가치가 있는 단어만을 추출하여 각 단어의 위치정보를 도치파일에 저장하는 전문검색 시스템의 내부구조를 나타내고 있다.



(그림 1) 전문 검색시스템의 정보저장방법

검색방법은 문헌 파일에 순차적으로 저장된 문헌들로 부터 구문분석을 한 색인어 후보를 시소러스에 부합시킨 시소러스 스템번호를 키(key)로 하여 지정된 색인어를 포함하는 문헌들을 효율적으로 추출하므로써 속도가 빠르다. 아래의 그림은 예)에서 설명한 바와 같이 전문 검색시스템의 검색절차를 나타내고 있다.



(그림 2) 전문 검색시스템의 검색절차

예) 이용자가 "검색"이라는 단어를 검색하였을 경우에 색인어사전(Dictionary)에서는 그 단어가 20 개 문서에서 30 번 출현한다는 정보를 출력해 주며, 이용자가 본문을 열람하고자 하면 도치파일(Inverted File)로 부터 그 색인어를 포함하는 문헌의 위치정보를 취하고, 본문색인파일(Text Index File)에서 본문(Text File)의 Pointer 로 이용자에게 본문과 검색한 단어를 출력한다.

III. 자연어 검색시스템

1. 개요

문헌 내용의 주요개념을 나타내는 단어나 문구등을 문헌에 나타난 그대로 검색어로 채택하는 경우로서, 언어의 통제 과정이 불필요하기 때문에 별도의 색인 작성을 위한 경비가 필요 없으며, 새로운 정보의 접근이 용이할 뿐만 아니라, 색인 작성자의 주관에 따라 같은 주제가 여러가지 다른 용어로 표현될 수 있는 것을 방지할 수 있다는 장점이 있다. 반면에 전혀 통제를 받지 않고 문헌에 나타난 그대로 검색어로 사용되기 때문에 같은 개념이 여러개의 다른 용어에 의해 표현될 가능성이 있으며, 심층 정보의 검색이 어려울 뿐 아니라 포괄적인 개념의 정보 검색이 어렵다는 단점이 있다[김재수 93]

결국 색인어 및 탐색어를 자연어로 하는 시스템의 가장 큰 문제는 검색의 정확율이 떨어질 수도 있다는 점인데, 그 가장 중요한 이유는 대개의 경우 표제나 초록 또는 텍스트의 본문속에 나타나 있는 단어는 기능어를 제외하고는 모두가 탐색어로 사용될 수 있기 때문에 탐색이 자유롭고 편리한 대신에 실제로 그 단어가 주제어가 아닌 문헌도 모두 검색되는 결과를 가져온다[정영미 93].

따라서 최근에 이르러 대표적인 대규모 정보 검색시스템에서는 자연어와 통제어를 병행하여 사용하는 시스템으로 설계되는 추세이기 때문에 어휘 통제를 위한 시소러스의 기능은 계속적으로 그 역할이 증대될 것으로 본다.

2. 검색효율을 높이기 위한 방안

가. 시소러스의 활용

시소러스의 어원은 희랍어에서 파생된 용어로 "지식의 보고" 또는 "백과사전"이라는 의미를 갖고 있는데, ISO 는 시소러스를 "구조적인 면에서는

상위개념의 용어와 하위개념의 용어를 의미론적으로 밝힌 어휘집이며, 기능적인면에서는 색인자 또는 정보 이용자가 자연언어를 통제언어로 변환하는데 사용하는 어휘집이다." 라고 정의하고 있다[ISO 2788, 86].

그러므로 검색의 효율을 높이기 위해서는 통제 어휘집인 시소러스를 활용하여 다양한 자연어를 일정하게 약속된 언어로 변환 시킴으로써 일관성을 유지하고, 나아가서 용어의 관계를 나타내 줌으로서 전체적인 체계를 조직화하는 것이 필요하다. 이에따라 용어의 부정확성이나 용어의 변화 및 다양성에 대처하게 되고, 본문중의 용어와 검색자의 용어를 조정하여 검색의 정확율을 향상 시킬 수 있다.

따라서 지금까지 이와 관련된 선행연구를 종합하여 보면, 국제적인 대규모 전문 검색시스템에서는 자연어와 어휘 통제집인 시소러스를 활용하여 상호보완적으로 검색할 수 있도록 지원되는 것이 많으며, 이러한 양자의 병용이 가장 바람직한 것으로 간주되고 있다[김은식 91].

또한 시소러스 기능은 대부분의 상용 IRS 엔진에서 지원되고 있는 실정이나, 도입하려는 기관에서는 관련 용어의 표준화에 효율을 기할 수 있는 표준제품(예: ANSI Z39.19)에 대한 고려를 중요시 해야 한다.

나. 검색 기법

지금까지 대부분의 상용 검색시스템에서 사용된 검색기법은 일반적으로 집합이론에 기초한 불리안 논리 검색기법이 기반을 이루고 있다.

불리안 논리 검색에서 탐색문은 개념을 나타내는 탐색어와, 이들의 논리적 관계를 나타내는 불리안 논리 연산기호(AND, OR, NOT, XOR)로 구성되며, 질문에 포함된 탐색어들의 정확한 조합을 통해 이와 완전히 일치하는 문헌 레코드를 검색하게 된다. 따라서 불리안 논리 연산을 통해 모든 조건을 만족하는 진리값 1인 문헌만을 검색하게 되는데, 이러한 검색 메카니즘은 매우 단순하면서도 종종 높은 검색성능을 보여주기 때문에 가장 널리 이용되고 있다.

그러나 불리안 논리 검색은 이용자 질문의 탐색어와 문헌의 색인어가 완전히 일치하는 문헌만을 검색함으로써 검색대상을 상호배타적인 체계(검색된 문헌.검색되지 않은 문헌)로 분리하게 된다. 따라서 검색된 문헌과 검색

되지 않은 문헌 사이에 해당되는 부분적으로 일치하는 문헌을 검색대상에서 제외시킴으로써 이용자 요구에 적합한 다수의 문헌을 누락시킨다는 문제점이 지적되어 왔다.

정보검색의 이론적 발전은 이러한 불리안 논리 검색의 한계에 대해 많은 대안들을 제시하였는데, 그중에서도 가장 강력한 대안이 퍼지검색 및 확률검색이라 할 수 있다.

퍼지검색 이론은 정보검색의 애매한 표현과 불확실한 상태를 표현한다는 의미에서 1965 년 자데(L.A. Zadeh)에 의해 처음으로 소개된 이론이다. 퍼지집합의 개념이 정보검색에 적용되는 경우 불리안 논리 검색을 확장 또는 일반화하기 위한 가중치시스템과 관련된다.

퍼지검색은 문헌의 색인어와 질문의 탐색어에 용어의 상대적 중요성을 나타내는 0 과 1 사이의 가중치를 부여함으로써 특정 색인어는 문헌에 부여될 수도, 부여되지 않을 수도 있으며, 중간정도의 중요성으로 부여될 수도 있다는 전제하에 불리안 검색을 확장 또는 일반화 하기 위한 검색기법으로 불리안 검색이 안고 있는 한계점들을 어느정도 해결할 수 있을 것으로 본다. [Bookstein,1985]

확률검색은 불확실성을 정보검색의 본질로 인식한 Maron & Kuhns 에 의해 1960 년대 부터 연구되기 시작하였으며,[Maron & Kuhns, 1960] 현재도 계속 연구개발 되고 있는데, 기본개념은 어떤 검색시스템도 어떤 문헌이 이용자의 요구에 적합한지 확실성을 가지고 예측할 수 없기 때문에 시스템은 반드시 확률을 다루어야 한다는 것으로 이론적으로는 가장 무장이 잘된 모델로 알려져 있으나, 지금까지 개발된 확률적 접근방법의 최적의 모형에 대해서는 일치가 이루어지지 않고 있다.

이 모델의 주안점은 어떤 문서의 주어진 질의에 대한 적절성(relevance)은 그 문서가 사용자에게 적절한 조건확률로서 표현되어야 한다는 것이며, 그 계산은 데이타베이스의 통계적인 특성과 표본(sampling)을 통해 추정되어진다. 이는 매칭 함수를 이용해 질문의 탐색어에 가중치를 구함으로써 적합성 확률의 순으로 문헌을 순위화 한다[Turtle, H. & Croft, W.B. 1991].

이 두가지 검색기법은 불리안 논리 검색의 문제점을 해결하기 위해 개발된 검색모형이라는 공통점을 가지고 있으면서도 검색과정에서 부여되는 가

중치에 대한 해석 및 접근방법에서는 서로 상이한 특성을 나타내고 있다. 따라서 대상이 되는 문헌집합이 클 경우에는 퍼지검색이 보다 효율적일 수 있으나, 특정 문헌집단에 대해 많은 경험과 통계적 데이타가 축적된 후에는 확률검색이 유용할 수 있다.

다. 자동색인 시스템의 지원

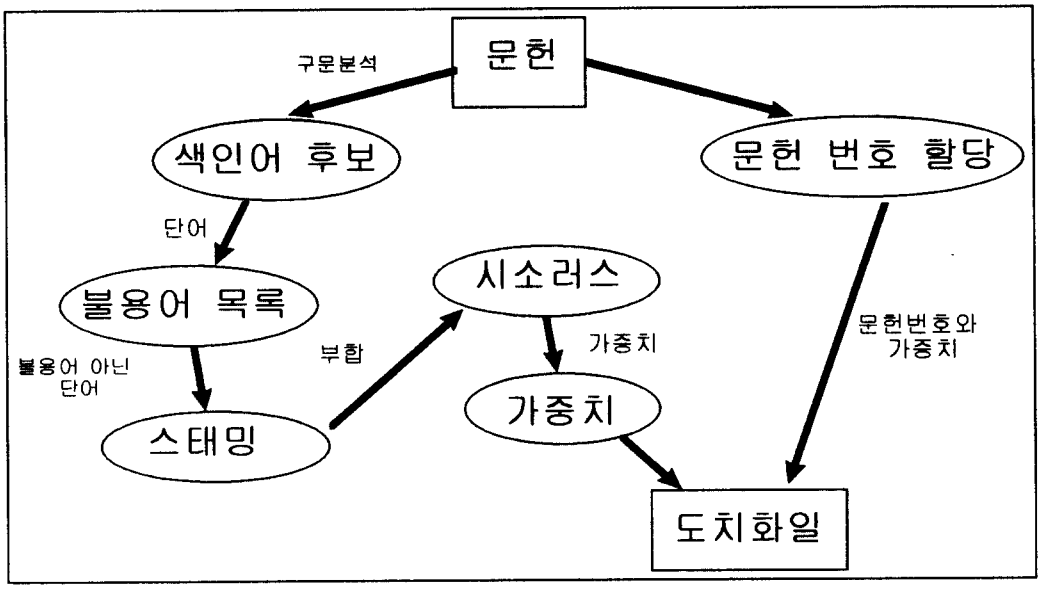
전문 검색시스템에서 정보검색의 요구를 효율적으로 처리하려면 색인 전문가가 수동으로 색인하던 수동색인 작업을 컴퓨터를 이용하여 자동색인(Automatic indexing)으로 처리하는 것이 필수적이다. 그런데 자동색인에서는 기계적으로 색인어(Index term)를 추출하기 때문에 수동색인 결과에 비해 색인의 질이 저하되는 문제가 발생하므로 불용어(Stop word) 리스트를 이용하여 검색어로 사용될 가능성이 매우 적은 용어를 제거하거나, 시소러스를 사용하여 색인어를 통제하여야 한다.

따라서 한국어 자동 색인의 경우 형태소 분석은 문서에 나타나는 명사들을 인식하고 이를 색인어로 추출하기 위한 경우나 불용어를 제거한 색인어의 선별을 위해서도 필요하다. 그러나 한국어의 경우에는 교착어라는 언어의 특성때문에 용언의 활용형이 다양하므로 불용어 리스트만으로 불용어 처리를 하기는 어려우며, 어형 통제보다도 색인어 선별을 위한 수단으로써 형태소 분석이 필수적이다.

한국어에서 색인어 선별이 중요한 이유는 복합명사와 고유명사, 외래어 등 사전 미등록어를 추정할 때 단어가 사전 미등록어인지를 알 수 없기 때문이다. 특히 자동색인에서는 미등록어를 정확하게 추정하는 기능이 그 성능을 좌우한다[강승식 94].

그러나 형태소 분석 기법만을 이용한 자동색인 과정에서는 복합 명사 처리, 사전 미등록어 추정, 모호성 문제, 오류어절의 인식 기능등과 같은 보완되어야 할 사항들이 있으므로 자연어 색인의 효율을 높이기 위해서는 구문분석이나 의미 분석 기법을 부분적으로 도입하여 활용하는 방안에 대해 실험해 보고 그 효용성을 검증해야 할 필요가 있다.[김판구, 94]

아래의 그림은 자동색인 시스템의 과정을 나타내고 있다.



(그림 3) 자동색인 시스템의 전체 구성도

o 스테밍(Stemming)이란?
정보검색 시스템의 성능을 향상시키고 색인 화일의 크기를 줄이기 위해 용어(Term)를 합성시키는 방법이다. 즉 형태가 다르지만 같은 어근을 가지고 있는 용어를 같은 용어로 인식하는 것을 말한다. 스테밍을 하면 용어대신에 스템을 저장하기 때문에 색인 파일의 크기를 약 50% 정도 줄일 수 있다.

IV. 구축된 DB 공유를 위한 기초방안

1. SGML(Standard Generalized Mark-up Language)을 이용한 DB 구축

SGML(Standard Generalized Markup Language)은 개괄적 마크업언어(Generic Markup Language)로서 원래는 표준적인 방법으로 상호 운영성 있게 전자문헌을 생성하고, 편집하고 출판하기 위해서 만든 표준언어이다. 개괄적 마크업이란 문헌에 제목, 초록, 본문, 장(章)과 같이 문헌요소의 속성을 표시하는 것으로서 하나의 문헌이 여러 문헌요소(Element)로 분해되고 이러한 문헌요소간의 관계가 표현되는 것이다[오민경 95].

따라서 SGML 을 이용한 전문 검색시스템의 구축은 문헌을 문헌요소라는 객체 단위로 이루어진 것으로 보고 표준적인 방법으로 구조화 시킬 수 있기 때문에 각 기관에 구축된 全文情報(Text Information)가 특정의 處理系(Processing)에 제한받지 않고 각 시스템간에 교환과 저장이 가능하며, 정보 검색용 표준 통신 프로토콜인 Z39.50 을 통해 오류없이 문헌을 전송하는 것이 가능하다.

즉 현재 Web 상에서 문서를 교환할 경우 인터넷 표준 문서형식인 HTML 로 변환해야 하듯이 SGML 은 이러한 HTML 을 포함하는 보다 상위 개념의 표준 문서형식으로서 향후 국가적 통합 디지탈 도서관을 구축하기 위해서는 각 기관의 다양한 형태의 정보를 수집하여 상호 교환할 수 있도록 통합 DB 구축시에 반드시 SGML 형식으로 변환하여 저장할 수 있도록 전단계에서 필터링(Filtering)해 줄 수 있는 솔루션이 전제가 되어야 한다.

2. 인터넷을 통한 검색구현 방법

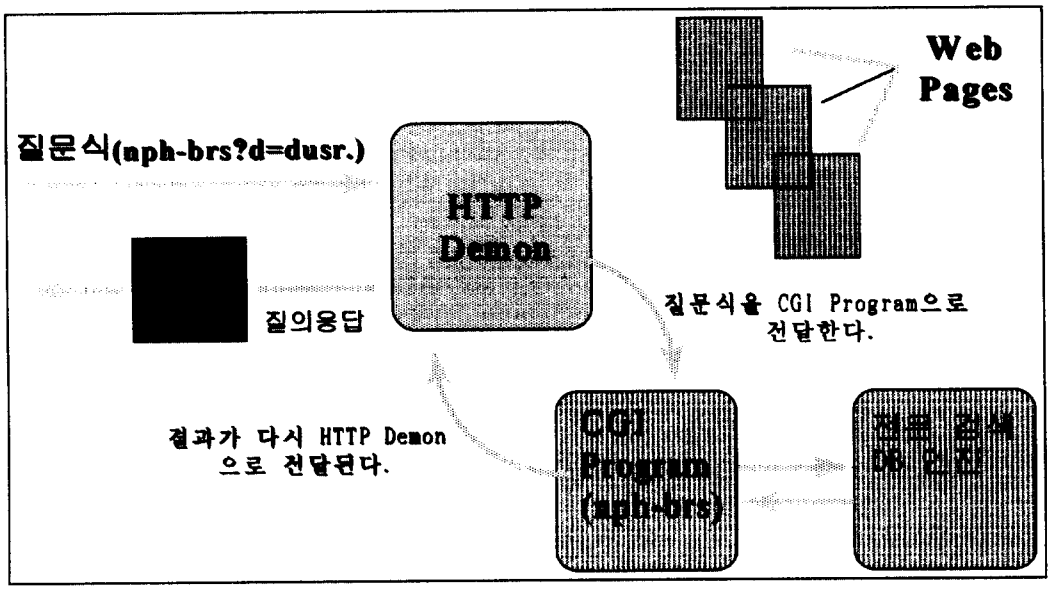
고속통신기술과 컴퓨터 테크놀로지 개발은 네트워크 접근에 의한 정보서비스를 기존의 것보다 어느면으로나 보다 효과적이고 효율적인 어떤 새로운 것으로 기대할 수 있게끔 유도해 오고 있다. 그 중에서도 네트워크를 통한 전자정보가 갖는 잇점, 즉 이용자 측면에서 주목하고 있는 구체적인 측면이라면 전세계의 각처에 제각기 개발되어 온 여러 형태와 규모의 정보 네트워크들을 접속할 수 있다는 점이다[유사라 95].

따라서 이기종간의 다양한 네트워크들의 운영체제가 정보, 통신 테크놀로지의 지원으로 상호운용(Inter Operation)되기 때문에 실제로 단순한 네트워크 서비스 뿐만 아니라 이용자들이 인터넷상에서 텍스트는 물론 그래픽과 멀티미디어 정보 까지를 원활하게 검색할 수 있으며, 전자출판을 원하는 사람들에게도 유용한 도구가 될 수 있는 솔루션이 지원되어야만 진정한 의미의 최종 이용자 탐색이 될 수 있다.

이러한 솔루션의 인터페이스 부문에서는 기존의 웹 브라우져(Web Browser)를 사용하거나, 이 솔루션이 독자적으로 제공하는 표준 CGI(Common Gateway Interface)를 이용하여 다른 인터페이스를 사용해도 된다.

당사에서 지원하고 있는 이 솔루션의 검색 구현과정을 상세히 기술하면, 표준 웹 서버에서는 사용자의 Browser 에서 요청되는 내용을 직접 다루게 되는데, 이 경우에 이 CGI Program 은 Browser 에서 요청되는 내용을 일단 Layer 에 저장한 후 검색엔진으로 보낸다. 보내진 질의를 가지고 검색엔진은 정보를 찾아내고, 찾아낸 정보를 다시 이 CGI Program 으로 보내면 받은 정보를 HTML 문서로 변환하여 Browser 로 출력하여 준다.

(그림 4)에서는 이 CGI Program 의 검색 구현과정을 나타내고 있다.



(그림 4) 인터넷상에서의 검색절차

이와 같은 인터넷을 통한 정보검색의 구현은 지역이나 시간등의 제약으로 한 곳이나 일정 부류의 이용자들만이 접속이 가능하던 환경을 초월하여 저작권에 위배되지 않는 범위내에서 사전에 이용권한의 제약조건이 가하여진 경우를 제외하고는 누구나 보다 용이하게 접근이 가능하다는 점과 여러가지 포맷으로 기존에 구축된 다양한 정보를 송수신 할 수 있다는 점이 무엇보다 큰 강점이다.

V. 결 론

디지탈 도서관의 기능을 정보 이용자 측면에서 살펴보면, 본고에서 제시한 바와 같이 이용자 인터페이스는 실시간으로 고속의 통신경로와 다양한 네트워크 방식을 통하여 접근 가능하며, 원격이나 로컬에서 모두 지원된다. 또한 이용자의 정보요구는 자연어 인터페이스를 전제로 하여 어떠한 형식(MARC 레코드, WP 포맷, DIF 등)으로라도 탐색되어지며, 색인과정에서나 질의탐색, 그리고 전거의 경우도 여러 형태의 정보자료, 즉 하이퍼링크에 의한 하이퍼미디어 정보를 처리할 수 있는 기능이 지원된다.

검색엔진은 외국어 정보를 처리할 수있는 다국어 지원 어휘사전을 비롯한 용어집이 주제별로 지원되고, 방대한 대규모 수준의 정보에 대비한 압축색인과 일련의 문헌순위에 대한 정보치를 포함한다. 또한 원격 데이타베이스 정보자원에 접속하여 검색하기 위한 메타데이타의 기능도 추가되어야 한다.

물론 아직까지도 이용자측에서 네트워크 접속과 서비스활용이 수월하지만은 않은 문제점을 안고 있는데, 이는 대규모 네트워크 시스템 관리측면에서 계속적으로 보완되어가야 할 과제이다.

참고 문헌

강승식, "한국어의 형태론적 특성과 형태소 분석기법" 정보과학회지, 12/8,1994

강일중, 정영미, "용어관계를 이용한 검색문헌의 순위 부여에 관한 연구," 정보관리학회지, 제 8 권 제 1 호 (1991).

김은식, "과학기술용어 시소러스 대역 데이타베이스 구축," 정보관리연구, 22/2 (1991)

김영택. 자연언어 처리, 교학사, 1994.

김재수. 자연언어와 통제언어에 의한 정보 검색의 효율성 비교 연구 -미사일 공학 분야를 중심으로. 박사학위 논문 중앙대학교, 1993.

김판구. 한국어 정보검색을 위한 상호 정보량에 기반한 복합어 자동색인. 박사학위 논문, 서울대학교, 1994.

사공철등저. 최신정보검색론. 서울 : 구미무역, 1990.

오민경. SGML 을 이용한 문헌구조화 및 텍스트 검색에 관한 연구. 석사학위논문 연세대학교, 1995.

유사라, "가상도서관 모형과 환경정보 가상도서관 서비스 사례", 국회도서관보, 제 32 권 제 7 호 (1995).

이젬마, 불논리검색, 퍼지검색, 확률검색의 효율 비교연구. 석사학위논문 숙명여자대학교, 1994.

정영미, 정보검색론. 서울 : 구미무역, 1993.

정영미, "하이퍼 텍스트의 개념과 응용에 관한 고찰," 정보관리학회지 6(2) 1989.

Belkin, N.J. and Croft, W.B. "Classification of Retrieval Techniques," ARIST 22 (1987).

Bookstein, Abraham. 1985. "Probability and Fuzzy-Set Applications to Information Retrieval." ARIST 20 : 117-149.

ISO 2788, Documentation-Guideline for the Establishment and Development of Monolingual Thesauri, 1986.

Maron, M.E. & Kuhns, J.L. 1960. " On Relevance, Probabilistic Indexing and Information Retrieval." J. of ACM 7(3) : 216-244.

Salton, G., Dynamic information and Library Processing. New Jersey, Premtice Hall, Inc; 1975.

Tenopir,C. " Full Text Database Retrieval Performance," Online Review, Vol.9, NO.2 (1985).

Tenopir, C., "Full Text Database Retrieval Performance," Online Review, Vol. 9, No, 2(1985). pp. 149-164.

Terry Noreault, et al., "Automatic Ranked Output from Boolean Searches in SIRE," Journal of the American Society for Information Science, Vol.28, NO.6.

Turtle, H. & Croft, W.B. Evaluation of an infoerence network-based retrieval model. ACM Transactions on Information Systems, 9(3),1991.