論文
2-12-13

디지털 도서관 환경에서의 정보 검색을 위한 자연어 문서 및 질의 처리기에 관한 연구 (A Study on Natural Language Document and Query Processor for Information Retrieval in Digital Library)

윤 성 희*
(Sung-Hee Yoon)

요 약

디지털 도서관은 자연어 문서와 멀티미디어 자료에 대한 정보 검색 엔진을 필요로 하는 가장 중요한 데이터베이스 시스템이다. 이 논문은 자연어 처리 기법의 정보 검색 엔진과 브라우저에 대한 설계와 실험 결과를 소개한다. 자연어 문서에 대한 정보 검색 과정은 어휘 분석, 구문 분석, 스테밍, 주제어 색인 등의 계산학적 처리를 포함한다. 많은 이미지와 이미지의 제목, 그리고 자연어로 기술된 설명 문서를 포함하는 실험적인 데이터베이스 'Earth and Space Science'를 통해서 자연어 문서 분석에 기반하는 정보 검색 기능을 실험하였다. 또한 디지털 도서관 환경에서의 멀티미디어 정보 검색 내용 기반의 이미지 검색 엔진과 병행하는 정보 검색 시스템으로서의 가능성을 보여준다.

ABSTRACT

Digital library is the most important database system that needs information retrieval engine for natural language documents and multimedia data. This paper describes the experimental results of information retrieval engine and browser based on natural language processing. It includes lexical analysis, syntax processing, stemming, and keyword indexing for the natural language text. With the experimental database 'Earth and Space Science'that has lots of images and titles and their descriptive text in natural language, text-based search engine was tested. Combined with content-based image search engine, it is expected to be a multimedia information retrieval system in digital library

1. 서론

현대의 급변하는 정보화 사회를 대변하는 가장 큰 현상은 폭발적인 양의 디지털 정보들의 범람과 계속되는 증가 현상이다. 그 정보의 양이 방대함은 물론이고 정보의 형태 또한 다양화되어 문서(text) 위주의 자료들로부터 이미지(image), 오디오(audio), 비디오(video), 음성(voice) 등의 모습으로 혼합되어 가고 있다. 이와 같이 다양하고 방대한 양의 데이터를 매체에 저장하고 사용자로 하여금 원하는 정보를 쉽게, 그리고 정확하게 얻을 수 있는 방법을 연구하는 것이 정보 검색(IR: Information Retrieval)이다. 특히 디지털 도서관(digital library) 시스템은 다량의

* 정회원 : 상명대학교 컴퓨터정보통신과학부 교수

※ 이 논문은 상명대학교 교내연구비 지원의 결과임

논문접수 : 2001. 12. 15.
심사완료 : 2001. 12. 22.

문헌을 집약적으로 관리하고, 다양한 집단에 큰 유익을 줄 수 있는 많은 자료를 제공하기 위해 정보 검색 기법이 가장 중요한 역할로 적용된 대표적인 현장이라고 할 수 있다. 디지털 도서관 시스템은 효율적인 정보 검색 기능에 의해 성공적으로 완성될 수 있으며, 따라서 디지털 도서관과 관련된 연구와 노력의 상당 부분이 정보 검색 기능에 집중되어 있다. 디지털 도서관 자료의 많은 부분은 오랫동안 축적되어온 문헌을 기반으로 하는 자연어로 기술된 문서들이며, 당연히 자연어 문서에 대한 정보 검색이 중요한 부분을 차지하게 된다.

최근의 웹(Web) 기반 환경과 멀티미디어 환경은 디지털 도서관을 비롯하여 방대한 양의 정보를 관리해야 하는 대형 정보 관리 시스템에 대해 멀티미디어 자료에 대한 검색 기능을 요구하고 있다. 과거 문서 자료 위주의 관리와 검색으로부터 멀티미디어 자료(multimedia data)를 통합하여 관리하고 검색하기 위한 시스템이 되도록 요구하고 있는 것이다. 그 중에서도 특히 문서와 문서에 포함된 이미지는 오랫동안 축적되어온 자료의 대표적인 형태이며, 디지털 도서관에서 제공하는 자료의 가장 많은 부분을 차지하고 있다[4][5].

최근의 웹 기반 환경은 디지털 도서관과 그 사용자들에 대해 커다란 변화를 가져오고 있다. 적은 비용으로 방대한 자료를 매우 자유롭게 접근할 기회를 갖게 된 것이다. 그에 따라, 정보 검색 서비스는 사용자가 원하는 정보를 정확하게, 신속하게 제공할 수 있어야 하며, 사용자의 반응에 능동적으로 대처할 수 있는 서비스 시스템이 되기 위해 노력하고 있다[2,7,8].

본 논문은 디지털 도서관 환경에서의 정보 검색을 위해 자연어 문서에 대한 자연어 처리 기법에 의한 정보 검색 방법과 그 실험을 소개한다. 자연어 처리 기반의 정보 검색 과정은 어휘 분석, 개략적인 구문 분석, 불용어 처리, 스테밍, 그리고 추출된 키워드의 색인 등의 계산학적 처리 과정을 포함한다. 또한 디지털 도서관 환경에 대한 실험적 배경으로서 이미지와 이미지의 제목, 그리고 자연어 설명 문서 등을 포함하는 'Earth and Space Science' 데이터베이스를 이용하여 설명 문서를 통한 정보 검색 기능을 실험하고, 디지털 도서관 환경에서의 멀티미디어 검색을 위해 이미지 자체의 특성에 기반하는 내용 기반(content-based) 이미지 검색 기능의 결합 가능성을 살펴본다.

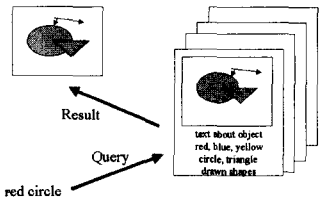
2. 연구 배경 및 목적

본 연구는 디지털 도서관 시스템과 관련하여 대규모 프로젝트(UMDL: University of Michigan Digital Library)를 수행하고 있는 미국 The University of Michigan의 SI와EECS에서 구축된 실험용 데이터베이스를 사용하여 협력 수행된 연구의 일부이다. 멀티미디어 디지털 자료를 갖는 대형 디지털 도서관 시스템을 설계-구현하는 연구로서, 사용자에게 효율적이고 편리한 이미지 검색 및 텍스트 검색 수단을 지원하고자 함이다. 이 연구의 진행은 내용 기반 이미지 검색 방식으로 진행되는 부분, 이미지 설명 문서에 대한 텍스트 기반 검색을 진행하는 부분, 사용자 인터페이스 구축 및 사용자 반응을 수집(user study)하는 세 부분으로 구성되었다. 본 논문의 내용은 이미지의 설명 문서를 자연어 처리 기법 기반으로 검색하는 방법을 중심으로 하고 있다.

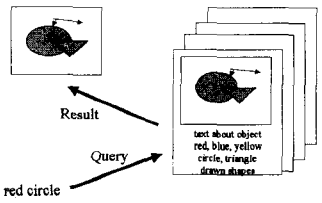
실험을 위한 데이터 베이스는 우주선 및 행성과 관련된 이미지 사진과 제목, 이미지에 대한 자연어 설명 문서들로 구성되어 있다. 현재 'Earth and Space Science'를 주제로 우주선, 행성 및 위성 등에 관련된 사진을 수집하여 약 900개 이상의 디지털 이미지가 실험용 오라클(Oracle) 데이터베이스로 구축되어 있다. 각 디지털 이미지는 이미지 데이터와 함께 자연어로 기술된 제목 및 설명 문서를 포함하고 있으며 텍스트 키워드는 실험 도메인의 특성을 반영하여 mission, subject, object 등 세 가지 카테고리로 구분한다. 여기서 이미지의 설명 문서는 이미지 검색을 위해 이미지 특성에 대해 작위적으로 기술된 문서가 아니며, 디지털 도서관의 자료가 되는 실세계의 각종 이미지 자료들이 자연적으로 포함하게 되는 링크된 설명 문서나 설명 문서에 링크된 이미지 등의 방법으로 구성되어 있는 문서를 말한다. 앞으로도 2000개 이상의 이미지를 수집하고, 보다 다양한 관련 연구를 지원할 수 있을 것으로 기대된다.

이미지의 관련 설명 문서를 고려하지 않을 경우, 디지털 형태로 변환되고 저장되는 이미지 자체에 대한 자료의 표현이나 검색 방식은 문서 기반으로 검

색하는 접근 방법과는 크게 다르다. 이미지 자체에 대한 데이터 표현과 검색을 위한 주된 접근 방식은 이미지 자체에 대한 색상(color), 질감(texture), 형태(shape) 등에 대한 이진적(binary) 계산에 의해 검색하는 내용 기반 검색 방식(content-based retrieval)이다[2]. <그림 1>은 이미지와 관련된 설명 문서를 이용하는 검색 방법과 이미지 특성 자체에 데이터 계산을 통해 유사 이미지를 검색하는 방법을 비교하고 있다.



자연어 설명 문서를 통한



이미지 검색 내용 기반 이미지 검색

[그림1] 자연어 설명 문서를 이용하는검색과 내용기반 이미지 검색

[Fig. 1] Natural language text-based retrieval and content-based image retrieval

그러나 사용자에게는 원하는 이미지를 색상과 질감, 형태와 같은 low-level feature 표현 방식으로 질의(image content query)하도록 하는 검색 방식은 질의 자체가 사용자에게 직관적이지 못한 방법으로 평가되고 있다. 사용자는 주로 자신이 찾는 이미지를 질의하기 위해 문서를 설명할 수 있는 키워드(keyword)를 입력(text query)하는 high-level의 질의에 익숙하다. 이러한 환경에서 볼 때, 검색의 대상이 되는 많은 디지털 이미지들은 이미지만 독립적으로 존재하기보다는 이미지에 대한 제목이나 관련 설명 등의 문서 자료를 동반하여 제공되고 있다는 사실은 매우 고무적이다[3]. 웹 환경에서도 많은 하이퍼텍스트는 문서의 내용과 직접적으로 관련되는 이미지를 많이 포함하고 있다. 반대로 이미지에 링크된 또다른 문서 페이지를 가질 수 있다. 특수하게 인공적 문법으로 표현된 설명이 아닌 자연어로 기술된 설명 문서들을 검색에 이용하기 위해서는 자연어 처리 기술이 적용되어야 한다. 이처럼 저장된 이미지에 관련된 문서 정보가 있다면 이 문서 정보를 이미지 검색을 위해 활용하면 이미지 검색 시스템 측면에서는 low-level 데이터 표현과 관리의 한계를 극복하기 위한 방법을 얻을 수 있고, 사용자에게 또 다른 방식의 이미지 검색 인터페이스(interface)를 제공할 수 있다.

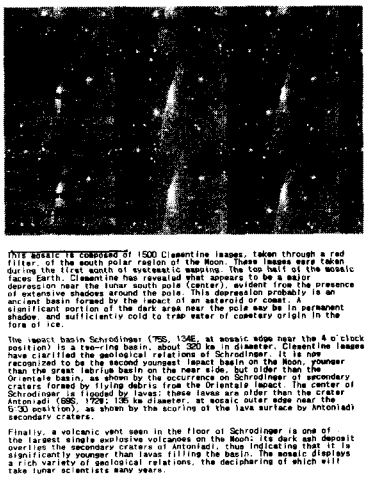
3. 이미지와 설명문서의 데이터베이스

본 연구는 디지털 도서관의 이미지 자료 관리와 검색을 위해 이미지와 동반되는 설명 텍스트를 이용한 검색을 위해 자연어 처리(natural language processing) 기술로 접근한다. 실험의 대상이 되는 이미지는 우주-항공 관련 사진과 사진의 제목 및 설명을 포함하는 1500여 이미지-텍스트 데이터베이스이다. Oracle database에 연구용으로 수집되고 구축된 이미지 데이터베이스는 디지털 이미지와 함께 영어 자연어(natural language)로 기술된 제목(title)과 관련 설명(description)을 포함한다. 이미지 검색을 위해 텍스트 데이터를 분석하고 주제어를 색인(indexing)하여 사용자가 원하는 이미지에 접근하는 방법을 취한다. 이미지 검색을 위한 텍스트 키워드는 다음과 같은 세가지 카테고리로 분류함으로써 사용자의 입력을 도울 수 있다.

Object
(예: deep space, visitors from space, man-made craft 등)
Subject
(예: solar radiation, Polar regions, Solar system 등)
Mission
(예: Apollo, Gemini, Mars 3 등)

다음의 <그림 2>은 데이터베이스에 저장된 이미지와 관련된 자연어 설명 문서의 예이다. 그림과 같

은 검색 결과 이미지는 오른쪽과 같은 자연어 설명 문서를 이미지와 함께 연결하고 있다. 이와 같은 설명 텍스트는 키워드 추출을 위해 어휘분석, 개략적 파싱(parsing)과 키워드 추출 및 색인(indexing)과 같은 자연어 처리 계산 과정을 거치게 된다.



This mosaic is composed of 1500 Clementine images, taken through a red filter, of the south polar region of the Moon. These images were taken during the first month of systematic mapping. The top half of the mosaic faces Earth. Clementine has revealed what appears to be a major depression near the lunar south pole (center), evident from the presence of extensive shadows around the pole. This depression probably is an ancient basin formed by the impact of an asteroid or comet. A significant portion of the dark area near the pole may be in permanent shadow, and sufficiently cold to trap water of cometary origin in the form of ice.

The impact basin Schrodinger (75S, 134E, at mosaic edge near the 4 o'clock position) is a two-ring basin, about 320 km in diameter. Clementine images have clarified the geological relations of Schrodinger. It is now recognized to be the second youngest impact basin on the Moon, younger than the great Iabrija basin on the near side, but older than the Orientale basin, as shown by the occurrence on Schrodinger of secondary craters formed by flying debris from the Orientale impact. The center of Schrodinger is flooded by lavas; these lavas are older than the crater Antoniadi (69S, 172E; 136 km diameter, at mosaic outer edge near the 6:30 position), as shown by the scoring of the lava surface by Antoniadi secondary craters.

Finally, a volcanic vent seen in the floor of Schrodinger is one of the largest single explosive volcanoes on the Moon; its dark ash deposit overlies the secondary craters of Antoniadi, thus indicating that it is significantly younger than lavas filling the basin. The mosaic displays a rich variety of geological relations, the deciphering of which will take lunar scientists many years.

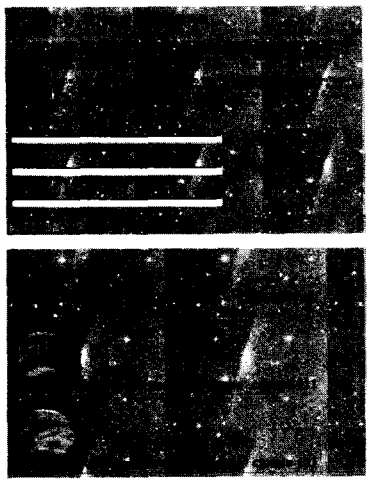
[그림 2] 이미지와 설명 문서의 예

[Fig. 2] An example of image with natural language description

위 이미지에 연결된 설명 텍스트의 내용은 다음과 같다.

" Mosaic View of the Lunar South Pole
This mosaic is composed of 1500 Clementine images, taken through a red filter, of the south polar region of the Moon. These images were taken during the first month of systematic mapping. The top half of the mosaic faces Earth. Clementine has revealed what appears to be a major depression near the lunar south pole (center), evident from the presence of extensive shadows around the pole."

<그림 3>은 키워드를 입력하여 사용자가 찾기를 원하는 이미지를 표현하는 화면과 검색 결과 이미지를 열거하는 화면이다.



[그림 3] 검색 키워드 입력과 설명 문서를 분석한 이미지 검색의 예

[Fig. 3] Keyword input and results of image retrieval based on text analysis

설명 문서에 대한 자연어 처리 과정의 중요한 목적은 효율적인 검색을 위한 색인을 자동적 구축하는 것이다. 색인은 방대한 양의 정보원으로부터 가장 유사한 내용의 정보 자료만을 선별해주는 역할을 한다. 즉, 정보를 탐색할 때 색인은 방대한 양의 정보로부터 사용자가 원하는 정보만을 걸러주는 여과기와 같은 구실을 하는 것이다.

사람에 의해 선별되고 통제되는 수동 색인(manual indexing)을 통해 색인을 만드는 과정은 해당 분야 전문가(subject expert)가 정보 자료에 포함되어 있는 중요한 개념을 추출한 후 적절한 코드, 즉 색인어로 변환시키는 방법이다. 수작업에 의한 색인 과정에서는 작업의 효율을 높이기 위하여 색인어를 통제하는 것이 보통이며, 동의어나 단수/복수형 통제 등을 통해 특정 개념은 항상 하나의 단어로 나타내고 동일한 색인어로 검색할 수 있다[1].

반면, 컴퓨터를 이용하여 자동 색인(automatic indexing)을 하는 경우에는 컴퓨터로 자료를 분석한 후 중요한 개념을 나타내는 단어를 일정한 기준에 따라 자동 추출한 다음 이를 색인어로 채택한다. 자동 색인의 경우에는 색인어를 사람이 직접 통제하지 않기 때문에 텍스트에 나타난 단어가 그대로 색인어로 채택될 수 있다. 따라서 자동 색인의 경우에는

검색의 효율을 높이기 위해 동의어 사전을 사용하거나 용어 절단법(term truncation)이나 용어 집단화 기법(term clustering) 등을 적용되기도 한다.

본 논문에서 제시하는 검색 방법은 자연어 처리 기법에 기반하여 유효한 색인어를 추출하는 자동 색인 방법이다. 자동 색인 기법에 대한 실험에 의하면 빈도수가 매우 높은 단어는 너무나 일반적이어서 주제어로서의 가치가 전혀 없으며, 반대로 빈도수가 지내치게 낮은 단어는 주제어로서 채택될 가능성이 거의 없다고 보인다. 따라서, 주어진 문서를 분석하여 단어의 사용 빈도수를 얻은 다음 최고 한계 빈도수를 넘거나 최저 한계 빈도수에 미치지 못하는 단어는 제외하고 이 두 한계 빈도수 사이에 속하는 단어들로부터 색인어를 선정하는 방법이 제안되기도 한다[1][7].

3.1 설명 문서에 대한 자연어 처리

본 연구에서의 자연어 문서 기반의 검색 방법은 자연어 처리 기술을 적용하여 어휘 분석, 구문 구조를 얻기 위한 개략적인 파싱, 색인 대상이 되지 못하는 불용어(stopword) 제거, 단어를 문법적인 원형으로 변환하는 스테밍(stemming), 색인이 될 주제어 선정의 과정을 포함한다.

- **가. 어휘분석(lexical analysis)** : 문장의 단어 분리, 문장 부호 분리, 숫자 및 특수 기호처리 등을 위한 과정이다.
- **나. 개략적 부분 파싱(parsing)** : 각 단어의 구문적 기능을 파악할 수 있도록 문장의 구문 구조를 분석한다.
- **다. 불용어(stopword) 제거** : 변별력이 매우 낮아서 검색에서 역할을 하지 못하는 단어들을 색인 대상에서 제외한다.
- **라. 스테밍(stemming)** : 단어들에 대해 구문적 변형 형태를 기본형으로 복원하여 기본형을 색인어로 취하되 변형 형태에 대한 검색이 가능하도록 한다. 예를 들면, "deploy"와 "deployed", 또, "land"와 "landing"등에 대한 기본형 복원 처리다.
- **마. 색인으로 사용할 단어/스템(stem)의 선정** : 그 단어의 구문적 성질을 중심으로 색인어를 지정한다. 일반적으로 형용사, 부사, 동사보다 명사가 색인어로서 중요한 의미를 갖게 되며, 복합 명사도 색인어로 채택된다.

3.2 개략적 파싱과 불용어 제거

자연어 처리 기술을 적용하여 어휘 수준의 분석, 문장 수준의 분석, 키워드 추출의 과정을 거쳐서 색인어를 선정하는 것이 문서 기반 검색 과정의 가장 중요한 부분이다. 컴퓨터에 의한 자동 색인 방법은 색인어를 선정하기 위해 단어의 출현 빈도수에 근거하여 주제어로서의 중요도를 측정하는 통계적 기법을 사용할 수도 있다. 현재로서는 색인어를 선정하기위한 신뢰성있는 통계 자료를 얻기 어려우므로, 본 연구에서는 언어학적 처리 과정을 통한 접근 방법을 택하고 있다.

형태소 분석 과정의 결과로부터 주제를 나타내는 단어나 구를 선택하는 간단한 방법이 적용될 수도 있으나, 보다 양질의 의미있는 주제어를 선별하기위해 구문 분석 수준의 처리과정을 거치는 것이 타당하다. 자연어 문장의 구문 구조를 분석 과정인 파싱(parsing)의 결과를 얻어 주어부, 서술부, 수식을 위한 전치사구 등을 구분하고 가장 중요한 역할을 하는 단일어 또는 복합어를 색인어로 선택하는 방법이다. 실제로 자연어 문장에 대한 완벽한 구문 분석은 모호성(ambiguity)의 처리 및 의미적인 분석(semantic analysis)을 포함하여 매우 어려운 과정임이 충분히 알려져 있다. 그러나, 정보 검색을 위한 자동 색인을 하는 경우에는 그러한 복잡하고 어려운 처리에 비해 구문 분석의 효과가 그다지 크지 못하다는 것이 여러 실험의 결과이다. 따라서 본 연구는 문장에 대한 비교적 단순한 통사적 수준의 구문 분석 과정을 거치도록 하며, 또한 파싱에 대한 결과가 얻어지지 못하는 경우에도 처리 효율을 높이기 위해 부분 파싱의 결과를 색인어 추출에 이용하고 있다. 구문 분석을 위해 확장된 문맥 자유 형식의 문법(context-free grammar)을 이용하고 있다.

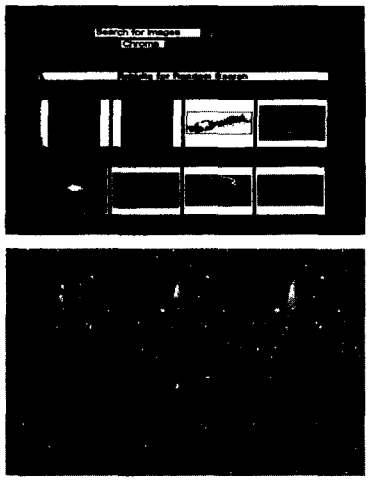
불용어 제거 과정은 자연어 문서에서 매우 빈번하게 출현하지만 색인어로서 지정될 가능성이 거의 없는 기능어들을 제외시키는 과정이다. 영어에서 'the', 'a/an', 'to'와 같은 단어들로서 불용어(stopword)들로서 자연어 문장에서 주제를 전달하지 않는 기능적 단어들이므로 검색 대상으로서 무시할 수 있다. 설명 문서의

분석으로부터 추출된 색인어는 가장 일반적인 방식인 역파일(inverted file) 구조로 저장되어, 색인어와 출현 빈도, 출현위치(이미지 또는 문서)를 저장한다. 이미지에 대한 제목(title)에 출현하는 단어들은 이미지 설명에 대한 주제어로서의 가치가 높기 때문에 설명 문서에 출현하는 단어들보다 높은 가중치를 부여할 수 있다. 예를 들면, "Mosaic view of the Lunar South Pole"이나 "Launch of Apollo 17 Lunar Landing Mission"등이 이미지에 대한 제목이다.

4. 내용 및 설명 문서 기반의 이미지 검색

설명 문서에 대한 자연어 분석과 색인을 통하지 않고 이미지 자체의 데이터 특성을 표현하고 계산하여 검색하는 방식은 내용기반 검색 방법이라고 구분된다. 색상(color), 질감(texture), 형태(shape)의 세 가지 특징을 디지털 자료로 계산하고 비교하는 방법을 취한다. 이미지 자체의 특성을 사용자들이 질의로 입력하기 어렵기 때문에 일반적으로 제시된 임의의 예제 이미지로부터 사용자가 원하는 이미지와 가장 유사한 이미지를 주어진 예제 이미지들로부터 선택하는 방법을 반복(relevance feedback)함으로써 목표 이미지에 접근하는 예제기반(example-based) 검색 방법이 채택된다[9][10]. 이미지의 내용에 기반하는 내용 기반 검색 기법에 대한 자세한 내용은 본 논문의 주제에 포함되지 않으므로 간략한 소개로 대신하기로 한다.

사용자가 반응하는 이미지의 유사한 정도는 컴퓨터로 계산된 방식의 유사함과는 크게 다를 수 있다. 비교 병행 연구를 위해 같은 데이터베이스를 대상으로 실험된 내용 기반 이미지 검색 시스템은 각각 PSIs(Perceptually Similar Images)와CSIs(Computationally Similar Images)로 구분하여 이와 같은 두 가지 측면의 이미지 유사도(similarity)를 판단한다. 색상, 질감, 형태 등 이미지 내용에 기반하여 컴퓨터로 계산된 유사한 이미지 CSI를 사용자에게 예제로 제공하게 되고, 사용자가 유사하다고 판단하여 선택하는 이미지는 PSI로서 재입력된다. 사용자의 반복적인 이미지 선택 과정은 가장 유사한 이미지를 검색하기 위해 다중모드(multi-mode)로 관련성 피드백(relevant feedback) 과정을 갖는다. <그림 4> 검색 시스템에 의해 제시된 예제 이미지로부터 사용자의 응답을 얻어 제시된 이미지의 예들을 보이는 화면이다.



[그림 4] 예제 이미지 선택과 검색된 이미지 집합
[Fig. 4] Example images and set of retrieved images

본 연구의 발전 방향은 설명 문서 기반 및 내용 기반 이미지 검색 엔진의 상호 지원 기법을 도입하는 것으로서 high-level 문서 정보와 low-level의 이미지 특성을 동시에 사용하고, 양방향의 기술을 효과적으로 결합하고자 하는 것이다. 이미지 검색 엔진에 대한 비전문 일반 사용자의 평가를 수렴한 결과로는 사용자들은 찾고자 하는 목표 이미지의 계산된 low-level feature에 대해서 질의를 입력할 만큼 충분히 이해하기 어려우며, 그 결과로 자신이 가장 유사하다고 판단한 이미지가 컴퓨터가 색상, 질감, 형태 등에 대한 계산을 통해 가장 유사하다고 판단하는 이미지와 매우 다르다는 것을 실감하게 된다는 것이다.

디지털 이미지 검색을 위한 내용 기반 검색 엔진(content-based search engine) 은 데이터베이스를 환경으로 관련성 피드백(relevance feedback) 방식을 채택하여 사용자가 원하는 이미지와 유사한 이미지를 점진적으로 접근해가는 과정을 반복한다[2][9]. 예제 이미지 자체의 색상, 질감 및 형태 정보를 계산하고 그 값을 이용하는 이와 같은 검색 엔진은 텍스트 정보를 활용하지 않는 이미지 검색 엔진으로서 그 자체만으로는 독립적으로 실용적인 성능을 기대하기 어렵다. 또한 내용 기반 이미지 검색 과정은 이미지

데이터에 대한 많은 계산량으로 인해 계산 시간 측면에서도 만족할 만한 효율의 검색 결과를 얻기가 어렵다.

결과적으로 비록 텍스트 기반의 검색 결과가 완전하게 만족스럽지 못하더라도 사용자들의 대부분은 검색 과정에서 그들에게 매우 익숙한 텍스트 방식의 검색어나 질의를 입력하여 원하는 이미지를 검색하기 위한 high-level feature로서 사용할 수 있기를 원한다. 따라서 현재의 이미지 검색 방식의 실용성 한계를 극복하기 위해 이미지의 관련된 설명 문서를 효과적으로 활용하는 문서 기반의 검색 방식을 동시에 적용하는 새로운 접근을 모색하고자 한다. 본 실험의 궁극적 방향인 디지털 이미지에 대한 내용 기반 검색과 텍스트 기반 검색의 병행 검색 시스템은 사용자에게 매우 융통적인 인터페이스를 제공할 수 있다고 기대된다. 디지털 도서관 환경에서 설명 문서가 동반되거나 링크된 이미지나 하이퍼텍스트에서 이미지에 링크된 문서가 존재하는 경우에 대해 모두 적용될 수 있다.

5. 결론

본 논문에서는 디지털 도서관 환경에서의 멀티미디어 정보 검색을 위해 이미지에 연결된 텍스트 자료, 즉 이미지에 대한 제목과 이미지 설명 문서를 중심으로 자연어 처리 접근 방식을 도입하여 이미지에 검색에 가장 핵심적인 키워드를 추출하고 색인을 관리하는 설명문서 기반 이미지 검색 방법과 실험 배경을 중심으로 소개하였다. 본 연구와 실험의 궁극적인 목적은 일반적인 사용자에게 직관적이고 익숙한 텍스트 입력 방식을 이미지의 내용 기반 검색 방법과 함께 검색 엔진으로 통합하고자 함이다. 즉, 이미지의 내용과 관련된 설명 문서의 검색을 동시에 적용하는 혼합 방식의 효율적인 검색 엔진(search engine) 및 브라우저(browser)를 개발하고자 하는 것이다. 이를 위해 위에 기술한 두 가지 기초 연구인 이미지 내용 기반 검색과 자연어 처리 방법을 적용한 설명문서 기반 검색 기술이 상호 협력하여 보다 향상된 효율의 시스템을 얻을 수 있기를 기대한다. 이미지 정보의 검색 키(key)를 색상, 질감, 형태 등의 low-level features를 기초로 하는 내용기반 검색 방식과 이미지에 링크된 high-level features인 일반 텍스트 분석에 의한 검색을 병행함으로써 사용자 인터페이스로서의 기능을 높이고 검색의 효율도 높일 수 있는 동시 효과를 얻는다. 디지털 도서관 정보 검색 환경 뿐만 아니라 웹 정보 검색, 디자인, 광고, 예술 분야 등 이미지를 검색 대상으로 하는 여러 분야에서 디지털 환경에 적응하기위해 대량의 이미지로부터 빠른 시간에 손쉽게 목표 이미지를 찾는 계산 과정은 가장 핵심적인 연구 부문이다.

국내에서도한국어 정보처리 분야의 연구에서 문서 검색에 관한 연구가 활발히 진행되고 있다. 자연어 문서로부터 주제어 추출과 인덱싱(indexing) 방법에 관한 연구가 주된 흐름이다. 내용 기반의 이미지 검색 또한 많은 연구가 이루어지고 있는데, 많은 현실 자료가 멀티미디어 데이터베이스로 제공되는 디지털 도서관 환경에서 정보의 검색은 이미지와 문서 정보를 분류할 수 없다. 본 연구의 방법과 실험은 국외 연구에 代한 협력 연구의 과정에서 영어 자연어 문서의 처리를 그 대상으로 하였으나, 한국어 환경을 위해 자연어 처리 및 한국어 정보 처리 측면에서 한국어 자연어로 기술된 이미지 설명 문서를 분석하고 검색하는 연구와 실험으로 이어질 계획이다.

※ 참고문헌

[1] 김영택, "자연어처리", 생능출판사, 2001.

[2] 노형기, 황본우, 문종섭, 이성환, "내용 기반 영상 정보 검색 기술의 현황", 전자공학회지, 제25권, 제8호, 1998.

[3] 박명선, 송병호, 이석호, "WISE : WWW 이미지 검색 엔진", 정보과학회논문지(C), 제4권, 제3호, 1998.

[4] 박선영, 강주영, 이인기, 용환승, "자연사 박물관 멀티미디어 웹 컨텐트 검색 시스템 구축", 정보관리학회지, 제16권, 제3호, 1999.

[5] 유성준, "멀티미디어 정보 검색 기술 동향", 전자공학회지, 제 25권, 제8호, 1998.

[6] 이해민, 김명호, "디지털 도서관 환경에서 일관성과 최근성을 고려한 메타데이타 관리 기법", 정보과학회논문지, 제27권, 제1호, 2000.

[7] Baeza Yates, Ribeiro Neto, "Modern Information Retireval", Addison Wesley Longman Inc.

[8] Carson, Chad, et al., "Region-based Image Querying[Blobworld]," Workshop on Content-based Access of Image and Video Libraries, Puertorico, June 1997.

[9] J. R. Smith and S. F. Chang, "Visually searching the web for content,"IEEE Multimedia Magazine, vol. 4, no.3, pp.12-20, Summer 1997, also Columbia U. CU/CTR Techical Report 459-96-25.

[10] J. R. Smith and S. F. Chang, "VisualSEEK: A Fully Automated Content-Based Image Query System,"ACM Multimedia Conference, Boston, MA, Nov. 1996.

[11] Karen S. Joes and Peter Willett, " Readings in Information Retireval", Morgan Kaufmann.

[12] W.I.Grosky and Y. Tao, "Multimedia Data Mining and Its Implications for Query Processing," Proceedings of the 9th International Workshop on Database and Expert Systems Applications, Vienna, Austria, 1998.

[13] Y. Rui, T. S. Huang, and S-F. Chang, "Past, Present, and Future, Journal of Visual Communication and Image Representation

윤 성 희

1987년 서울대학교 컴퓨터공학과 졸업
1989년 서울대학교 대학원 컴퓨터공학과 석사
1993년 서울대학교 대학원 컴퓨터공학과 박사
1994년-현재 상명대학교 컴퓨터정보통신학부 교수
관심분야 : 자연어처리 및 한국어정보처리, 기계번역, 멀티미디어 정보검색 등