영상검색엔진을 위한가중치 N-Gram색인 방법

이상열*, 정성호**, 황병곤*
*대구대학교 컴퓨터정보공학과
**포항1대학 컴퓨터정보처리

Weighted N-Gram Indexing for Image Search Engine

Sang-youl Lee*, Sung-Ho Jung**, Byung-kon Hwang*
*Dept. of Computer & Information Engineering, Taegu University
**Dept. of Computer & Information Processing, Pohang College

요 약

멀티미디어 검색 시스템들은 아직까지 내용 기반에 의한 검색기술이 실용적으로 쓰일 만큼 높은 성능을 보이고 있지 않기 때문에 텍스트에 의한 검색만을 지원하고 있는 실정이다. HTML 문서에 나타나는 텍스트 중 이미지 아래에 붙은 표제나 이미지 링크에 붙어 있는 텍스트를 골라내어 이미지의 색인 정보로 이용하여 텍스트를 추출하는 기법을 제안하였다. 텍스트를 추출하기위해 N-Gram 색인 방법을 사용하였으며 한편 검색 효율을 높이기위해서 질의 의도가 큰 단어에 가중치를 부여하였다.

1. 서론

최근 컴퓨터 기술과 통신망 기술의 발전으로 인하여 정지 화상(Image), 음성 자료(Audio data), 동화상 자료(Video data) 등의 멀티미디어 자료가 대량으로 발생되고 있다. 이에 따라 전자 신문, 홈 쇼핑 등과 같은 새로운 멀티미디어 정보 서비스가 인터넷을 통하여 확산되고 있고, 문헌 정보, 인물 정보, 법률 정보 등의 정보 검색 시스템에서도 기존의 텍스트 위주의 정보 검색 서비스에서 멀티미디어 정보 검색 서비스로 급격히 변화되고 있다. 멀티미디어 검색은 멀티미디어의 내용을 파악하여 검색하는 내용기반 검색과 멀티미디어를 설명한 텍스트를 이용하여 텍스트 기반 검색이 있다.[2,5]

본 논문은 웹 문서에서 멀티미디어를 설명하는 텍스트를 자동으로 추출하고, 추출된 텍스트를 이용하여 자동으로 분류하여 데이터베이스에 저장하는 웹 멀티미디어 검색엔진을 제안한다. 제안되는 웹 멀티미디어 검색엔진은 설명하는 텍스트뿐만 아니라 파일명에 기술된 확장자를 이용하여 멀티미디어 종류를 분류한다.

본 논문의 구성으로는 2장에서는 관련연구로써 웹 멀티미디어의 기존의 색인방법에 대하여 설명하고, 3장에서는 N-Gram기반의 색인방법에 대하여 설명한다.[8,9] 4장에서 시스템 구조와 실험 결과에 대하여 설명하고 마지막으로 5장에서 본 논문의 결론과 개선점, 향후 연구 과제를 제시한다.

2. 관련 연구

웹상에서 이미지를 검색하는 시스템이 최근 다양하게 연구 되어 왔으며 Yahoo의 Image surfer는 키워드로 분류 트리(Category tree)를 생성하고 이를 이용하여 웹에서 이미지를 검색한다.[1] 이 시스템은 이미지에 관련된 텍스트를 반자동으로 분류하여 데이터베이스에 저장한다. 이 시스템은 색인 검색뿐만 아니라 색상 히스토그램으로 이미지를 검색할 수 있다. Lycos의 미디어 검색 툴은 이미지 URL과 이미지가 포함된 웹 문서에서 키워드를 자동으로 추출한다. WebSeek 시스템은 웹 이미지 내용을 분석하여 이미지 헤더, 파일 종류, 크기, 날짜 등과 같은 이미지에 연결된 텍스트 정보를 검색 키워드로 사용한다. 또한 이 시스템은 사람의 얼굴이나 수평선 등의 객체를 이미지에서 자동으로 인식하고, 색상과 질감 등의 정보를 사용하여 하나의 이미지를 여러 조각으로 분할한다. ImageRover는 이미지 수집, 해석을 담당하는 이미지 수집 시스템과, 질의 서버와 사용자 인터페이스로 이루어진 이미지 질의 서버시스템으로 이루어진다.[4] WISE는 이미지 설명 텍스트와 이미지 색상 히

스토그램을 이용하여 이미지 검색을 지원하는 시스템이다. 이미지 설명 텍스트는 이미지 표제, 이미지 주위의 텍스트, 링크, Alt 텍스트 등을 추출하여 구성하며, 사용자가 AND 또는 OR로 연결한 키워드를 이용하여 검색한다.[7]

본 논문에서는 웹 문서 탐색 에이전트와 탐색된 URL을 분석하여 이미지를 수집하여 키워드별로 분류하여 데이터베이스에 저장하는 이미지 수집시스템과 여기에서 파생되어 나온 키워드로 검색하거나 이미지 색상 히스토그램을 이용하여 이미지 검색하는 검색 시스템으로 구성되어 있다. 질의자가 키워드와 색상 히스토그램을 입력하면 검색엔진은 가장 근접한 키워드에 해당하는 멀티미디어 정보를 제공한다.

3. 멀티미디어 자료의 색인 방법

웹 기술은 다양한 인터넷 서비스를 제공하기 위해서 어절 단위 색인 법에서의 복합명사 띄어쓰기 문제를 완화할 수 있으며, 형태소 단위 해석에서와 같은 복잡한 문장 해석 규칙이나 언어 정보의 개발을 요구하지 않는 *N*-Gram 방법을 제안한다. 표1은 제안하는 방법의 색인 과정을 간략히 보여주며, 각 단계에 대한 자세한 설명은 다음과 같다.

표1 *N*-Gram 기반의 색인 과정

① 문서의 모든 어절들을 추출한다.
② 불용어를 삭제한다.
③ 각 어절에서 비색인 분절들을 삭제한다.
④ 나머지 색인 분절을 *N*-Gram들로 분할하여 색인어로 설정한다.
⑤ 가중치를 설정한다.

1) 문서의 내용을 색인하기 위해 먼저 빈칸, 마침표, 쉼표, 따옴표 등을 구분자로 하여 모든 어절들을 추출한다.

2) 불용어 리스트를 이용하여 색인어로서 무의미한 어절들을 삭제한다.

3) 나머지 어절들에 대해 비색인 분절을 삭제한다. 비색인 분절은 단일 조사(-가, -이, -를, -으로, -부터), 복합조사(-으로부터, -에서부터), 어미, 접미사 등이 결합된 다양한 형태의 음절 들을 포함한다. 예를 들면, 다음과 같은 어절들에서 "색인" 뒤에 오는 모든 문자열이 여기에 포함된다.

색인*을*	색인*하여*	색인*하였는데*
색인*되어*	색인*되었으니*	색인*임을*
색인이기에	색인*이라고*	색인*이지만*

4) 생성된 각각의 색인 분절에 대해 *N*-Gram방법을 적용한다. *N*-Gram방법이란 인접한 *N*개의 음절을 말한다. 예를 들면 "정보검색서비스"이란 어절에 대해 2-Gram은 "정보", "보검", "검색", "색서", "서비", "비스"이며 3-Gram은 "정보검", "보검색", "검색서", "색서비", "서비스"이다. 색인 분절의 음수가 *N*보다 큰 경우에는 색인 분절의 여러 개의 *N*-Gram들로 분리하고, 작은 경우에는 색인 분절 전체를 하나의 *N*-Gram으로 취한다.

표2는 제안하는 방법을 이용한 색인 과정을 보여준다.

표2 *N*-Gram 기반의 색인 방법의 예

내년 중반부터 정보검색서비스가 실시된다.
① 문장내의 어절 인식
내년, 중반부터, 정보검색서비스가, 실시된다.
② 불용어 제거
정보검색서비스가, 실시된다.
③ 비색인 분절의 절단
정보검색서비스, 실시
④ n-gram의 적용
정보, 보검, 검색, 색서, 서비, 비스, 실시

5) 의미 없는 *N*-Gram의 생성으로 인해 질의에 부적합한 문서들이 검색된 가능성이 있다. 예를 들면 아래와 같은 문서는 같은 의미의 문서로 나타나게 된다.

① 유기효율
② 소기효율

{유기, 기효, 효율}과 {소기, 기효, 효율}의 색인어를 형성한다. 여기서 '기효'와 '효율'이 60% 일치하므로 ①과②는 서로 연관성이 있는 단어이므로 검색결과에 출력될 것이다. 이러한 현상을 제거하기 위하여 각각의 단어에 가중치를 질의자가 부여한다. '유기'와 '효율'의 단어에 가중치를 50% 이상을 두어서 검색을 하면 '기효'란 단어는 일치하더라도 가중치가 적기 때문에 두개의 단어는 유사성이 떨어지게 된다.

4. 시스템 구조와 실험결과

웹 멀티미디어 검색엔진은 크게 웹 문서 탐색 에이전트와 멀티미디어 수집기, 멀티미디어 검색엔진, URL 과 이미지를 저장하는 데이터베이스로 구성된다. 웹 멀티미디어 검색엔진의 기본 구조는 그림 1과 같다.

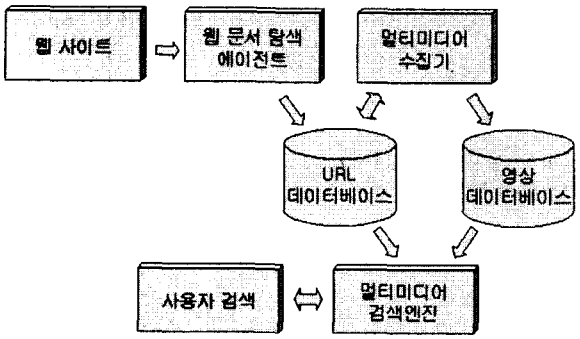

그림 1. 웹 멀티미디어 검색엔진의 기본 구조

1) 웹 문서 탐색 에이전트

문서 탐색 에이전트는 주어진 웹 사이트를 방문하여 멀티미디어 자료를 수집하는 역할을 한다. 수집하는 방법은 너비 우선 탐색을 먼저 수행하고 사용자의 미리 정해 놓은 깊이에 따라 깊이 우선 탐색을 한다. 깊이 우선 탬색은 사용자가 정하지 않으면 계속 연결되는 링크를 찾다보면 시스템이 과부하 또는 무한 루프에 빠질 수 있기 때문이다. 본 논문에서는 웹 문서의 URL을 수집하여 탐색 에이전트 모듈을 그림2와 같이 구성하였다.

(1) 초기 사용자가 지정한 웹 사이트를 탐색한다.
(2) 웹 사이트를 방문하여 연결된 모든 웹 문서를 다운로드한다.
(3) 다운로드한 웹 문서는 이미 방문 했거나 중복된 것을 제거하고 URL저장소에 저장한다.
(4) 사용자가 선택한 웹 문서를 저장 다음 연결된 문서를 분석한다.
(5) 연결된 문서에서 사용자가 지정한 카운트를 벗어나지 않는 범위에서 (2)에서 (4)까지 반복하여 깊이 우선 탐색을 한다.

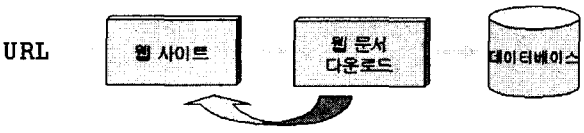

그림 2. 웹 문서 탐색 에이전트의 모듈 구성

2) 멀티미디어 수집

멀티미디어 수집은 그림 3와 같이 멀티미디어 다운로드, 설명추출, 색인, 다운로드한 자료체크로 구성되어 있다.

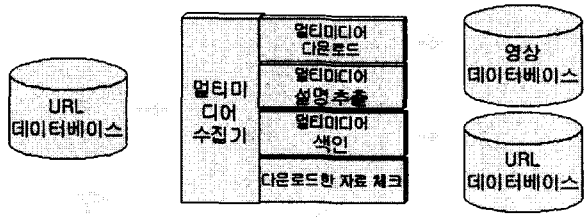

그림 3. 멀티미디어 수집

(1) 내용 구성

① 데이터베이스에 등록된 URL을 참조하여 웹 문서에 있는 멀티미디어의 자료를 다운로드 한다.
② 정상적으로 다운로드 되면 멀티미디어를 설명하고 있는 자료와 파일명을 이용하여 색인을 한다.
③ 다운로드가 진행되지 않으면 URL주소가 잘못되었는 것으로 판단하여 URL 데이터베이스에서 오류 표시를 한다.
④ 색인작업이 끝나면 이미지와 색인된 내용을 데이터베이스에 저장한다.
⑤ URL 데이터베이스에 정상적으로 멀티미디어 자료가 다운 완료했다는 표시와 시각을 기록한다.

(2) 멀티미디어 자료 수집

① 다운로드한 웹 문서를 분석하여 연결된 웹 문서의 멀티미디어 자료를 찾아낸다.
② 웹 문서의 종류는 MIME헤더를 조사한다.
content-type:text/html 이면 HTML
image/jpeg 또는 image/gif 이면 이미지
video/mpeg 이면 동영상이다.
③ 다운로드 한 후 파일명의 확장자에 맞게 이미지와 동영상을 구분하여 데이터베이스에 저장한다.

(3) 멀티미디어 설명 추출

① 멀티미디어 파일 옆에 있는 텍스트를 분석한다.
② 하나의 웹 페이지에 여러 멀티미디어가 들어 있는 프레임과 그 이미지의 링크를 가지고 있는 프레임으로 구성되어 있는 경우는 깊이 우선 탐색하기위해 URL사이트를 URL 데이터베이스에 저장한다.
③ 일반적으로 멀티미디어 링크에 덧붙인 텍스트

<img src="..."> 에 나타나는 alt="..."에서 단어를 추출한다.

⑦ 추출된 인덱스 자료를 N-Gram방법을 이용하여 색인 정보를 만들어 영상데이터베이스에 저장한다.

멀티미디어 설명 추출하는 모듈은 그림4와 같이 구성하였다.

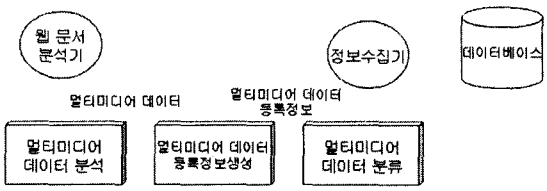

그림 4 멀티미디어 설명 추출기

3) 멀티미디어 검색엔진

사용자가 멀티미디어를 질의 방법은 다음과 같다. 이미지의 파일 이름 또는 설명 텍스트를 각각 AND나 OR로 연결한 후 파일의 확장자를 입력한다. 이때 검색하고자 하는 단어의 가중치를 사용자가 입력하게 되어 N-Gram의 단점을 어느 정도 해결하였다. 출력된 그림은 큰 그림을 왼쪽에 한개 보여주고 오른쪽 화면에는 5개의 작은 그림으로 나타냈다. 화면 밑에 "이전"과 "다음"의 버튼을 만들어 여러 그림을 찾을 수 있도록 했다. 또한 질의자가 선택한 그림을 다운받을때 카운트를 하여 가장 많은 카운트를 받은 멀티미디어 자료를 화면에 먼저 나타나도록 하였다.

5. 결론

현재 멀티미디어 검색 시스템들은 대부분이 텍스트에 의한 검색만을 지원하고 있는 실정이다. 이는 사용자들이 그동안 텍스트를 이용한 검색 방법에 익숙해져 있기 때문으로 볼 수 있지만 아직 내용 기반에 의한 검색 기술이 실용적으로 쓰일 만큼 높은 성능을 보이고 있지 않기 때문이다. 본 논문은 로봇에이전트가 웹상의 멀티미디어 정보를 찾아 데이터베이스에 저장할 때 N-Gram방법을 이용하여 색인하는 기법을 제안하였다. HTML페이지에 나타나는 텍스트 중 이미지 아래에 붙은 표제나 이미지 링크에 붙어 있는 텍스트를 골라내어 이미지의 색인 정보로 이용하였다. 그러나, 웹 문서에 포함되거나 연결된 이미지를 설명하는 텍스트를 더 지능적으로 추출하는 방법이 필요하고 내용 검색에 좀 더 빠른 방법의 특징 검출 시스템이 필요하다. 또한 대용량의 멀티미디어 정보를 저장하기위해서 신속하게 검색하기 위한 인덱스 설정문제와 데이터베이스의 저장 공간 문제에 대한 연구가 필요하다.

그림 4 멀티미디어 검색결과

[참고문헌]

[1] 야후의 이미지 검색. http://imagesearch.yahoo.co.kr/

[2] J.R. Sm. S.F. Chang, "An Image and Video Engine for the Word-Wide Web", Symposium on Electronic Image: Science and Technology Storage & Retrieval for Image and Video Database V. San Jose, CA, February 1997.

[3] Charles Frankel, Michael J. Swain, Vassilis, "Webseer: An Image Search Engine for the World-Wide Web", TR 96-14, U. Chicago, 1996.

[4] Stan Sclaroff, Leniod Tayche, Macro La Cascia, "ImageRover: A Content-Base Image Browser for the World Wide Web", Proc. IEEE Workshop on Content-base Access of Omage and Video

Labraries, 1997.
[5] V. Harmandas. M. Sanderson. M. D. Dunlop, "Image retrieval by hypertext links", ACM SIGIR '97, 1997.
[6] 이성민, 이형우, 최창원, "주문형 뉴스 서비스를 위한 에이전트의 설계", 한국정보처리학회, 1998, 봄
[7] Martijn Koster, ConneXions, "Robot in the Web: threat or treat?", Volumn 9, No.4, April, 1995
[8] 차병래, 서재현, "Web/DOI 기반 멀티미디어 검색 엔진 설계". 목포대학교 정보처리연구소, 제8권, pp.101-112, 2000.
[9] 이준호, 안정수, 박현주, 김명호, "한글문서의 효과적인 검색을 위한 n-gram기반의 색인 방법", 정보처리학회지, 제13권 제1호, pp.47-63, 1996