인터넷기반 멀티미디어 정보검색 시스템 : 옥서'95 의 색인 및 검색

강현규, 장호욱, 전미선, 박세영

ETRI, 자연어처리 연구실

Design and Implementation of a Multimedia Information Retrieval System based on Internet

Hyun-Kyu Kang, Ho-Wook Jang, Mi-Seon Jun, Se-Young Park

Natural Language Processing Section, ETRI

요 약

본 논문은 인터넷 기반의 멀티미디어 정보 검색 시스템인 옥서 '95의 정보 색인 및 검색에 대한 설계 및 구현에 대하여 논한다. 정보 구축시 키워드의 확장 개념으로서의 키팩트 추출과 모호성 해소 그리고 키팩트, 하이퍼 문서 및 멀티미디어 데이타의 색인을 한다. 또한 검색시 자연언어 질의에 대한 키팩트의 추출, 확장 및 서열처리를 통하여 사용자가 원하는 정보를 검색하게 한다. 검색 대상의 문서로서 백과사전, 신문기사, 기술문서를 다루었으며 여러가지 검색 기능을 설계 및 구현하였다. 전문을 대상으로 색인 및 검색을 하였으며 앞으로 전자도서관이나 정보통신 서비스에 활용할 예정이다.

1. 서론

현대 정보화 사회에서 정보는 인간의 관리가 불가능 할 정도로 많이 쏟아져 나오고 있다. 따라서 이러한 수많은 정보들을 컴퓨터에 저장하고 이를 관리하여 필요에 따라 사용자에게 서비스하는 정보 검색 시스템(Information Retrieval System)이 널리 이용되고 있다[1-3].

최근 몇년들어 급속히 확산되기 시작한 인터넷은 그의 구축의 용이성과 대중성을 기반으로 정착하게 되었으며 정보의 공유라는 대전제하에 수많은 인터넷 접속 및 대량의 정보를 접할 수 있게 되었다. 그러나 이러한 수많은 정보가 있음에도 불구하고 정보의 바다라는 인터넷에서 원하는 정보를 찾기란 그리 쉽지 않다. 우선 어디에 정보가 있는지, 그리고 어떻게 정보를 찾아야하는지 또한 보다 원하는 정보를 정확하게 찾기란 쉽지않다[16,17].

일반적인 인터넷에서의 한글 정보검색 시스템은 키워드 및 불리안 질의를 기반으로 일반인들이 정확하게 사용하기가 어렵다. 또한 검색되는 대상의 정보들이 방대하기 때문에 사용자가 원하는 내용을 정밀하게 찾아주기란 쉽지 않다[16,17]. 다양한 정보환경 속에서 필요한 정보의 신속 정확한 검색 및 이를 위한 정보의 효율적 구축은 대단히 중요한 기능이다. 다가오는 초고속통신망 사회에서 정보검색은 기본적인 서비스로 활용될 것이며, 정보검색의 대상은 텍스트뿐만아니라 멀티미디어 정보가 주류가 될 것이다. 따라서 미래에는 다양한 미디어를 중심으로 방대한 정보를 구축하고 사용자의 요구에 맞는 다양한 정보를 제공하는 디지탈 도서관구축 및 다양한 검색이 필요하다[14].

본 논문은 인터넷을 기반으로하는 멀티미디어 정보 검색 시스템인 옥서 '95를 설계 및 구현한다. 먼저 다양성을 위하여 방대한 자료 및 멀티미디어 데이터를 가질 수 있는 백과사전을 다룬다. 또한 일반적으로 새로운 정보를 갖는 신문 기사를 다룬다. 그리고 실제 정보로서 일반 도서 등과 같이 문단 구조에 의한 일정한 깊이를 갖는 기술문서를 대상으로한다[14,15].

인터넷 환경에서 위에 열거한 정보들을 어떻게 구축하고 어떻게 검색하는지를 중심으로 기술한다. 사용자에게 보다 친숙하게 하기 위하여 일상적으로 사용하는 자연언어로 질의할 수 있는 인터페이스를 갖도록 설계 구현되었다. 또한 보다 정확한 내용의 의미를 다루기 위하여 기존의 키워드와는 달리 키팩트라는 개념을 도입하고 여러가지 지식정보들을 사용하여 사용자가 원하는 바를 실제 내용정보에 충실하도록 설계 구현되었다. 다양한 멀티미디어 정보 색인 및 여러가지 지식 정보로서 사전(dictionary) 정보를 색인 검색한다. 또한 문서적인 구조를 갖는 경우 보다 정확하고 원하는 정보를 검색해 주기 위하여 구조적인 정보를 자동으로 나누고 나뉘어진 정보를 대상으로 검색하고 그 이외 주변정보를 얻을수 있도록하였다. 마지막으로 방대한 문서에서 보다 신속하

고 정확하게 정보를 찾기 위하여 중요도에 따른 순서화를 통한 서열처리를 하였다.

제 2장에서는 인터넷에서의 시스템 구축 및 제공하기 위한 시스템 개념 구조 및 실제 설계 구현된 기능들을 나열한다. 제 3장에서는 정보 구축 구조 및 흐름에 대하여 상세히 설명한다. 제 4장에서는 정보 검색기에 대하여 설명하고 정확한 검색을 위한 부분들을 설명한다. 마지막으로 제 5장에서 결론을 맺는다.

2. 시스템 개념 구조 및 기능

2.1 시스템 개념 구조

사용자들의 관점에서 클라이언트와 서버의 기능에 사용자들의 역할을 고려하면 인터넷 정보검색의 시스템 개념 구조는 일반사용자 클라이언트, 정보제공 클라이언트, 서버 그리고 통신망으로 구성된다[15]. 그림 1에 전체 시스템 개념 구조가 나타나 있다.

- 일반사용자 클라이언트

일반사용자 클라이언트는 단순하게는 간단한 검색인터페이스(Browser) 만을 갖고서 서버로 검색 요구를 하고 그 결과를 받아 사용자에게 전달하는 것이다,

예를들면 베토벤의 운명 교향곡을 찾기위해 사용자가 검색인터페이스에 <베토벤>, <운명>이라는 키워드를 질의어로 하여 검색요구를 하고, 그 결과를 받았을때 문자 및 화상 정보는 화면에 보여주고 음악 정보는 스피커를 통해 사용자에게 전달하는 것이다.

- 정보제공 클라이언트

정보제공 클라이언트는 서버에 구축되어 있는 정보에 새로운 정보를 추가하거나 아예 처음부터 새로운 정보를 구축하는 기능을 담당한다. 서버가 지원하는 구축기능 혹은 별도의 정보구축 도구등을 이용하여 수행되며, 독립적으로 수행될 수 있는 일들로서는 서버에 구축될 정보를 미리 마크업 해놓는 것들은 가능하다.

위의 예의 경우 음악에 관련된 정보를 마크업하고, 디지털화하여 서버에 전달하거나 마크업 과정 중에 다른 연결된 정보가 필요한 경우 검색인터페이스를 통해 검색 확인하여 정보를 추가할 수 있는 기능이 있다.

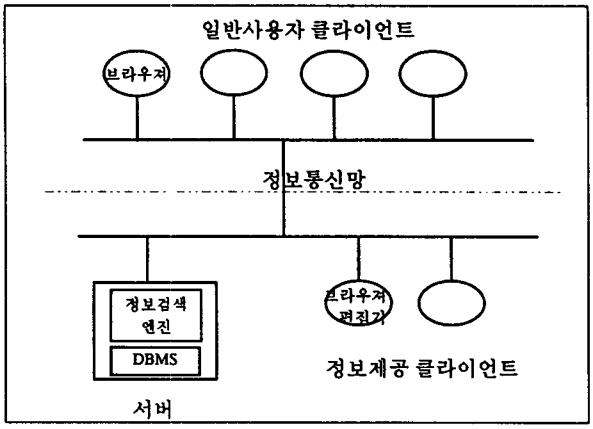
- 서버

검색요구의 대상이 되는 정보의 구축 기능 및 정보검색 엔진을 갖고 있으며 여러가지 서비스를 제공할 수 있다. 정보검색에 필요한 색인 정보를 생성 관리하고 자연언어 처리에 필요한 사전 및 키워드 목록들도 관리된다.

위의 예의 경우 마크업된 정보 및 디지털화된 음악 정보를 클라이언트로 부터 받아 색인과정등을 거쳐 서버에 구축되며, 이 구축된 정보는 클라이언트로 부터의 요구 즉 질의어 <베토벤>, <운명> 등을 통해 검색엔진에 의해 검색되어진 후 클라이언트로 전달된다.

- 정보통신망

클라이언트들과 서버간의 통신은 정보통신망을 통한다. 정보통신망은 계속 발전하고 있으며 미래는 유선뿐만 아니라 무선을 이용한 통신도 이용될 것이다. 정보검색을 위해 이용되는 통신망에서 흐르는 데이타는 정보검색 요구, 정보구축 정보, 정보검색 정보등이다.



<그림 1> 시스템 개념 구조

2.2 시스템 기능

본 시스템은 백과사전, 기술문서, 신문기사를 다루었으며 다양한 정보 검색 기능을 제공한다. 클라이언트에 있는 사용자는 정보검색 인터페이스를 이용하여 서버에 존재하는 대량의 자료 중 필요한 정보를 얻을 수 있다. 본 시스템에서 제공하는 검색 기능은 자연언어 검색 기능, 표제어 검색 기능, 제목별 검색 기능, 날짜 검색 기능을 통한 검색 기능, 그리고 문서에 대한 하이퍼 텍스트 검색 등을 제공한다[14].

- 자연언어 색인 검색

자연언어 색인 검색 기능은 사용자가 자연언어 질의를 통해 원하는 정보에 접근할 수 있도록 해주는 것으로 다음과 같은 특징을 가진다.

- 백과 사전 자연언어 검색 기능
- 기술 문서 및 신문에 대한자연언어 검색 기능
- 복합 명사 처리 기능 지원
- 전거어 처리 기능
- 가중치에 의한 문서 나열

- 표제어 색인 검색

표제어 색인 검색 기능은 표제어를 입력하여 여기에 해당하는 정보를 얻을 수 있도록 해주는 기능으로 아래와 같은 특징이 있다.

- 백과 사전 표제어 검색 기능

- 전거어 처리 지원
- 다어절 표제어 처리 기능
- 수 포함 표제어 처리 기능

● 전문 색인 검색

전문색인 검색은 백과사전, 기술문서, 신문기사의 내용에 대하여 색인하고 보다 그 의미를 정확히 다루기 위하여 다음의 기능들을 지원한다.

- 키팩트 검색 지원
- 모호성 해소에의한 검색
- 가중치에 의한 문서 나열
- 전거어 처리 지원
- 질의 확장 처리 지원

● 기술 문서 색인 검색

기술 문서 색인 검색 기능은 기술 문서 제목이나기술 문서의 구조에 따른 문단이나 계층적으로 문서를 검색할 수 있도록 해주는 기능이다.

- 기술문서에 대한 문단단위의 검색 지원
- 기술 문서 구조의 계층별 검색 지원
- 제목에 의한 검색

● 신문 색인 검색

신문 색인 검색 기능은 기사 제목이나 특정한 날짜를 입력하여 사용자가 직접 보고자하는 내용으로 접근할 수 있도록 해주는 기능이다.

- 신문 기사 제목별 검색 기능
- 날짜에 의한 검색
- 신문 자연언어 검색

● 하이퍼텍스트 색인 검색

문서들 사이에 하이퍼텍스트 기능을 추가하여 사용자가 원하는 정보로 직접 링크를 통해 접근할 수 있도록 한다.

- 표제어들에 대한 하이퍼텍스트 기능
- 인명에 대한 참조기능
- 시사용어에 대한 참조 기능
- 기술용어에 대한 참조 기능

● 멀티미디어 색인 검색

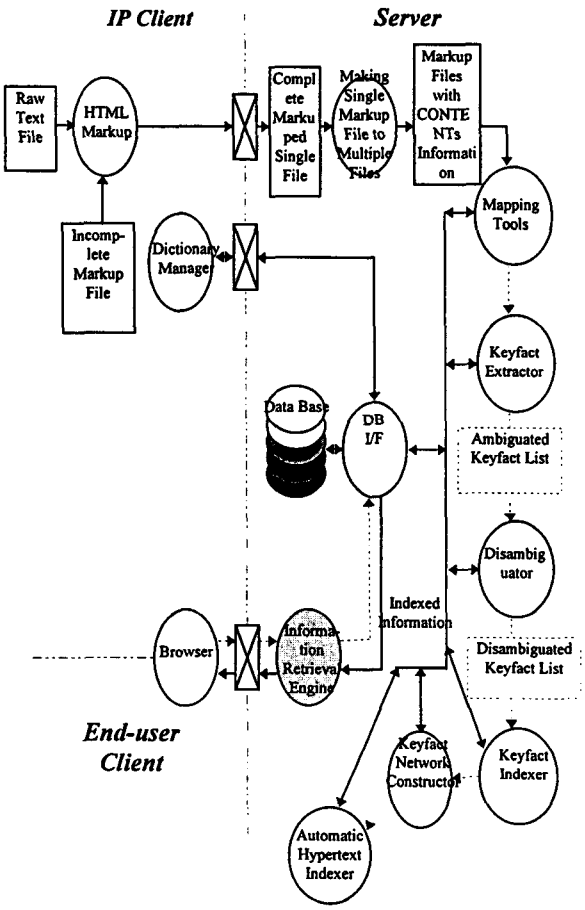
다양한 멀티미디어에 대한 색인을 자동으로 할 수 있도록 하여 멀티미디어 자료를 접근할 수 있도록 해주는 기능이다.

- 멀티미디어 자동 색인
- 이미지, 비디오, 사운드
- 국기, 지도, 애니메이션

3. 정보 구축기

정보검색을 위한 정보구축과정은 여러단계의 과정을 거쳐 이루어진다. 원래의 미디어(문서)에 대한 마크업 과정, 카드 생성 과정, 데이타베이스 변환(구축) 과정, 그리고 자동색인 과정이 있다. 자동색인 과정은 키팩트 추출 과정, 모호성 해소 과정, 그리고 키팩트 색인 과정을 거친 후 자동하이퍼문서 색인 과정을 마치면 멀티미디어 정보검색을 위한 정보구축이 완료된다[15].

그림 2와 같이 이러한 일련의 과정을 통하여 정보검색의 대상 정보를 구축하고 이 정보로 부터 기존의 참조정보를 갱신하여 다음 정보 구축시에 다시 복귀적용되도록 한다.



<그림 2> 정보 구축 흐름도

3.1 정보 구축 구조

정보제공 클라이언트에 의해 마크업된 문서는 서버로 전달되고 이 마크업된 문서는 문서의 특성에 따라 적당한 논리적 단위(카드, Card)로 나누어진다. 나누는 과정에서 다시 원래대로 복원할 수 있는 부가적인 정보를 생성한다. 이 카드와 복원정보(Recovery Information)는 DB 변환도구 (Mapping Tool)에 의해 데이타베이스에 저장된다.

다음 키팩트 추출기(Keyfact Extractor)는 데이타베이스에 저장된 카드들에 대해 사전등의 참조정보를 이용하여 키팩트를 추출하고 모호성이 잔재하는 키팩트 목록(Ambiguated Keyfact List)를 생성하여 모호성 해소기 (Disambiguator)로 넘긴다. 모호성 해소기는 키팩트 망의 정보를 참조하여 모호성 없는 키팩트 목록(Disambiguated Keyfact List)을 만들어 키팩트 색인기(Keyfact Indexer)로 넘긴다. 키팩트 색인기는 넘겨받은 키팩트 목록을 키팩트 역색인 화일(IDF; Inverse Document File)에 갱신한다. 그리고 키팩트 망 구축기(Keyfact Network Constructor)는 키팩트 역색인 화일을 참조하여 키팩트 망을 갱신한다.

다음은 자동 하이퍼텍스트 색인기 (Automatic Hypertext Indexer)가 자동화로 가능한 색인을 수행한다.

3.2 카드 생성기

정보 구축자에 의하여 입력된 문서는 일련의 과정을 통하여 문서의 논리적 구조를 고려한 카드의 형태로 바뀌게 되는데 이러한 작업이 카드생성기에서 이루어진다.

기본적으로 HTML(Hyper Text Markup Language) 문서는 생성당시 제작자에 의하여 통신 병목을 고려하여 적당한 기준으로 나뉘어져 있는 것이 사실이다. 그러나 본 개발 시스템은 이미 구조적으로 나뉘어져 있는 문서 뿐만 아니라 무작위로 입력되는 HTML 문서에 대하여 효율적인 관리, 저장 및 정보검색에 필요한 알맞은 구조를 갖추는 과정이 필요하다. 기본적으로 입력된 하나의 문서가 여러개의 카드로 나뉘어 지는 기준이 되는 것은 HTML의 태그로 사용된 <Hn> ... </Hn> (where 0<n<7)를 이용한다.

카드생성기 실행 결과의 예

- 입력되는 사용자의 문서가 다음과 같을때 :

```
<HTML>
<title> 기술문서 구조의 예 </title>
<h1> 기술문서의 제목 </h1>
<p> 요약문과 내용
  <h2> 1. 1의 제목 </h2>
  <p> 1 의 내용
    <h3> 1.1 1.1의 제목 </h3>
    <p> 1.1의 내용
    <h3> 1.2 1.2의 제목 </h3>
    <p> 1.2의 내용
  <h2> 2. 2의 제목 </h2>
  <p> 2의 내용
</HTML>
```

- 카드생성기의 실행결과로 출력되는 카드는 다음과 같다.

```
<h1> 기술문서의 제목 </h1>
<p> 요약문과 내용
<h2> 1. 1의 제목 </h2>
<h2> 2. 2의 제목 </h2>
```

```
<h1> 기술문서의 제목 </h1>
<h2> 1. 1의 제목 </h2>
  <p> 1의 내용
  <h3> 1.1 1.1의 제목 <h3>
  <h3> 1.2 1.2의 제목 </h3>
```

```
<h1> 기술문서의 제목 </h1>
<h2> 1. 1의 제목 </h2>
  <h3> 1.1 1.1의 제목 </h3>
  <p> 1.1 의 내용
```

```
<h1> 기술문서의 제목 </h1>
<h2> 1. 1의 제목 </h2>
  <h3> 1.2 1.2의 제목 </h3>
  <p> 1.2 의 내용
```

```
<h1> 기술문서의 제목 </h1>
<h2> 2. 2의 제목 </h2>
<p> 2의 내용
```

3.3 키팩트 추출기

사용자의 정보검색 요구를 처리하기 위해서는 먼저 정보검색 구축의 대상이 되는 텍스트 화일에 대한 색인 화일을 만들어야 한다. 기존의 정보검색 시스템에서는 '명사'와 '복합 명사'로 구성되는 키워드를 텍스트에서 추출하여 키워드의 빈도수에 의한 색인 화일을 구성하지만 한글 키워드가 가지는 모호성과 대표성의 한계를 보완하며 텍스트의 내용을 상세히 표현하고 불필요한 키워드의 추출을 막아 키워드만 사용할때 보다 정확한 정보검색을 제공하기 위해 키워드에 동사정보를 사용하여 '키워드 + (동사정보)'로 구성되는 키팩트 개념을 도입하여 키워드와 간단한 문장을 정보검색 시스템에 이용한다[9].

사전은 명사 사전 단어수는 약 10만개, 조사 사전의 단어수는 현재 221개, 동사 사전은 "파생형:기본형"의 형태로 구성되어 있는데 현재까지 작성된 파생형 동사의 갯수는 약 22,000개이며 기본형 동사는 7,000개이다. 동사 사전은 계몽사 학생 백과 사전과 일간 신문의 일부를 참조하여 작성하였으며 사전의 내용을 수정, 보완했다.

다음은 하나의 입력 화일에 대해 키팩트 추출을 수행한 예이다. 키팩트 추출의 장점은 동사 정보를 사용함으로써 불필요한 키워드 추출을 막고 키워드 사용시보다 텍스트의 내용을 보다 상세히 표현한다는 점이다. 키팩트의 갯수는 키워드에 동사 정보를 추가함으로써 조금 늘어나게 된다.

- 입력

```
<s1> 아가미 </s2>
<p> 무척추동물에서는 환형동물 이상의 동물이
가지고 있다. 빗살 모양으로 갈라져 있고 많은 모세 혈관이
있어서 붉은색을 띤다. 물 속에 녹아 있는 산소를
받아들이고 몸 안에서 생긴 이산화탄소를 물 속으로 걸러
낸다.</p>
```

- 키팩트 추출 결과

무척추동물
환형동물
동물
동물이 가지다
빗살 모양
빗살
모양
빗살 모양으로 갈라지다
모세 혈관
모세
혈관
모세 혈관이 있다
붉은색
붉은색을 띠다
물
산소
산소를 받아들이다
몸
이산화탄소
물

추출된 키팩트의 갯수 : 20

불필요한 키워드 추출의 방지 : 동물, 가지 -> 동물이 가지다
텍스트의 내용을 상세히 표현 :
빗살 모양 -> 빗살 모양, 빗살 모양으로 갈라지다
모세 혈관 -> 모세 혈관, 모세 혈관이 있다
붉은색 -> 붉은색, 붉은색을 띠다
산소 -> 산소, 산소를 받아들이다
키팩트의 갯수는 키워드의 갯수보다 '키워드 + 동사'의 수만큼 늘어난다.

3.4 모호성 해소기

저장된 카드들에 대해 키팩트 추출기가 명사사전, 조사 사전, 동사 사전을 이용하여 키팩트를 추출할 때에는 단어의 형태 비교를 근간으로 하므로 의미 모호성을 가진 키팩트들이 존재할 수 있다. 모호성은 그 단어가 포함된 문장내의 다른 단어들에 의해 해소될 수 있다. 모호성을 해소하기 위해서는 한 문장내에 최소한 두 개 이상의 단어가 나타나야 하며 표면에 나타난 단어들만으로는 모호성 해소 지식이 부족할 수도 있다. 그러므로 표면에 나타난 단어들은 키팩트망을 참조하여 의미적으로 관련있는 단어들로 확장한 후, 즉 좀 더 많은 모호성 해소 지식을 얻은 후 해소하였다[8].

3.5 키팩트 색인기

먼저 각각은 이미 카드별로 구분된 HTML의 개개 화일들로 구성되어 있다. 각각의 HTML 문서들로 부터 키팩트가 추출되고 모호성이 해소된 키팩트 들에 대하여 자연언어 검색을 위해 자동 색인한다. 키팩트 색인기는 크게 1차 자연언어 검색을 색인하는 1차 검색 색인기, 2차 검색의 지식정보를 색인하는 상호정보 색인기 그리고 2차적으로 문서의 순위를 조정하기 위해 색인하는 2차 검색 색인기로 구성된다[10,12].

3.5.1 1차 검색 색인기

1차 검색 색인기는 먼저 키팩트 추출기 및 모호성 해소기로 부터 추출되고 모호성이 해소된 키팩트 리스트를 입력 받는다. 입력된 키팩트 리스트들에 대하여 각 키팩트의 빈도를 계산한다. 먼저 키팩트 빈도를 계산하기 위해서 입력 받은 키팩트들에 대하여 정렬을 한다. 각 HTML 문서들에 대하여 정렬된 키팩트 리스트들은 그의 빈도가 계산된다. 빈도가 계산된 키팩트들은 키팩트 , HTML ID, 및 빈도의 정보를 갖는 키팩트 카드빈도에 저장된다. 그런 후 전체적인 키팩트들을 정렬하여 빈도가 계산된 키팩트 리스트를 만든다. 마지막으로 빈도가 계산된 키팩트들은 키팩트, 키팩트 문서 출현 빈도, HTML ID 및 빈도의 쌍 정보를 갖는 키팩트 색인에 저장된다.

- 문서 빈도 추출

모호성이 해소된 키팩트 리스트를 입력받아 정렬된 키팩트 리스트를 만들고 같은 키팩트끼리 병합을하고 그의 빈도를 계산한다.

- 키팩트 정렬

빈도가 계산된 키팩트 리스트를 바탕으로 키팩트, HTML ID, 빈도를 정렬하여 빈도가 계산된 키팩트 리스트를 만든다.

- 키팩트 색인

빈도가 계산된 키팩트 리스트를 바탕으로 최종 키팩트 색인 한다. 키팩트 색인은 우선 정형화된 키팩트 및 HTML 문서에서의 키팩트 출현 빈도를 가지며 이들은 또한 실제 그 갯수 만큼의 HTML 문서 번호, 빈도를 갖는다.

3.5.2 상호정보 (Mutual Information) 색인기

상호 정보망이란 사용할 문헌의 집합에서 용어들간의 관련도를 구하는 방법이다. 문장에 쓰인 모든 단어는 서로 유기적으로 관계를 가진다. 단어는 그 자체로서 의미를 가지기 보다는 다른 단어와 함께 쓰일 때 더욱 명확하게 의미를 전달하게 되는 것이다. 따라서 한 단어의 의미는 그 자체에 내재되었다기 보다는 다른 단어에 의하여 정의된다고 볼 수 있는 것이다. 문서들을 보면 단어와 단어 사이에는 의미적, 구문적 관계가 있음을 알 수 있다[5,11].

예를 들어 '의사'라는 단어어는 간호원,, 병원, 환자, 치료 등의 단어들과 함께 쓰이는데 이러한 단어의 연관성을 측정하여 질의어 확장시에, 질의어에 나타나는 용어와 관련된 용어들을 어느 정도까지 포함할 것인가에 대한 척도를 상호 정보망에 의해 제공된 용어간의 관련도에 의해 조절 가능하다. 또한 질의어에 나타난 용어와 검색된 문서의 내용내의 용어에 대한 관련도를 알수 있어 검색문서의 순위를 재조절 할 수 있다[10].

3.6 하이퍼문서 색인기

하이퍼 문서인 경우 일반적으로는 문서에 나타난 정보 즉 앵커 포인트로 부터 관련된 다른 문서로의 이동을 의미한다. 일반적인 하이퍼 문서인 경우는 수동에 의해 앵커 포인트와 목적 문서를 정의하게 된다. 자동 하이퍼 문서인 경우에는 사실상 정의가 어렵게 된다. 그러나 기술 문서인 경우 이미 언급한 카드 생성기의 절편

(segmentation)된 HTML 문서에 따라서 목차 내지는 상위 구조(chpter 나 section)의 앵커 포인트를 가지는 HTML 문서에 의하여 다른 문서로의 이동이 가능하도록 구현되었다.

또한 현재 시스템에서 구현된것은 백과사전, 신문, 기술 문서를 고려 할 때 공통적으로 고려할 수 있는 것은 사전류이다. 즉, 백과사전의 경우 표제어에 의한 백과사전의 내용 문서로의 이동, 신문의 경우 시사용어에 대한 시사용어 사전 내용 문서로의 이동, 그리고 기술 문서의 경우 기술 용어 사전 내용의 문서로의 이동을 의미한다.

전자신문이나 기술문서의 제목검색을 위한 색인 그리고 전자신문의 날짜 검색을 위한 색인이 구현되었다. 먼저 하이퍼 텍스트 색인을 위해 사전을 구축하였다. 마찬가지로 사전들도 각각의 표제어별 하나의 HTML 문서를 구성하였다. 또한 하이퍼텍스트의 하이퍼 링크를 위해 백과사전, 신문, 기술 HTML 문서에 각각 표제어, 시사용어, 기술용어의 앵커를 달았다. 백과사전의 경우 애니메이션, 비디오, 사운드, 국기, 지도, 그래픽, 테이블, 사진의 데이타 및 캡션 정보를 이용하여 앵커를 달았다.

3.6.1 사전 구축

- 백과사전 표제어 사전
  백과사전의 각 표제어 별로 표제어순의 일련번호의 HTML 문서에 대하여 표제어와 그에대한 ID 인덱스를 구성하였다.

- 인명, 기술용어, 시사용어 사전 구축
  일반 텍스트 화일로 구성된 각 사전들에 대하여 먼저 표제어(제목)와 시작 주소 그리고 offset을 구성한다. 구성된 정보를 정렬한다. 정렬된 정보를 이용하여 원 텍스트들을 가지고 일련의 완전한 HTML 문서를 구성하였다. 또한 구성된 HTML 문서의 표제어와 ID 인덱스를 구성 하였다.

3.6.2 하이퍼텍스트 마크

- 백과사전 : 표제어
  이미 표제어가 마크된 백과사전의 HTML 문서들에 대하여 백과사전 표제어 사전 정보를 이용하여 앵커를 달았다.
- 백과사전 : 멀티미디어 데이타
  표제어 앵커가 달린 백과사전의 HTML 문서들에 대하여 표제어, 멀티미디어 데이타 화일 이름, 설명정보를 이용하여 앵커를 달았다.
- 신 문 : 인명사전
  신문의 HTML 문서들에 대하여 인명 사전 정보를 이용하여 앵커를 달았다.
- 신 문 : 시사용어 사전
  인명의 앵커가 달린 신문의 HTML 문서들에 대하여 시사용어 사전 정보를 이용하여 앵커를 달았다.
- 기술 문서 : 기술용어 사전
  기술문서의 HTML 문서들에 대하여 기술용어 사전 정보를 이용하여 앵커를 달았다.

3.6.3 제목검색을 위한 색인

- 신 문
  인명 및 사사용어의 앵커가 달린 신문의 HTML 문서들에 대하여 그의 제목과 해당 HTML 문서로의 앵커가 달린 제목을 위한 HTML 문서를 구성하였다.

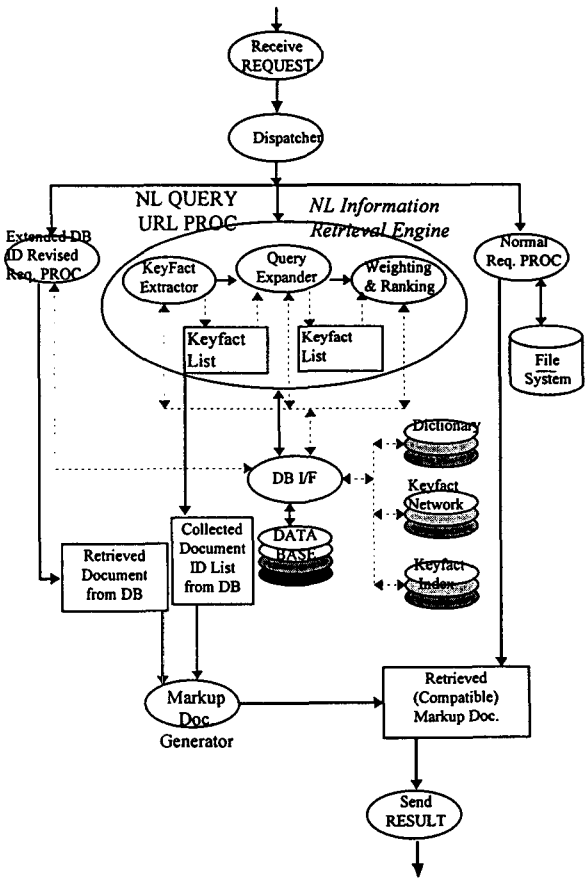
- 기술 문서
  기술용어가 달린 기술문서의 HTML 문서들에 대하여 그의 제목과 해당 HTML 문서로의 앵커가 달린 제목을 위한 HTML 문서를 구성하였다.

4. 정보검색기

정보검색은 정보구축과정에서 생성된 색인 정보를 이용하여 효율적이고도 신속한 문서탐색이 가능해 지며, 특히 자연언어 검색의 경우 키팩트 추출과정, 질의 확장 과정, 및 서열 처리 과정을 통해 사용자가 원하는 정확한 정보를 제공하여 준다[15].

4.1 정보 검색기 구조

정보검색의 유형은 일반 검색, 확장 DB 검색, 자연언어 검색으로 나뉘어진다. 그림 3은 정보 검색 흐름을 보여주고 있다.



<그림 3> 정보 검색 흐름도

일반 검색은 화일 시스템에 디렉토리의 경로에 따라 구축된 정보를 사용자가 제공하는 경로정보를 이용하여 검색하여 결과를

전달한다. 확장 DB 검색은 일반검색에서 경로명이 이용되듯이 데이타베이스에 저장되어 있는 정보를 카드 단위로 검색하는 경우는 DB 식별자로 검색이 이루어진다. 이 DB 식별자는 정보구축시 DB 변환도구에 의해 부여된 것이다. 자연언어 검색은 일반사용자는 좀더 편한 자연언어로 무엇인가를 찾기 원한다. 자연언어로 검색명령이 전달되면 자연언어 정보검색 엔진은 일련의 과정을 거쳐 검색된 정보들의 DB 식별자 목록을 생성하여 클라이언트로 전달한다.

정보검색 엔진의 자연언어 처리과정은 질의어에서 키팩트를 추출하고(Keyfact Extractor), 정확한 검색을 위해 질의어를 확장하고(Query Expander), 그리고 가중치 부여 및 서열화를 (Weighting and Ranking) 한다.

4.2 키팩트 추출기

정보검색의 요구로 자연언어 질의어가 들어오면 자연언어 검색기는 키팩트 추출기에서 질의어내의 키팩트를 추출하며 추출된 키팩트 리스트는 질의 확장기로 넘겨져 추출된 키팩트들을 확장하여 질의어에서 추출된 키팩트와 전체 텍스트를 대상으로 키팩트 색인화일과의 비교에 의해 순서화된 결과 리스트를 산출해 낸다.

- 키팩트 추출의 예

다음은 사용자의 자연언어 질의에 대해 키워드 추출과 키팩트 추출을 비교한 예이다.

입력
　　태양에 가까운 행성은?

키워드 추출 결과
　　태양
　　행성

키팩트 추출 결과
　　태양
　　태양에 가깝다
　　행성

키워드 추출과 키팩트 추출의 차이는 '태양에 가깝다'라는 문장을 추출했다는 점이다. '태양에 가깝다'를 포함한 문서는 사용자 질의에 대해 태양과 행성이라는 단어만을 가진 문서보다 정확한 내용을 제공해 줄 수 있다. 이와 같이 문장을 정보검색에 사용하면 키워드만을 사용하는 경우와는 달리 사용자에게 제공되는 문서 순위가 달라지게 된다.

4.3 질의 확장기

기존의 정보 검색 시스템에서는 문서와 질의에 나타난 명사류 단어만을 색인어로 추출한 후 색인어의 형태 비교를 근간으로 하는 검색이었다. 그러나 명사뿐만 아니라 동사와 형용사도 문장에서 핵심적인 역할을 수행한다. 질의에 나타난 색인어와 똑같은 형태의 색인어가 저장 문서에는 없더라도, 질의에 나타난 사용자 의도와 같은 의미를 가지는 문서가 있다면 찾아줄 수 있어야 한다. 단어의 형태 비교가 아닌 의미를 기반으로 하는 검색 시스템을 구현하기 위해서는 원래의 질의(original query)에 나타난 명사류 단어외에도 용언류도 색인어로 고려해야 하며, 질의에서 추출한 색인어들을 의미적으로 관련이 있는 명사류와 용언류들의 집합으로 확장할 필요가 있다.

키팩트망은 어떠한 동사나 형용사 또는 명사가 통사적 구성에 쓰이게 될 때 어떤 대상과 함께 쓰이는지 그 대상의 종류와 틀을 수집한 후, 단어와 단어의 관계를 밝히는 정보(어휘의 의미 범주)를 간직한 계층적 구조를 가지도록 하였다. 관계 정보는 보통 동의어(Used For), 하위어(Narrow Term), 상위어(Broader Term), 소속어(Has Part), 관련 명사(Related Term), 관련 용언(Fact Term) 등을 가진다. 또한 의미적으로 관련있는 대상들이 실제 말뭉치에서 갖는 상호 정보값을 포함한다. 그러므로 키팩트망은 의미 정보 이외에 통사와 통계적 정보를 포함한다는 점에서 시소러스와 차별화된다[4,8].

예를 들어 '당나귀와 말의 차이점은 무엇인가?' 라는 자연언어 질의에서 '말'은 짐승, 언어, 측량 단위 등의 여러가지 뜻을 갖는 다의어이다. 질의 확장시에는 당나귀와 3가지 뜻을 갖는 '말'의 관련 단어들을 비교하여 당나귀와 짐승 말의 관련 단어들의 교집합에서 짐승의 의미로 쓰인 말임을 알게된다. 그러므로 질의에 나타난 키워드 당나귀와 말은 당나귀와 짐승 말의 의미로 확장되어 정보 검색에 이용된다.

당나귀(a *donkey*): 갈기(HP,3,4,1), 갈색(RT,3,235,1), *말과*(BT,3,5,1), *꼬리*(HP,3,96,1), *나귀*(UF,3,5,2), *짐승*(BT,3,75,1), 갈기가 있다(FT,3,2,1), 말과에 속하다(FT,3,4,1), 몸이 튼튼하다(FT,3,3,1), 부리기에 알맞다(FT,3,2,1), 온순하다(FT,3,4,1), 털이 있다(FT,3,39,1)

말(horse): *짐승*(BT,198,75,1), 가축(BT,198,140,3), 굴레(HP,198,1,1), *꼬리*(HP,198,96,1), *나귀*(NT,198,3,1), *말과*(BT,198,4,1), 말발굽(HP,198,1,1), 조랑말(NT,198,3,1), 암말(RT,198,2,1),...

말(language):고립어, 문학, 언론과 출판의 자유, 상형 문자, 언어

말(a mal): 측량(BT,198,6,1), 그릇(RT,198,6,1), 측량 단위(RT,198,7,3), 부피(RT,198,99,1), 양(RT,198,16,1), 미터법을 도입하다(FT,198,1,1), ...

4.4 서열 처리기

서열처리기란 자연언어 검색시 질의한 질의어에 대하여 키팩트 추출 및 질의 확장된 키팩트 리스트를 입력 받아 중요도 계산 및 검색 가중치에 따라 해당 카드를 찾아 카드의 순서를 나열하는 것을 말한다. 서열 처리는 최종적으로 자연언어 질의에 대한 결과 리스트를 제공하는 것으로서 미리 색인된 키팩트들의 중요도 및 유사도 계산 방법에 따라 최종 결과가 달라지게 된다.

서열 처리를 위해서는 우선 해당되는 키팩트 들에 대하여 키팩트색인에 있는지의 여부와 있는 경우 각 키팩트에 대한 부가 가중치 부여 방법에 따라 부가 가중치 값을 얻고, 각 키팩트에 대한 유사도 값을 얻은 다음 전체적인 가중치에 따라 정렬하게 된다.

- 자연언어 검색

질의 키팩트 리스트 및 그의 갯수를 가지고 가중치 값에 따른 정렬된 카드 테이블 및 전체 카드 갯수를 계산한다.

- 유사도 계산

유사도 계산에서는 각 키팩트들에 대하여 해당 카드의 번호 및 중요도를 계산한다. 이때에 여러 키팩트들이 소속된 카드인 경우 키팩트의 중요도들이 계속 합산되게 된다.

- 결과 정렬

최종적으로 모든 키팩트들이 처리된 후 각 카드들이 가지고 있는 가중치 값에 따라 정렬함으로써 최종적으로 이 순서에 따른 카드들의 리스트를 보여 주게 된다.

- 2차 서열처리기

1차 서열처리기에서 나온 자연언어 질의에 대한 결과 리스트를 가지고 MI (Mutual Information : 상호정보) 값을 이용하여 다시 검색을 한다. MI를 이용하여 질의 내의 검색어들과 문서내의 키워드들간의 구도화된 상호 정보값을 구하고, 그 값과 문서내의 키워드 정보를 이용하여 문서순위를 재조정한다.

5. 결 론

본 논문에서는 인터넷을 기반으로하는 멀티미디어 정보검색 시스템인 옥서 '95의 정보 색인 및 검색에 대하여 논하였다. 먼저 인터넷 기반의 시스템 개념 구조 및 설계 구현된 기능에 대하여 설명하였다. 정보 구축에 관한 전반적인 흐름을 다루었으며 특히 HTML의 문서 카드 생성, 키워드의 확장 개념으로서의 키팩트의 개념을 기반으로하는 키팩트 추출, 어의 모호성 해소, 색인을 하였다.

또한 백과사전, 기술문서, 신문기사에 대한 사전 및 멀티미디어를 위한 하이퍼 텍스트 색인을 하였다. 마찬가지로 정보 검색시에도 키팩트를 기반으로하는 추출, 확장, 그리고 최종적으로 문서를 검색하여 순서적으로 나열하는 서열처리에 대하여 설계 및 구현하였다.

기존의 인테넷에서의 한글정보를 검색하여주는 로봇과는 달리 내용을 기반으로하는 전문검색, 멀티미디어 색인, 기술문서등의 구조에따른 검색, 그리고 자연언어 질의를 통한 보다 의미적인 정확성 향상에 촛점을 두어 설계 구현되었다.

앞으로는 다양한 형태의 검색 인터페이스를 갖는 새로운 형태의 정보검색 모델을 연구할 예정이며 보다 의미적인 처리에 역점을 두어 연구할 예정이다.

설계 구현된 내용은 앞으로 정보통신 서비스나 전자도서관 구축에 활용할 계획이다.

감사의 글

본 논문은 Gigabit 정보통신 시스템 소프트웨어 개발 사업의 멀티미디어 정보검색 기술 개발 과제의 결과로서 색인 및 자연어 검색 부분이외에, 과제에 참여한 정경택, 최동시, 김만수씨께 감사를 드립니다.

참고 문헌

[1] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, 1989.

[2] D. Harman and G. Candela, "Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking," *Journal of the American Society for Information Science*, Vol. 41, No. 8, pp. 581-589, 1990.

[3] T. Noreault, M. Koll, and J. J. McGill, "Automatic Ranked Output from Boolean Searches in SIRE," *Journal of the American Society for Information Science*, Vol. 28, No. 6, pp. 333-339, 1977.

[4] J. H. Lee, M. H. Kim, and Y. J. Lee, "Ranking Documents in Thesaurus-based Boolean Retrieval Systems," *Information Processing and Management*, Vol. 30. No. 1, pp. 79-91, 1994.

[5] K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29, 1990.

[6] Hyun-Kyu Kang, Se-Young Park, Key-Sun Choi, "A Precision Improvement of Korean Natural Language Retrieval System using a Two-level Ranking Method," Proc. of the 1995 International Confrence on Computer Processing Oriental Language, Hawaii, USA, 1995.11.25.

[7] Hyun-Kyu Kang, Chang-Yeol Lee, Ho-Wook Jang, Se-Young Park, "An Implementation of an Automatic Keyword Extraction System," Proc. of the 3rd Pacific Rim International Conference on Artificial Intelligence, Beijing, China, pp. 708-711, 1994.08.

[8] Mee-Sun Jeon, Se-Young Park, "Information Retrieval System Based on Keyfact Network," Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS) '95, Seoul, Korea, pp.199-204, 1995.12.

[9] Ho-Wook Jang, Se-Young Park, "Keyfact Concept for an Information Retrieval System," Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS) '95, Seoul, Korea, pp.510-513, 1995.12.

[10] 강현규, 박세영, 최기선, "자동키워드망과 2단계 문서순위 결정에 의한 자연어 정보검색 모델," 1995년도 제7회 한글 및 한국어 정보처리 학술대회, 한국 정보 과학회, 한국 인지 과학회, 연세 대학교, 서울, 1995. 10.

[11] 김명철, 이운재, 최기선, 김길창, "시소러스 작성을 위한 개념 획득 도구", 한글 및 한국어 정보처리, pp39-50, 1992.

[12] 김정세, 강현규, "동적 정보 검색 시스템," 1995 춘계 학술발표회, 한국 정보 과학회, 조선 대학교, 광주 , pp. 953-956, 1995. 04.

[13] 강현규, 장호욱, 이승률, 박세영, "옥서에서의 표제어와 자연어 검색의 설계 및 구현," 1994가을 학술 발표회, 한국 정보 과학회, 연세대학교, 서울, pp.633-636, 1994. 10.

[14] 한국전자통신연구소, Gigabit 정보통신 시스템 소프트웨어 개발사업, 보고서, 정보통신부, 1995.12.

[15] 정경택, 강현규, 최동시, 장호욱, 전미선, 김만수, 박세영, 정보검색 검증 시스템 개략 설계서, TD95-6240-09, 자연어처리연구실, ETRI, 95.03.

[16] 정인성, "한글로 인터넷을 뒤진다? 한글 검색 엔진", 마이크로소프트웨어, 정보시대, pp. 230-232, 96.05.

[17] 김성훈, "로봇 사용한 한국형 검색엔진 '까치네'", INTERNET, 정보시대, pp. 257-259, 96.03