

그린 환경을 위한 웹기반 대용량 이미지 콘텐츠 검색 시스템 설계 및 구현[†]

(Design and Implementation of Web-based Retrieval
System for Massive Image Contents in Green
Computing Environment)

나 문 성*, 이 재 동*

(Moon-Sung Na, Jae-Dong Lee)

요 약 환경에 대한 문제가 대두됨에 따라, 에너지와 자원의 소비를 억제하고 탄소 배출을 줄이기 위한 노력들이 진행되고 있다. 그 중 다양한 형태의 문서들을 디지털로 변환하여 컴퓨터상에 파일로 보관하고, 이를 전자 문서 또는 전자 이미지 형태로 제공하여 에너지와 자원을 줄이기 위한 노력들이 진행되고 있다. 하지만, 디지털화 된 이미지 데이터들은 인터넷상에 표준화되지 못한 채 산재해 있어 이를 활용하기 위해서는 많은 시간과 노력이 필요하게 된다. 이에 따라 전력 소비 및 자원의 소비가 다시 증가하게 된다. 본 논문에서는 인터넷 상에 존재하는 디지털화된 이미지 데이터의 효율적인 관리와 제공을 통해 에너지와 자원의 소비를 줄이기 위한 시스템을 설계하고 구현한다. 웹상에 존재하는 대용량의 이미지 콘텐츠들을 보다 정확하게 분류하고 사용자가 원하는 이미지 콘텐츠를 정확하고 빠르게 제공해 줌으로써 이미지 데이터에 대한 에너지와 자원의 소비를 줄일 수 있다.

핵심주제어 : 그린 환경, 웹기반, 이미지 검색

Abstract As environmental issues are emerging, many efforts are globally conducted to reduce waste of energies and resources for green growth, as well as low-carbon emitting and replacement of document papers with digital files and images. On the other hand, it may require much time and efforts for users to find the proper image files on the web, where enormous un-standardized digital files are flourishing. Therefore, power and resource consumption may also grow up again in searching and retrieving files. This paper suggests efficient system design and implementation for fast and precise massive image contents retrieval for saving the energies and resources. Eventually it will contribute to green growth in computing environment.

Key Words : Green Environment, Web-based, Image Retrieval, HTML

1. 서 론

전 세계적으로 지구 온난화로부터 환경을 보호하기 위해 온실가스 배출을 규제하고, 화석 연료 및 원자재 가격이 급증하자 이에 대한 대응으로 국내외적으로 녹색성장 요구가 거세지고

[†] 이 연구는 2009학년도 단국대학교 미디어콘텐츠연구원(IMC)의 연구비 지원으로 이루어졌음.

* 단국대학교 정보컴퓨터학과

있다. 이러한 녹색성장의 방안으로 디지털 기술을 활용한 그린 IT 기술이 대두되고 있다. 그린 IT기술은 이전의 생산-유통-사용-폐기물 과정에서 발생하는 에너지와 자원의 소비를 억제하기 위해 디지털 기술을 활용하고 디지털화된 데이터를 사용함에 따라 에너지와 자원의 효율화를 위한 방안으로 활용될 수 있다[1]. 실제 일본 경제산업성 자료에 따르면 IT기술의 활용으로 약 11조kWh의 에너지를 절감할 수 있을 것으로 예상하고 있으며, 유럽통신사업자협회에 따르면 IT기술을 활용하였을 경우 약 10배 이상의 탄소 배출 억제 효과를 가져올 수 있을 것으로 예상한다[2].

그린 IT 기술의 요구 증가와 함께 디지털 카메라, MP3, PMP 등 다양한 디지털 단말기들의 발전으로 아날로그 시대에서 디지털 시대로 빠르게 변화하고 있다. 이러한 변화 속에서 데이터의 형태 및 콘텐츠 소비 양상 역시 이전에 아날로그 형태의 제작, 구매, 소비 패턴에서 디지털 형태의 패턴으로 변화하고 있다. 사진의 경우 필름 카메라를 이용하여 사진을 찍고 이를 현상하여 보관하던 아날로그 형태의 제작, 구매, 소비 패턴에서 디지털 카메라의 보급 및 활용을 통해 사용자 자신이 직접 사진을 찍고, 그 사진을 PC로 전송하여 보관, 관리 할 수 있는 형태로 변화하고 있다.

이러한 디지털화를 통한 변화는 이미지 데이터들의 에너지와 자원의 소비를 줄이고 다양한 형태의 서비스 모델을 제공하고 있다. 하지만 디지털화된 이미지 데이터들이 인터넷 상에 무분별하게 존재함에 따라 대용량의 이미지 데이터들을 효율적으로 관리하고 제공해 주기 위해 많은 양의 에너지와 자원을 소비하게 된다. 이를 위해 본 논문에서는 웹상에 존재하는 디지털화된 이미지 데이터들을 효율적으로 관리하고 제공해 주기 위한 시스템을 설계하고 구현한다. 구현된 시스템은 에너지와 자원의 소비를 줄임으로써 그린 환경을 위한 방안으로 활용될 수 있을 것이다.

본 논문의 2장에서는 관련 연구에 대해서 기술한다. 3장에서는 에너지와 자원의 소비를 줄이기 위한 이미지 관리 및 제공을 위한 시스템

을 설계하고 구현한다. 4장에서는 구현된 시스템의 성능 평가를 진행한다. 마지막으로 5장에서는 향후 연구 과제에 대해 기술한다.

2. 관련 연구

2.1 그린 IT 기술

2.1.1 그린 IT 개념

그린 IT는 크게 IT 부분의 친환경활동(Green of IT)과 IT를 활용한 친환경 활동(Green by IT)으로 구분된다[1]. Green IT는 비용절감, 지속적인 부가가치 및 경쟁우위를 위한 IT 산업의 친화적 경영 활동으로 IT의 에너지 절감과 유해물질, 폐자원 관리를 통해 IT 부문 자체에 대한 환경규제 준수 및 에너지를 최소화하는 활동이다. Green by IT는 기업운영·공급자관리 과정에서 환경 지속가능성을 위해 상품·서비스·자원의 라이프사이클에 걸쳐 최적의 IT를 사용하는 것이다. 곧, 물리적 활동을 온라인으로 대체함으로써 탄소 등 환경오염 물질의 배출과 에너지 자원의 소비를 최소화하는 활동이다[1].

2.1.2 IT 부문을 활용한 친환경 활동

IT 부문을 활용한 친환경 활동으로 실천할 수 있는 구체적인 방안은 사람의 이동감소, 공간 절약화, 무종이화, 사물의 이동 감소 등으로 나누어 볼 수 있다[3]. 사람의 이동을 줄이기 위한 방안은 네트워크를 통한 유비쿼터스 오피스의 실현으로 가능하다. 또한, 재택근무 등을 통해서 공간 절약화를 통한 에너지 소비량을 줄일 수 있다. 전자 문서, 전자 이미지, 전자 액자 등의 IT 기술은 기존 종이 또는 인화지 등을 통해 제공되는 정보를 컴퓨터상의 문서, 이미지로 제공함으로써 자원의 소비를 감소시킬 수 있다. 또한, 물건의 이동 역시 SCM(Supply Chain Management), TMS(Transportation Management System) 등을 통해 이동 경로의 최적화, 이동 방법의 최적화를 통해 실현가능하다.

2.2 Web 기반 이미지 관리/검색 기술

이미지 콘텐츠에 대한 검색은 크게 텍스트에 의한 이미지 콘텐츠 검색과 내용 기반 이미지 검색으로 나뉠 수 있다[4]. 과거에는 주로 텍스트 기반의 이미지 검색이 주를 이루었고 최근 들어 내용 기반 이미지 검색 분야에도 활발한 연구가 이루어지고 있다. 최근에 등장한 이미지의 색상, 질감 형태 등의 특성을 이용한 내용 기반 이미지 검색은 텍스트 기반 이미지 검색에 비해 고차원적 검색이라고 보기보다는 이미지 검색의 새로운 분야로 보는 것이 좋을 것이다. 텍스트 기반 이미지 검색은 과거에서부터 현재까지 가장 널리 이용되는 이미지 검색 기법이며 이미지와 관련된 텍스트 정보, 혹은 메타데이터가 존재하는 한 이미지 검색에 필수적인 검색 기법일 것이다.

표준화된 메타데이터를 기반을 한 이미지 검색은 이미지에 대한 메타데이터를 대상으로 검색이 이루어지므로 이는 텍스트 기반의 이미지 검색 방식이다. HTML에 의한 검색 역시 텍스트 기반 검색이지만 여기에는 차이점이 있다. HTML은 문서 구조나 내용 정보를 기계가 이해할 수 없는 형식이며 전문검색(full-text search)을 통해서 검색이 가능하다. 그러나 이러한 방식으로 검색된 수천, 수만 건의 문서는 검색의 정확성이 떨어지게 된다. 그러나 만약 표준화된 이미지 메타데이터가 존재하고 이를 대상으로 검색을 하게 된다면 검색 정확성의 향상을 가져올 수 있을 것이다. 예를 들어 누가, 언제, 어떠한 주제로 문서가 작성되었는지 알고 이와 같은 부분을 한정지어 검색할 수 있다면 검색의 정확성은 훨씬 향상될 것이다[10]. HTML에서도 <Meta> 태그를 통해 <Meta Name="description" Content="....."> 또는 <Meta Name="keyword" Content=".....">와 같은 방식으로 약간의 내용정보나 키워드 등을 입력할 수 있는 메타태그를 지원하고 있지만 이를 지원하는 웹 검색 엔진은 일부에 불과하다[5]. 또한 통합된 표준이 없이 각 검색 엔진마다 지원하는 태그의 속성이 달라 사용자가 메타정보를 표시하는데 어려움이 있으며 문서 종류에

따라 고유한 메타정보를 가질 수 있는데 이를 새롭게 정의할 수가 없다[6].

3. 대용량 이미지 관리 및 검색시스템 설계 및 구현

본 장에서는 웹상에 존재하는 다양한 형태의 대용량 이미지 콘텐츠를 관리하고 검색하기 위해서 필요한 요소들을 정의하고, 이를 표준화하기 위한 메타데이터를 제안한다. 그리고 이미지 관리 및 검색을 위한 시스템을 설계하고 이를 구현한다.

3.1 웹 문서 기반 이미지 정보 관리

웹상에 존재하는 웹 문서들에서 이미지가 차지하는 비율은 전체 문서의 약 70% 정도이다. 이것은 브라우저를 통해 사용자들이 보게 되는 부분이 대개 이미지로 작업된 내용들이라는 이야기이다. 웹 문서 상에는 사용자들이 보기 편하게 하기 위해 배경 및 메뉴 등 다양한 내용을 이미지로 작성하게 되고, 그 중 의미 있는 정보를 가지고 있는 이미지 콘텐츠는 대개 1~3개 정도이다[7][8].

이렇게 작성된 페이지의 경우 이미지의 역할은 텍스트의 내용을 보조하는 수단으로 사용된다. 브라우저 상에 빨간색 네모 박스의 내용은 이미지를 대표할 수 있는 단어들로서 수작업으로 선정한 내용이다. 이러한 단어들을 볼 때 이미지의 내용을 텍스트가 설명하는 구조보다는 텍스트의 내용을 보다 사실감 있게 또는 현장감 있게 제공하기 위한 수단으로 이미지가 사용된다. 이에 웹 문서상의 일반적인 텍스트 내용만을 가지고 이미지의 의미를 추론하는 것보다 특정 태그에 의한 의미 추출, 특정 태그 내에서 단어의 출현 빈도, 이미지의 URL, 이미지의 파일명, 웹 페이지의 타이틀, 이미지 캡션, <ALT> 태그, HTML 제목, 하이퍼링크 등을 이용하여 이미지의 의미를 나타내는 단어를 선정해야 한다[9].

HTML 태그에는 다양한 의미를 가지고 있는

태그들이 존재한다. 그 중 문서의 내용을 대표하는 태그로 <Title>, <Meta> 태그가 있다. <Title>태그는 브라우저 상단에 표시되는 텍스트를 나타내기 위한 태그로 보통 문서 전체의 의미 또는 키워드의 내용을 적는다. 또한 <Meta> 태그는 다양한 문서의 정보를 정의하게 된다. [그림 1]은 이러한 <Title>, <Meta> 태그의 사용 예이다.

```

read[1] - 네이버
<html>
<head>
<meta http-equiv="Content-type" content="text/html; charset=euc-kr"></meta>
<link rel="shortcut icon" href="http://imgnews.naver.com/image/favicon.ico">
<title>네이버 뉴스</title>
<link rel="stylesheet" type="text/css" href="/css/news200705291500.css">
<script type="text/javascript" src="/js/news_200705161645.js"></script>
<script type="text/javascript" src="/js/keyword_20070418.js"></script>
<!-- han95a -->
<script>
document.domain = 'news.naver.com';
</script>
</head>
<body bgcolor=#ffffff leftmargin=5 topmargin=0>
<center>
<style>

```

[그림 1] <Title>, <Meta> 태그 사용 예

이러한 <Title>, <Meta> 태그는 문서 전체의 내용을 나타내므로 그 문서에 속한 이미지의 내용을 대표하는 단어로 사용 될 수 있다. 사용자가 뉴스 상의 한 이미지를 검색하고자 할 때 [그림 1]의 예에서와 같이 <Title> 태그 상에 '네이버 뉴스'라는 문서를 대표하여 설명하는 단어를 통해 관련 내용을 추론 할 수 있다.

또한 태그 내에는 alt 속성이 존재한다. alt 속성의 경우 이미지가 브라우저 상에서

```

read[1] - 네이버
<table width="15" cellspacing="0" cellpadding="0" border="0">
<tr>
<td>
<a href="http://news.naver.com/moviescene">

</a></td></tr>
<tr><td background="http://imgnews.naver.com/image/news/2007/leedongjin/lee_leftbg.gif" align="center">
<table width="91" cellspacing="0" cellpadding="0" border="0" border="1" style="border-color:#ffffff">
<tr><td height="23" style="padding:3px 0 0 4px"><a href="http://news.naver.com/movies">
<tr><td height="1" bgcolor="#000000"></td></tr>
<tr><td height="23" style="padding:3px 0 0 4px"><a href="http://news.naver.com/movies">
<tr><td height="1" bgcolor="#000000"></td></tr>

```

[그림 2] <Alt> 태그 사용 예

표시되지 못할 경우 이미지의 텍스트를 표시하기 위한 속성으로 이미지에 대한 설명 또는 주제를 적는다.

이미지의 파일명과 이미지에 하이퍼링크 또는 주석이 달려 있는 경우도 이미지의 의미를 나타내는 단어로 사용할 수 있다. 이미지의 파일명은 대개 이미지와 관련된 한글, 영문으로 작성하며, 이미지에 하이퍼링크가 걸려 있는 경우 하이퍼링크로 연결되는 문서의 Title이 이미지의 내용을 대표할 수 있다. 또한 이미지 앞뒤에 주석이 있는 경우 이미지에 대한 추가적인 설명 또는 기능으로 볼 수 있다. HTML 문서상의 의미 있는 태그들로부터 이미지에 대한 의미를 추론하기 위한 단어들을 추출할 수 있다. 이렇게 추출된 단어들은 태그가 나타내는 의미 및 내용에 따라 검색 시 다른 가중치로 계산하여 검색랭킹을 결정하게 된다[10][11].

$$W_{total} = W_{word} + W_{태그}$$

$$W_{word} = 1 * W_{TF} * W_{태그}$$

질의에 대한 단어가 들어 올 경우 해당 단어와 매칭되는 횟수(WTF)와 태그별 가중치(W태그)를 두어 해당 이미지가 질의된 질의어와 얼마나 매칭이 되는지 계산할 수 있다.

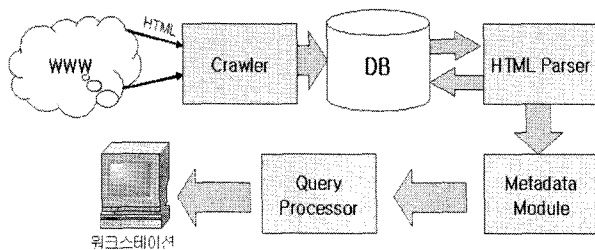
3.2 검색 시스템 설계 및 구현

본 절에서는 웹 페이지 상에 있는 이미지들을 자동으로 수집하고, 이미지에 대한 정보를 자동으로 추론하여 메타데이터로 생성하여 데이터베이스에 저장 한 후 이를 검색에 이용하는 시스템 전반에 대해 설계하고 구현한다.

3.2.1 검색 시스템 구조

[그림 3]은 검색 시스템의 구성 모듈 및 흐름을 나타낸다. 이미지 검색 시스템은 웹 문서를 자동으로 수집하여 데이터베이스에 저장하는 수집기(Crawler)와 모아진 웹문서의 내용을 토대로 이미지를 분리시키고 이미지의 의미를 추론하기 위해 HTML 문서를 정리하는 HTML 파

서(HTML Parser), 이를 의미 있는 태그의 텍스트를 모으고, 이미지의 정보를 추출하여 메타데이터 형식으로 정리하고 메타데이터화 하는 메타데이터 모듈(Metadata module), 그리고 사용자의 질의를 처리하는 질의 처리기(Query Processor), 마지막으로 질의 결과를 시각화 하여 보여주는 인터페이스로 구성된다.



[그림 3] 검색 시스템 구성 모듈 및 흐름도

3.2.2 시스템 S/W 설계

3.2.2.1 HTML Parser

HTML Parser는 수집기(Crawler)를 통해 수집한 웹 문서를 정리하여 각각의 태그 내용별로 정리하여 저장하는 모듈이다. Crawler를 통해 수집한 웹 문서는 작성자의 패턴에 따라 여러 가지 형태를 가지게 된다. 이러한 형태의 문서들을 3.1절에서 살펴본 바와 같이 의미 있는 태그와 이미지 데이터로 분리하여 이미지 콘텐츠에 대한 정보를 추론하기 위해서는 HTML 문서에 대한 정리가 필요하다. 또한 실제 검색서비스에서 SQL문을 이용하여 검색을 하게 되는데, 이때 정리되지 않은 문서상에서 잦은 Select, Join문을 사용하게 되면 많은 시간이 소요되게 된다. 이를 위해 웹 문서에 대한 정리가 필요하고, Metadata module을 통해 메타데이터를 구축하여 보다 빠르고 쉽게 접근이 가능토록 한다.

HTML Parser는 태그와 태그 내에 있는 텍스트들을 정리하는 역할을 한다. 태그와 텍스트의 분리는 '<'문자로 시작하여 '>'문자로 끝나는 텍스트의 경우 태그로 인식하고, 그 외 내용의 경우 텍스트로 분리하여 배열 리스트에 저장을 하게 된다. 이렇게 함으로써 각각의 태그별로 태

그 사이에 존재하는 텍스트를 추출할 수 있다. HTML Parser에 대한 알고리즘은 [그림 4]와 같다.

```

buffer = ""
while read each Html in Database
{
    int a=0;
    for(int l=0; l < Html.Length; l++){
        //태그 부분만을 저장시키는 알고리즘
        if(Html[l] == "<"){
            buffer is 태그 and 태그 start;
            for(int j=l; Html[j] != ">"; j++){
                buffer = Html[j]
            }
            //텍스트 부분만을 저장하는 알고리즘
        } else{
            buffer is Text and Text start;
            for(int j=l; Html[j+1] != "<"; j++){
                buffer = Html[j]
            }
        }
        ArrayList[a] = Buffer;
        a++;
    }
    save ArrayList in Database
}
  
```

[그림 4] HTML Parser 동작 알고리즘

3.2.2.2 질의 처리기(Query Processor)

질의 처리기는 사용자가 입력한 쿼리(Query)와 메타데이터의 내용을 비교하여 점수를 계산하고 계산된 점수에 따라 순위를 정해 사용자 인터페이스에 제공하는 모듈이다.

점수 계산을 위해 사용자가 입력한 쿼리를 메타데이터 상의 데이터와 비교하고 사용자의 쿼리와 일치하는 데이터가 메타데이터 상에 존재할 경우 아래 계산식에서와 같이 계산을 하여 점수를 계산토록 한다. 또한 <표 1>과 같이 각각의 메타데이터의 내용에 따라 가중치를 달리 주어 각 태그가 의미하는 바를 보다 자세히 반영토록 하였다.

<표 1>과 같이 메타데이터별 가중치 계산은 2가지 방법에 따라 결정하였다. 첫째, 메타데이터별 가중치의 계산은 이미지 콘텐츠의 메타데이터 200개를 선정하여 수작업으로 그 내용을

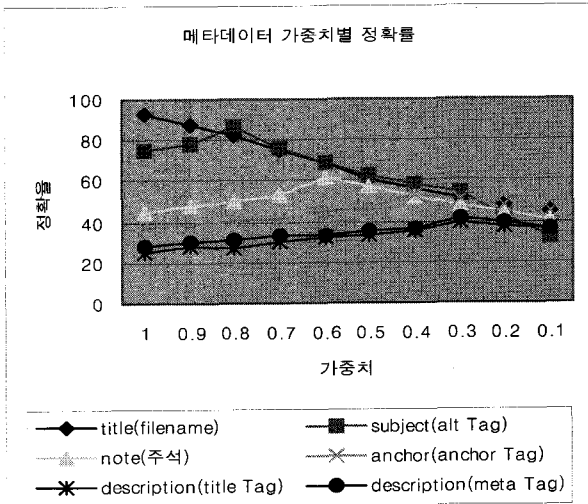
비교하여 결정하였다.

약200개의 표본 메타데이터 중 Title(이미지의 파일명)이 의미성을 가지는 단어일 경우(숫자로 표기한 경우 제외) 이미지의 내용을 대표하는 경우가 100%이고, Subject(<Alt> 태그)가 존재할 경우 이미지의 내용을 대표하는 경우가 약 80%정도였다. 또한 Description(note, 주석)이 존재할 경우 임지의 내용을 대표하는 경우가 약 60%정도였다. 이것은 주석이 이미지의 설명 외에 다른 HTML상의 주석 내용을 대표하는 경우가 있기 때문이다. Description(<A> 태그)의 경우는 약 40%, Description(<Title>, <Meta> 태그)의 경우는 약 30%이다.

<표 1> 메타데이터별 가중치

$W_{word} = 1$ $* W_{TF} * W_{태그}$ W_{word} : 단어의 점수 W_{TF} : 빈도수 $W_{태그}$: 메타데이터별 가중치	
메타데이터	가중치
title	1.0
subject	0.8
description(note)	0.6
description(anchor)	0.4
description(title, meta)	0.3

둘째, 각 메타데이터 속성 별 가중치의 값을 0.1~ 1.0까지 유동적으로 변경해 가며 각 가중치 값별 정확률을 계산하였다. [그림 5]는 각 메타데이터 속성별 가중치 값을 달리 하였을 경우 나오게 되는 이미지 검색 시 정확률이다.



[그림 5] 메타데이터 가중치별 정확률

4. 실험 환경 및 성능분석

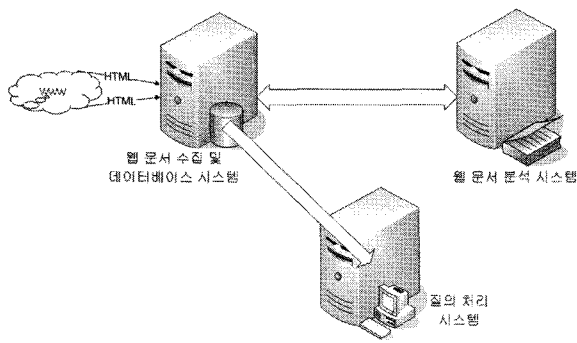
4.1 실험 환경

이미지 검색시스템을 구현하기 위해서는 [그림 6]과 같이 크게 3 부분의 실험용 시스템이 필요하였다.

첫째로, 웹상의 문서를 수집하여 데이터베이스에 저장하는 웹 문서 수집 시스템(Crawler & DB)과 수집된 문서들을 Parsing하고 메타데이터로 정리하는 웹 문서 분석 시스템, 마지막으로 사용자 질의가 들어 왔을 때 질의를 처리하는 질의 처리 시스템으로 나뉘어 질 수 있다. 또한 각 시스템의 소프트웨어 구성은 <표 2>와 같다.

<표 2> 실험을 위한 시스템의 소프트웨어 구성도

구분	웹 문서 수집 및 데이터베이스 시스템	웹 문서 분석 시스템	질의 처리 시스템
OS	Windows 2003	Windows 2003	Windows XP
사용 툴	Visual Studio 2003	Visual Studio 2003	Visual Studio 2003
DB	SQL 2000	-	-
사용 언어	C#	C#	C#



[그림 6] 실험을 위한 시스템 구성도

4.2 성능평가

이미지 검색 시스템의 성능평가를 위해서는 일정 URL 주소를 선정하고 이에 대한 웹 문서를 약 900개를 수집하였다. 수집된 웹 문서는 성능평가를 위해 현재 검색 엔진 성능이 가장 좋다고 알려져 있는 구글(www.google.co.kr)에서 특정 키워드 30개를 입력하여 나온 결과 중 1위부터 30위에 해당하는 URL을 선정하여 900개의 문서를 수집하였다.

특정 키워드의 내용은 <표 3>과 같으며, 이 내용은 2006년도 인기 검색어 리스트에서 랜덤하게 30개를 선정하여 사용하였다.

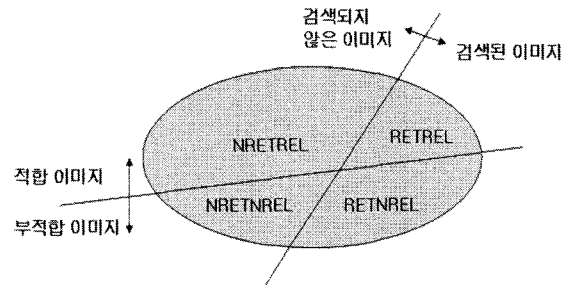
<표 3> 특정 키워드 리스트

Ace Golf	GS 스쿼어	G-마켓	Job Korea
MBC라디오	MSN KOREA	휴대폰	강원대
검찰청	과일나라	농협비씨	교원나라
금호렌트카	나이키	대한건축사	대한생명
LG	거북선	굿데이신문	동대문시장
란	디자이너싸이드	레이디경향	마리아병원
무대조명	바다넷	바둑정석	변호사
복음성가	대한교직원공제회		

이렇게 특정 검색 엔진에서 검색 결과로 나온 URL을 선정하여 웹 문서를 가져온 것은 검색 엔진의 정확성을 비교하기 위함이다. 검색 성능이 가장 좋다고 알려진 검색 엔진으로부터 나온 결과 문서 중 이 문서 안에 있는 이미지 콘텐츠를 추출하고 이들에 대한 이미지 정보를 수집하여 검색을 실시한다.

또한 검색엔진의 일반적인 성능평가는 응답속

도, 시간당 처리율, 랭킹 조작성, 색인어 추출의 정확도, 검색의 재현율, 정확률 등 여러 가지 성능평가 항목들이 있다. 그 중 본 논문에서는 [그림 7]과 같이 검색된 이미지, 검색되지 않은 이미지, 적합 이미지, 부적합 이미지로 나누어서 계산하여 재현율과 정확도를 측정하여 성능 평가를 실시한다.



[그림 7] 성능 평가 항목

또한 재현율과 정확도를 계산하기 위한 식은 아래와 같다.

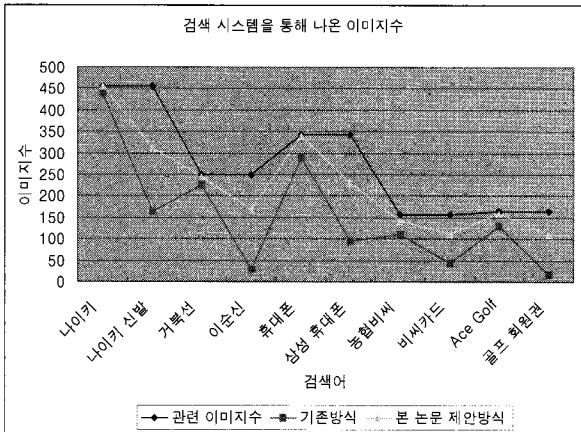
$$\text{재현율} = \frac{\text{검색된 적합 이미지 수}}{\text{적합 이미지 수}} = \frac{\text{RETREL}}{\text{RETREL} + \text{NRETREL}}$$

$$\text{정확률} = \frac{\text{검색된 적합 이미지 수}}{\text{검색된 이미지 수}} = \frac{\text{RETREL}}{\text{RETREL} + \text{RETNREL}}$$

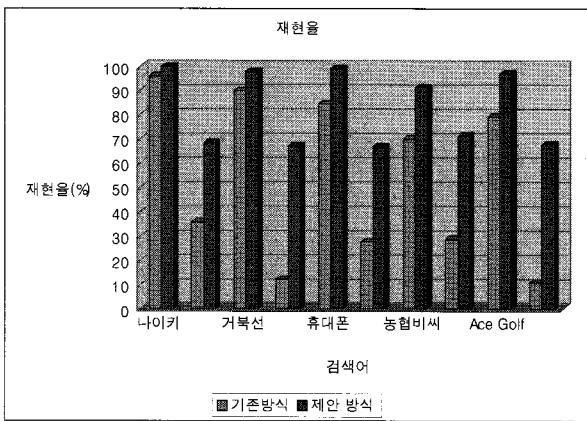
4.2.1 재현율

재현율 측정을 위해 일부 검색어를 입력하여 기존 방식과 본 논문에서 구현한 방식을 통해 검색되어 나온 이미지 수를 측정하였다. (그림 8)은 검색을 통해 결과로 나온 이미지들의 수이며 관련 이미지수는 원본 문서를 통해 추출된 이미지의 수이다. 또한 (그림 9)는 검색을 통해 나온 결과를 통해 재현율을 계산하여 막대그래프로 도식화한 그림이다.

재현율 성능분석은 단순 질의에 대한 검색과 복합 질의 또는 유사 의미에 대한 질의에 대한 분석이다. '나이키', '거북선', '휴대폰', '농협비씨', 'Ace Golf'와 같은 단순 질의에 대한 검색 결과는 재현율의 경우 기존방식과 본 논문에서 제안한 방식이 약 10% 정도의 차이를 보인다. 하지만 복합 질의 또는 비슷한 의미의 검색('나이키 신발', '이순신', '삼성 휴대폰', '비씨카드', '골프



[그림 8] 검색을 통해 나온 이미지수



[그림 9] 본 논문에서 제안한 검색시스템의 재현율

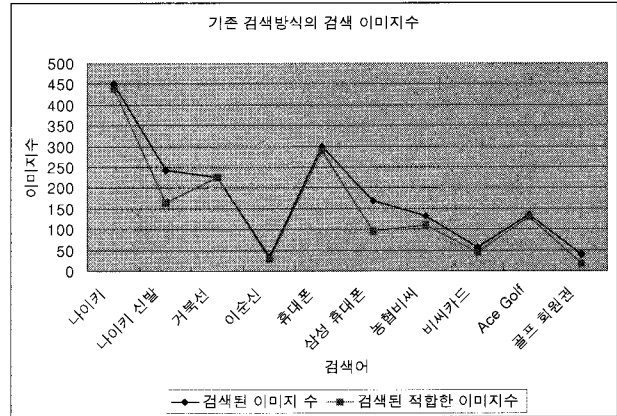
회원권)에 대한 재현율의 검색 결과는 작게는 30%에서 약 50%의 성능 차이를 보인다. 이는 기존 방식에서는 제공하지 못하는 이미지 의미에 대한 내용을 웹 문서를 통해 추론함으로써 본 논문에서 구현한 검색 시스템이 보다 평균 30% 정도의 성능이 향상됨을 알 수 있다.

4.2.2 정확률

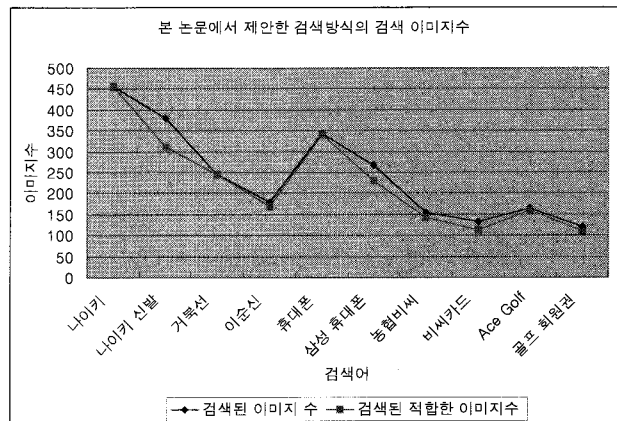
정확률 측정을 위해 재현율을 구하기 위해 검색한 결과에서 결과값으로 나온 이미지들을 수작업으로 분석하여 관련 검색어와 적합한 이미지가 나왔는지 분석하였다.

이를 통해 기존 방식과 본 논문에서 제안한 시스템간의 정확률 비교가 가능하다. [그림 10]은 기존 방식을 통해 검색한 결과 검색된 이미

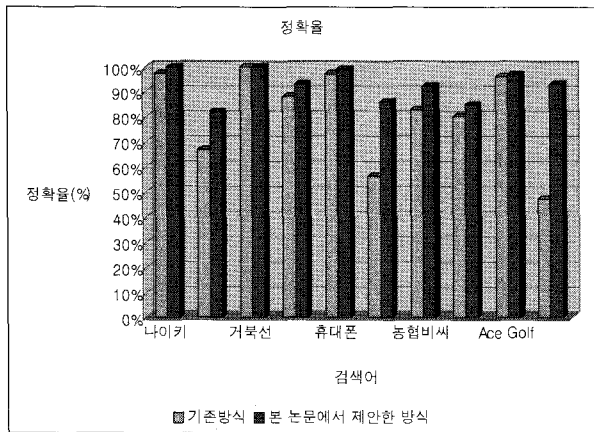
지의 수와 그 중 적합한 이미지 수에 대한 그래프이다. [그림 11]은 본 논문에서 구현한 시스템을 통해 검색한 결과 검색된 이미지 수와 그 중 적합한 이미지 수에 대한 그래프이다. 그리고 [그림 12]는 [그림 10]과 [그림 11]을 통해 계산된 정확률에 대한 그래프이다.



[그림 10] 기존검색방식에 의해 검색된 이미지수와 적합한 이미지 수



[그림 11] 본 논문에서 제안한 검색 방식에 의해 검색된 이미지수와 적합한 이미지 수



[그림 12] 검색 시스템에 대한 정확률

5. 결론 및 향후 연구방향

최근 환경에 대한 문제들이 대두됨에 따라 에너지 소비를 줄이고 탄소 배출을 줄이기 위한 노력들이 진행되고 있다. 이러한 노력의 일환으로 다양한 형태의 문서들을 전자 문서 또는 전자 이미지 형태의 디지털 데이터로 변환하여 컴퓨터상에서 제공 받아 에너지 자원의 소비를 감소시키기 위한 노력들이 진행되고 있다. 하지만, 디지털화 된 멀티미디어 콘텐츠들이 웹상에 무수히 많이 존재하게 됨으로써 멀티미디어 콘텐츠를 사용하기 위해서 많은 시간의 콘텐츠 검색을 위한 시간이 소요되고, 이를 통해 에너지 소비가 증가하는 문제점이 발생하게 된다. 이에 사용자가 원하는 내용의 콘텐츠를 찾기 위한 검색 시스템이 필요하게 되고, 보다 효율적이고 정확성 있게 사용자의 요구에 부합하는 콘텐츠를 찾기 위한 기술들이 필요하다.

현재 멀티미디어 콘텐츠 중에서 이미지 콘텐츠들에 대한 검색은 몇몇 검색 엔진들에 의해 제공 되고 있다. 하지만 주로 관련 업체에서 제공하는 데이터베이스 상에 존재하는 이미지에 대한 검색만을 수행하거나, 단순히 키워드 매칭의 텍스트 위주(파일명, 제목 등) 검색만이 제공되고 있어 검색의 범위가 제한적이며 검색의 정확성과 효율성이 높지 않다.

따라서 본 논문에서는 웹상에서 이미지 콘텐

츠들에 대한 검색을 위해 웹 이미지를 자동적으로 가져와 이미지의 내용을 추론하고 이를 메타 데이터로 생성하여 보다 정확성 있고 확장성 높은 검색이 되도록 이미지 검색 시스템을 설계하고 구현하였다.

이 시스템을 통해서 얻을 수 있는 장점은 Crawler에 의해서 수집된 모든 문서들에 대한 이미지 검색이 가능하여 관련 업체의 데이터만을 검색하여 제공하는 방법보다 많은 데이터를 검색하여 제공할 수 있다는 점과, 이미지의 내용을 이미지가 있는 웹 문서상의 내용을 토대로 추론하여 이를 검색에 이용함으로써 사용자의 요구에 부합하는 이미지 콘텐츠를 빠르고 정확하게 제공할 수 있다는 점이다. 성능분석과 같이 기존 방식에 의한 검색이 제한적이었지만 반해 본 논문에서 설계한 시스템이 약 40% 이상의 재현율과 약 20% 정도의 정확률에서 높은 성능 향상을 가져옴을 알 수 있다. 즉 HTML 문서상의 데이터를 통해 복잡한 사용자의 질의에 대해서 보다 확장된 검색이 가능하며, 단어 간 가중치에 따른 알고리즘에 의해 보다 정확한 검색이 가능하다.

향후 연구 과제로는 가중치의 값을 현재 표본 데이터의 정확율을 통해 선결정하여 알고리즘을 작성하였는데, 질의어에 따라 가중치 값이 변동되어 보다 정확한 값을 결정하기 위한 알고리즘 연구가 필요하다. 또한 이미지뿐만 아니라 동영상, 음악 등의 다양한 멀티미디어 콘텐츠를 검색하기 위한 검색 시스템의 확장이 필요하다. 즉 HTML 문서의 구조를 분석하고 이를 통해 각각의 멀티미디어 콘텐츠에 대한 내용을 추론하여 보다 다양한 형태의 콘텐츠를 검색하기 위한 검색 방법이 필요하다. 이를 통해 컴퓨터상에서 소비되는 에너지 소비를 줄이고 전력 감소 효과를 증대시킬 수 있다.

참 고 문 헌

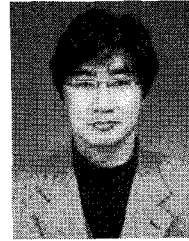
- [1] 이혜정, "그린경영과 그린IT," 전자공학회지 제35권 11호, pp.43~55, 2008.
- [2] 이지훈 외, "녹색성장시대의 도래," CEO

- Information, 제675호, 삼성경제연구소, 2008.
- [3] 김병관, “그린 IT를 통한 환경과 경제의 상생 방안,” 한국경영정보학회, Vol.2009, No.11, pp.691~696, 2009.
 - [4] Goodrum, A. A. “Image Information Retrieval : An Overview of Current Research,” Informing Science, Vol.3, No 2, pp.63~67, 2000.
 - [5] 남호현, 양광민. “웹 문서에서 메타태그 활용도에 관한 연구.” 경영학논문집. Vol.26, No.2, pp.1~20, 2000.
 - [6] 윤승구, 1998. “구조와 메타정보 검색을 지원하는 SGML 웹 검색 엔진의 개발.” 연세대학교 대학원 석사학위논문, 1998.
 - [7] 남승희, “이미지 검색시스템을 위한 메타데이터 구축에 관한 연구,” 연세대학교 대학원 석사학위논문, 2001.
 - [8] 남호현, 양광민. “웹 문서에서 메타태그 활용도에 관한 연구.” 경영학논문집. Vol.26, No.2, pp.1~20. 2000.
 - [9] Lassila, O. “Web Metadata : A Matter of Semantics.” IEEE Internet Computing, pp.30~37, 1998.
 - [10] Lew, M. S. “Next-Generation Web Searches for Visual Content,” IEEE Computer, pp.46~53, 2000.
 - [11] Abbadeni, N., D. Ziou, and S. Wang. “Image Classification and Retrieval in the World Wide Web,” Proceedings of the fourth ACM conference on Digital libraries, pp.208~209, 1999.



나 문 성 (Moon-Sung Na)

- 정회원
- 1989년 02월: 조선대학교 전자공학과 (공학사)
- 2003년 08월: 한양대학교 경영대학원 (석사)
- 2006년 9월~2009년 9월: 단국대학교 컴퓨터과 학과 (박사수료)
- 1996년 9월~1997년 3월: 운전시뮬레이터 개발
- 2005년 3월~2007년 2월: 단국대학교 정보통신대학원 겸임교수
- 2000년12월~현재: 한국콘텐츠진흥원 문화기술본부장
- 관심분야 : 멀티미디어 콘텐츠, 가상화, 하드웨어 시스템 구현



이 재 동 (Jae-Dong Lee)

- 정회원
- 1985년: 인하대학교 전자계산학과 (학사)
- 1991년: Cleveland State University (석사)
- 1996년: Kent State University (박사)
- 1997년 3월~현재: 단국대학교 컴퓨터학부 컴퓨터공학전공 교수
- 2009.8~현재 단국대학교 미디어콘텐츠기술센터 센터장
- 2009.7~현재 한국문화콘텐츠기술학회 학회장
- 2006.4~현재 국가지정 단국대학교 CT연구소 소장
- 2002.11~현재 농협중앙회 전산고문
- 2007.2~2009.3 Dream economy leader 포럼위원
- 2006.7~2007.12 민관확대 콘텐츠 정책협의회 위원
- 2005.8~2006.8 문화관광부 KOCCA CT포럼/전략기획 운영위원/분과위원장
- 2005.3~2009.6 단국대학교 콘텐츠&컨버전스기술연구소 소장
- 2005.1~2006.12 전국대학정보화 협의회 이사
- 2004.7~2006.6 단국대학교 정보통신원 원장 (C.I.O)
- 2004.1~2006.6 (사)이러닝 산업협회 이사
- 관심분야 : Ubiquitous Computing, Contents & Entertainment Technologies, (Mobile) Internet Technology/Applications, Cloud Computing