Journal of Korea Multimedia Society Vol. 15, No. 9, September 2012(pp. 1165-1173)

# 3-태그 기반의 웹 이미지 검색 기법

이시화†, 황대훈††


## 요 약

웹2.0 환경에서의 대중적인 기술 중 하나는 태깅이며, 현재 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 적용되고 있다. 그러나 태깅에 사용된 태그가 정보 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 실제로는 부정확한 태그로 인해 낮은 검색 결과를 제공 하고 있다. 이에 선행 연구에서는 웹상에 산재되어있는 다양한 리소스 및 그에 따른 태그 정보들을 수집하여 태그들 간의 연관성에 따라 맵핑하고, 이를 클러스터링 하기 위한 연구를 진행하였다. 본 논문에서는 클러스터링된 태그들을 검색에 활용하는 3-태그 기반 검색 알고리즘을 제안하였다. 제안 알고리즘의 성능평가를 위하여, 태그 기반 대표 사이트인 Flicker 사이트의 이미지 검색 결과와 정확성 및 재현율을 비교 평가하였다.


# 3-tag-based Web Image Retrieval Technique

Si-Hwa Lee†, Dae-Hoon Hwang††


## ABSTRACT

One of the most popular technologies in Web2.0 is tagging, and it widely applies to Web content as well as multimedia data such as image and video. Web users have expected that tags by themselves would be reused in information search and maximize the search efficiency, but wrong tag by irresponsible Web users really has brought forth a incorrect search results. In past papers, we have gathered various information resources and tags scattered in Web, mapped one tag onto other tags, and clustered these tags according to the corelation between them. A 3-tag based search algorithm which use the clustered tags of past papers, is proposed in this paper. For performance evaluation of the proposed algorithm, our algorithm is compared with image search result of Flickr, typical tag based site, and is evaluated in accuracy and recall factor.


**Key words:** Web2.0(웹2.0), Tag(태그), Clustering(클러스터링), Image Retrieval(이미지 검색)

## 1. 서 론

인터넷의 발달과 사용자의 적극적인 참여에 힘입어 웹서비스 환경은 다양하게 변화하고 있으며, 이러한 변화의 흐름을 잘 반영하는 것이 웹2.0이다.

웹2.0에서 대부분의 정보는 사용자에 의해 생산되고, 사용자가 붙인 태그에 의해 분류되어진다[1]. 이와 같이 현재 큰 반향을 일으키고 있는 웹2.0은 이제 모든 인터넷 사이트들의 필수 전략이 되었으며, 웹2.0을 성공적으로 구현하기 위한 다양한 기법들이 등장하고 있다. 이러한 기법들 중 핵심적인 기술이 바로 태깅이다[2,3].

※ 교신저자(Corresponding Author) : 황대훈, 주소 : 경기도 성남시 수정구 복정동 산 65번지 가천대학교 새롬관 5-14호(461-701), 전화 : 010)5458-8111, FAX : 031)757-6715, E-mail : hwangdh@gachon.ac.kr
접수일 : 2012년 4월 19일, 수정일 : 2012년 7월 12일
완료일 : 2012년 8월 10일

† 준회원, 가천대학교 일반대학원
(E-mail : leesihwaman@gmail.com)
†† 종신회원, 가천대학교 IT대학 인터렉티브미디어학과
※ "이논문은 2012년도 가천대학교 교내연구비 지원에 의한 결과임".(GCU-2012-R046).

태깅은 현재 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 적용되고 있다[4]. 그러나 태깅에 사용된 태그가 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 실제로는 태그가 가지는 부정확한 태그로 인해 만족스럽지 못한 검색결과가 도출되고 있다[5,6].

이와 같은 문제점을 해결하기 위해 본 논문의 선행 연구에서는 웹상에 산재되어 있는 리소스와 그에 따른 태그 정보들을 수집하여 태그들 간의 연관성에 따라 맵핑하고, 이를 클러스터링하기 위한 시스템에 관한 연구를 진행하였으며[7,8], 본 논문에서는 생성된 클러스터내의 태그들을 기반으로 향상된 검색 서비스 제공을 위한 3-태그 기반 검색 알고리즘을 제안 및 구현하였다. 제안한 3-태그 기반 검색 알고리즘의 성능평가를 위해 기존 단일 태그 매칭 검색 기법과 선행 연구에서 제안한 2태그 검색 기법과의 이미지 검색 결과에 따른 비교평가를 수행하였다.

## 2. 관련 연구

### 2.1 웹2.0 환경에서의 태그

웹2.0이란 정보의 개방을 통해 인터넷 사용자들 간의 정보 공유와 참여를 이끌어내고, 이를 통해 정보의 가치를 지속적으로 증대시키는 것을 목표로 하는 일련의 움직임을 말한다. 즉, 웹2.0은 개방적인 웹 환경을 기반으로 네티즌이 자유롭게 참여해 스스로 콘텐츠를 생산, 재창조, 공유하는 기능을 수행한다[2].

웹2.0은 2004년 IT관련 컨퍼런스에서 O′Reilly사와 MediaLive사 간의 아이디어를 협의하는 과정에서 그 개념이 도출되었다[1]. 이는 2001년 닷컴 버블 붕괴 이후에도 생존하면서 지속적으로 성장한 구글, 아마존닷컴 등과 같은 성공한 인터넷 기업들이 공통점을 분석함으로써 도출되는 웹상의 새로운 트렌드를 포괄적으로 일컫는 말이기도 하다. 즉, 웹2.0은 인터넷의 새로운 패러다임으로서, 블로그와 검색으로 대표되는 인터넷 환경을 웹1.0으로 본다면 UGC(User Generated Contents)가 중심이 되는 새로운 인터넷 환경을 웹2.0이라 할 수 있으며, 지금보다 더욱 사용자가 중심이 된다는 것이 특징이다. 이와 같이 현재 큰 반향을 일으키고 있는 웹2.0 현상은 이제 모든 인터넷 사이트들의 필수 전략이 되었으며, 웹2.0을 성공적으로 구현하기 위한 다양한 기법들이 쏟아져 나오고 있다. 이러한 기법들 중 핵심적인 기술이 바로 태깅(tagging)이다[4,5]. 태깅은 현재 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 활용되고 있다.

### 2.2 태그 기반 검색에 관한 연구

현재 웹상의 모든 태그 기반 시스템들이 태그를 사용하는 여러 가지 목적 중 하나는 사용자들이 자발적으로 태깅한 태그 정보들을 기반으로 기존 검색 시스템보다 쉽고 정확한 검색결과를 제공할 수 있기 때문이다.

윤기상[9]은 태그를 기반으로 개인화 검색을 위한 연관태그 카테고리를 이용한 사용자 관심 프로파일 생성 방법 및 앞으로 더 많이 활용이 될 것으로 예상되는 태그를 웹 페이지 분석에 이용하였다. 또한 사용자의 선호도가 낮은 분야의 태그를 가진 웹 페이지들은 하위로 조정해주기 위한 방법을 제안하였으며, 제안된 방법이 기존 검색엔진보다 더 사용자의 요구를 만족시키는 효과가 있음을 보였다.

권대현[10] 등의 연구에서는 의미정보를 얻기 위해 엄격한 온톨로지를 적용하는 대신 간편한 온라인 어휘사전을 적용함으로써 그 효과를 높이려는 시도를 전개하였다. 이 연구에서는 프린스턴 대학에서 개발한 온라인 어휘사전인 ‘워드넷(WordNet)’을 사용하여 검색어 및 이미지 태그들의 의미정보를 획득한 후, 이를 바탕으로 한 웹 이미지 검색 결과를 보여주는데, Flickr 검색 결과에 비해 정확도의 향상이 뚜렷한 것으로 나타났다. 하지만 워드넷 자체의 한계로 말미암아 제한적인 의미 정보만 사용 할 수 있을 뿐이고, 따라서 첫 한, 두 페이지에서 월등히 높았던 정확도가 검색 페이지가 증가할수록 빠르게 줄어드는 것이 단점이라 할 수 있다.

강필구[11] 등은 태그를 효율적으로 관리하기 위한 시스템을 설계 및 구현하였다. 또한 태그 기반 시스템에 적합한 데이터베이스를 제시하고 연관 태그 및 대표 태그를 추출하는 방법 및 그를 통해 트리화된 구조로 검색 결과를 제시하는 연구를 진행하였다. 그러나 미리 구조화된 데이터베이스 및 연관 태그와 대표 태그를 사용자가 선택하는 방식을 사용한다는 단점을 가지고 있다.

### 2.3 태그 클러스터링 기법

Christopher[12]의 연구에서는 블로그 상에 존재하는 뉴스 문서들을 수집하여 TF-IDF(Term Frequency-Inverse Document Frequency)의 평가 방법론을 이용하여 유사문서 추출 및 그 과정에서 추출된 유사한 키워드들을 기반으로 자동 태깅을 위한 연구를 진행하였다. 또한 자동 태깅된 태그 정보들을 기반으로 계층적 클러스터링 알고리즘을 적용하여 중요도가 높은 태그부터 낮은 태그 순으로 구조화된 클러스터를 생성하는 방법론을 제안하였다.

그러나 이 연구에서 제안한 방법론은 특정 콘텐츠들에 대해서만 효율적인 것으로 분석되었으며, 또한 클러스터링 과정 중 어느 시점이 높은 태그이고, 어느 시점부터가 낮은 태그인지에 대한 명확한 제시는 못하고 있다. 이는 계층적 클러스터링이 가지는 특성에 의한 것으로서, Sahoo[13] 등에 따르면 계층적 클러스터링은 데이터 쌍들을 기반으로 하나의 계층적인 큰 클러스터를 생성하기 위한 알고리즘이라고 정의되어 있다. 이와 같은 계층적 클러스터링의 알고리즘의 경우는 비구조화된 정보인 태그들을 계층적인 구조로 구조화시키기에는 적합한 알고리즘이지만, 웹의 특성상 태깅된 태그들에는 부정확한 태그들이 많이 존재한다. 이와 같은 이유에서 하나의 큰 클러스터로 생성되는 계층적 클러스터링 알고리즘의 경우는 부정확하게 태깅된 태그들을 처리하기 위한 별도의 방법론이 필요하다.

Begelman[14] 등은 RSS를 이용하여 태그들을 수집하고 이를 기반으로 태그들을 좌표평면상에 표현하였다는 가정 하에 연관 태그들을 클러스터링 하기 위해 Spectral Bisection 알고리즘을 이용하는 방법론을 제안하였다. 이 방법은 태그 그래프를 양분하는 과정과 양분된 그래프가 또다시 양분될 필요가 있는지 검증하는 과정으로 구성된다. 따라서 위의 알고리즘은 재귀적으로 반복 진행되며, 태그 그래프를 클러스터 단위로 충분히 클러스터링 될 때까지 반복되는 방법이다. 그러나 그래프를 기반으로 하는 클러스터링 알고리즘들의 경우 태그가 가지는 근본적인 두 번째 문제점인 비구조화된 태그 정보를 좌표평면 상에 표현해야 된다는 큰 단점을 가지고 있다.

## 3. 3-태그 기반 검색

본 논문에서는 단일 태그로 인한 낮은 웹 리소스 검색의 효율성 향상을 위해 선행 연구에서는 웹상에 산재되어 있는 태그들을 연관성이 높은 태그들로 클러스터링 하기 위한 연구와 2-태그를 기반으로 검색에 활용하기 위한 연구를 [7,8]에서 진행하였다. 본 논문에서는 선행연구의 연관 태그들로 구성된 클러스터 내의 태그들을 검색에 활용하기 위해 3-태그 검색알고리즘을 제안 및 구현하였다.

### 3.1 실험 데이터

실험 데이터는 웹2.0의 선두주자라고 할 수 있는 Flickr Open API를 이용하여 이미지 및 그에 태깅된 태그 정보들을 수집하였으며, 키워드 'apple'을 통해 검색된 상위 리소스 120개 및 그에 태깅된 836개의 태그들을 실험 데이터로 선정하였다. 다음 그림 1은 실험 데이터의 일부를 보여주고 있다.

### 3.2 태그 클러스터링

선행연구[7,8]을 통해 생성된 가중치 행렬을 기반으로 연관도가 높은 태그들을 클러스터링 하기 위해 그림 2의 클러스터링 알고리즘을 제안하였다. 알고리즘은 CAST 알고리즘의 $\theta$값을 통해 클러스터링 해나가는 방법론을 응용하여 단일 연결 관계로 구성되는 방법론을, 연관 태그 추출을 통해 검색에 활용하기 위해 다중 연결 관계로 클러스터링한다.

클러스터링 알고리즘의 동작과정을 살펴보면, 첫 번째 단계로 가중치 행렬 중 최대 가중치를 가지는 tag $i$와 tag $j$를 클러스터 $C(i)$에 추가한다. 그 후, $C(i)$에 추가된 tag $i$와 tag $j$에 동시에 incident한 tag들 중 가중치 평균이 $\theta$보다 크거나 같은 원소 $T(i,j)$를 가중치 행렬 $T_G$에서 선택하여 $C(i)$에 추가하며, 태그들 간의 가중치 $T(i,j)$가 $\theta$보다 작게 될 때까지 이를 반복 진행하게 된다. 이러한 진행과정은 $\theta$(threshold)보다 큰 가중치를 가지는 모든 tag들이 클러스터에 포함될 때까지 반복 수행하며, 이를 통해 연관관계가 높은 태그들로 클러스터링하게 된다.

알고리즘을 적용한 결과 'apple-mac'과 관련된 20개의 태그로 구성된 클러스터 1(그림 3)과 'apple-ipod'과 관련된 14개의 태그로 구성된 클러스터 2(그림 4) 및 'apple-fruit'와 관련된 8개의 태그로 구성된 클러스터 3이 생성되었다. 그림 3과 4는 클러스터 1과 2의 결과 중 일부를 나타내었다.

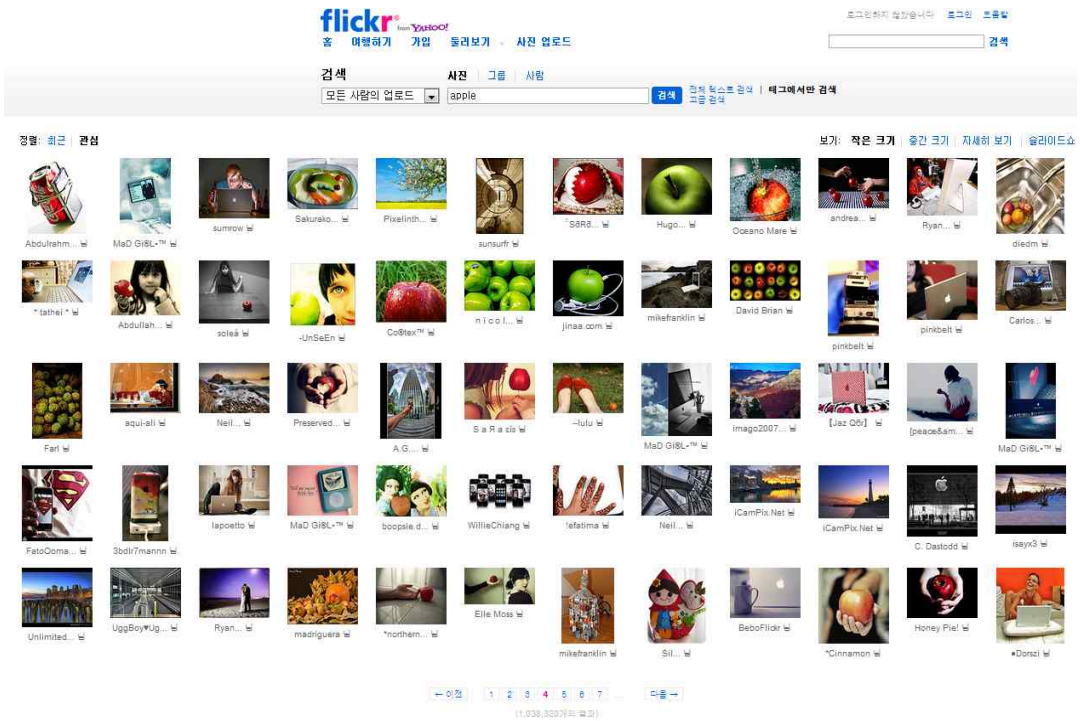


그림 1. 실험 데이터

```
// i : 클러스터 번호
// C(i) : i 번째 클러스터
// T(i,j) : tag i와 tag j간의 빈도수, 즉 가중치 행렬 T_G의 i행 j열
// Max(i,j) : T_G의 원소 중 최대 가중치를 가지는 원소
// A_i : 클러스터 C(i)에 포함된 태그들의 가중치 행렬

i=1
// Threshold θ보다 큰 가중치를 가지는 모든 tag들이 cluster에 포함될 때까지 반복

Initialize A_i

Repeat {

    //T_G에서 최대가중치를 가지는 원소 Max(i,j)의 두 태그 tag i와 tag j를 선택하여 클러스터 C(i)에 추가
    Select Max(i,j) Add tag i and tag j to C(i) Add element Max(i,j) to A_i

    //T(i,j) ≥ θ인 가중치를 가지는 tag i와 tag j가 C(i)에 모두 포함될 때까지 반복
     While(T(i,j) ≥ θ) {

        // 클러스터 C(i)의 가중치 행렬 A_i에 추가된 tag i와 tag j에 동시에 incident한 tag중 가중치 평균이
        θ보다 크거나 같은 원소 T(i,j)를 가중치 행렬 T_G에서 선택하여 C(i)에 추가
       Add  tag i and tag j of T_G to C(i) Add element T(i,j) to A_i

     }

     i= i+1

} until (all (T(i,j) ≥ θ)∈C(i))
```

그림 2. 클러스터링 알고리즘

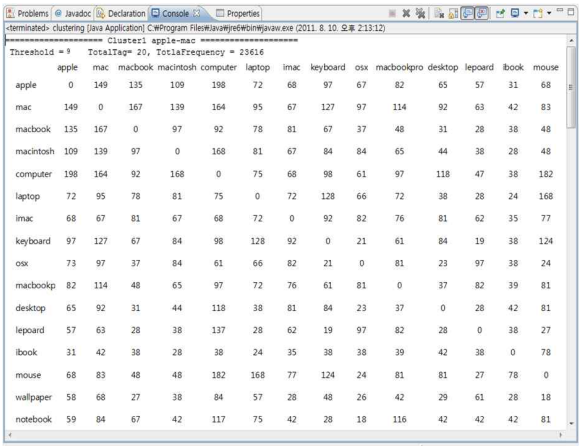


그림 3. 생성된 태그 클러스터 1

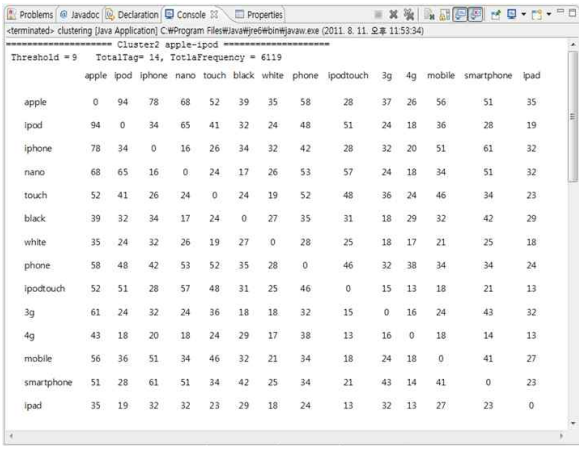


그림 4. 생성된 태그 클러스터 2

### 3.1 3-태그 검색

생성된 클러스터 내의 태그들을 검색 시스템에 적용하기 위해서는 그에 적합한 검색 기법이 필요하며, 본 논문에서는 빈도수로 구성된 클러스터 내의 태그들 간의 연관관계를 활용하여 다중 태그를 추출하기 위해 3-태그 검색 알고리즘을 제안하였다.

알고리즘의 진행 과정은 크게 세 단계로 구성되며, 첫 번째 과정으로 생성된 클러스터 내의 가중치로 연결된 태그 $A(i,j)$중 최대 가중치를 가지는 두 태그 Max($A(i,j)$)를 선택한다. 두 번째로 Max($A(i,j)$)와 동시에 incident한 태그들 중 가중치 평균이 가장 높은 tag $k$를 추출하여 MaxWeight($A(i,j,k)$)를 추출하며, 마지막으로 3개의 태그로 구성된 MaxWeight ($A(i,j,k)$)를 포함하고 있는 리소스를 TagDB에서 추출한다. 이러한 과정은 $A(i,j)$에서 MaxWeight ($A(i,j,k)$)가 empty될 때가지 반복 진행된다.

3-태그 추출 결과 클러스터 1은 42쌍의 연관 태그 쌍과 클러스터 2는 24쌍의 연관 태그 쌍을 추출하였으며, 클러스터 3은 18쌍의 연관 태그 쌍을 추출하였다.

표 1은 클러스터 1의 연관 태그 12쌍 중 3-태그 추출 알고리즘의 MaxWeight(A(i,j,k))에 해당하는 연관 태그로서 Tag1, Tag2, Tag3는 각각의 가중치로 구성되어 있으며, Weight 값은 각각의 태그들 간의 가중치의 합이다.

클러스터 내의 연관 태그를 확장하여 3-태그 간의 관계(가중치)를 활용함에 따라 사용자가 의도하는 연관 검색 태그들을 추출 가능하다. 그림 5는 클러스터 1의 'apple-computer' 태그를 중심으로 추출된 3-

표 1. 클러스터 1의 3-태그 추출 결과

| Cluster 1(3-Tag) MaxWeight(A(i,j,k)) | | | |
|---|---|---|---|
| Tag1 | Tag2 | Tag3 | Weight |
| apple | computer | mac | 170 |
| apple | computer | macintosh | 158 |
| apple | computer | mouse | 149 |
| apple | computer | macbook | 141 |
| apple | computer | keyboard | 141 |
| apple | computer | desktop | 127 |
| apple | computer | macbookpro | 117 |
| apple | computer | laptop | 115 |
| apple | computer | imac | 111 |
| apple | computer | osx | 108 |
| apple | computer | lepoard | 100 |
| apple | computer | ibook | 89 |

```
//n : cluster의 개수
//C(num) : num번째 클러스터
//A(i,j) : 클러스터 C(num)에 포함된 태그들의 가중치 행렬
//Max(A(i,j)) : 가중치 행렬 A(i,j)의 원소 중 최대값을 가지는 원소로서, tag i와 tag j의 가중치
//MaxWeight(A(i,j,k)) : Max(A(i,j))와 incident한 tag들 중 가중치 평균이 가장 큰 tag k
//TagDB : A(i,j)의 tag i와 tag j를 가지는 모든 tag들의 집합

//클러스터 C(num)의 개수만큼 반복
for (num=1; num>=n; num++) {

    //A(i,j)에서 Max(A(i,j))가 empty가 될 때까지 반복
    Repeat {

        //가중치 행렬 A(i,j)의 원소 중 최대값을 가지는 Max(A(i,j))선택
        Find Max(A(i,j)) in A(i,j)

        //A(i,j)에서 MaxWeight(A(i,j,k))가 empty가 될 때까지 반복
        Repeat {

            //Max(A(i,j))와 incident한 tag들 중 가중치 평균이 가장큰 MaxWeight(A(i,j,k))를 A(i,j)에서 선택
            Find MaxWeight(A(i,j,k)) in A(i,j)

            //TagDB에서 MaxWeight(A(i,j,k))의 tag i, j, k를 가지는 모든 이미지 검색
            Search all images including tag i, tag j and tag k in TagDB

            //A(i,j)에서 MaxWeight(A(i,j,k))는 삭제
            Remove MaxWeight(A(i,j,k)) from A(i,j))

        } until (empty MaxWeight(A(i,j,k)) from A(i,j))

        //A(i,j)에서 Max(A(i,j))는 삭제
        Remove Max(A(i,j)) from A(i,j))

    } until (empty Max(A(i,j)) from A(i,j))

}
```

그림 5. 3-태그 기반 검색 알고리즘

태그들을 기반으로 검색에 적용한 결과를 나타내었으며, 그림 6은 클러스터 3의 이미지 검색 결과를 나타내었다.

이와 같이 추출된 3-태그 기반 검색은 연관 태그 확장을 통해 검색 태그들이 늘어날수록 검색 결과 또한 향상되는 결과를 도출하였다. 그러나 다수의 태그를 활용함에 따라 그에 부합하는 검색 결과 또한 줄어들었다. 이는 3-태그 검색 방식은 리소스에 태깅된 태그가 추출된 3-태그를 모두 포함해야 검색되어지기 때문이다.

## 4. 실험 및 평가

본 논문에서는 태그가 가지고 있는 근본적인 문제점인 부정확한 태그로 인한 낮은 검색 결과의 문제점을 해결하기 위한 방안으로 3-태그 기반 검색 기법을 제안하였다. 제안한 알고리즘의 성능 및 평가를 위해 기존 태그 기반시스템의 단일태그 매칭 기법 및 선행연구에서 제안한 2-태그 쌍 검색기법과의 이미지 검색결과에 대한 평가를 진행하였으며, 평가항목은 다음 표 2와 같다.

실험 데이터로는 Flickr Open API를 이용하여 'computer' 태그를 키워드로 1-10page에 해당하는 상위 이미지 480개와 그에 해당하는 태그 3,261개를 실험 데이터로 활용하였다.

표 2. 평가 항목

| |
|---|
| 1. 단일 태그 검색(기존 태그 검색 방식) |
| 2. 2-태그 쌍 검색(기존 선행연구 방식) |
| 3. 3-태그 검색(제안 방식) |

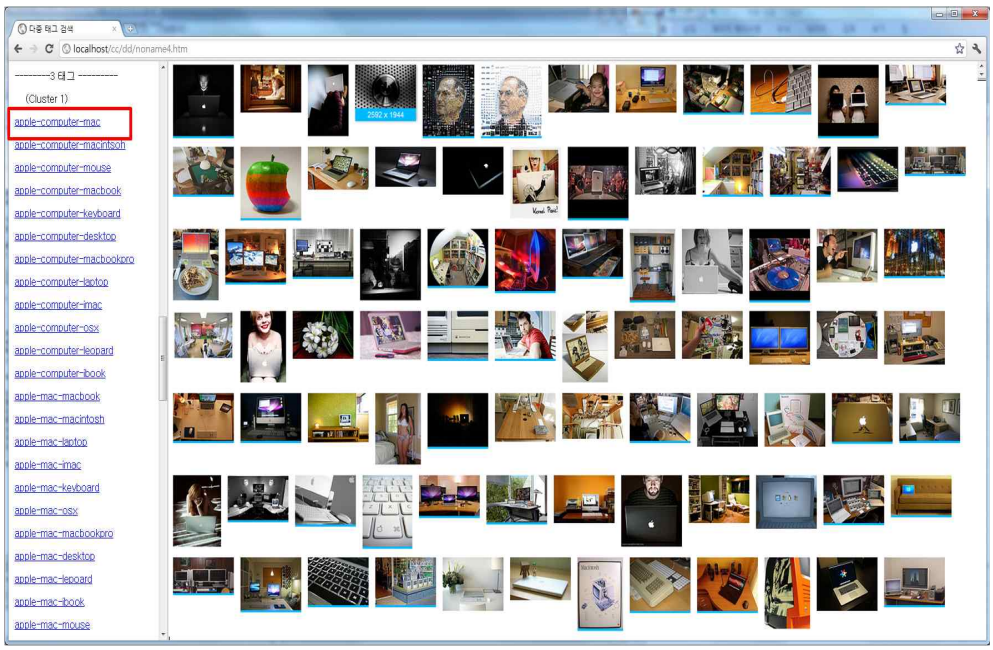


그림 6. 클러스터 1의 3-태그 검색 결과

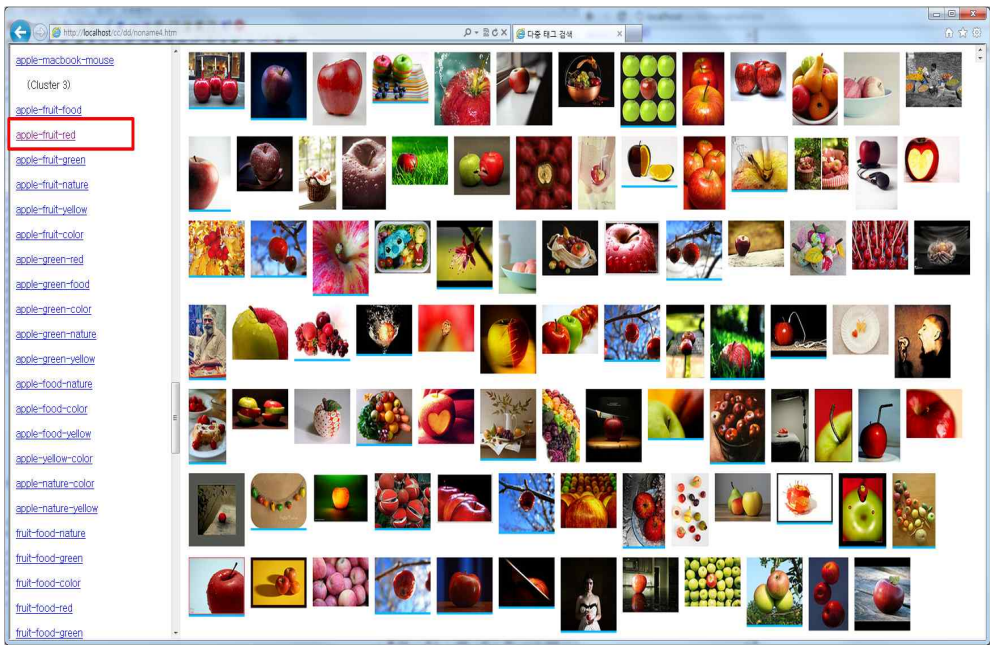


그림 7. 클러스터 3의 3-태그 검색 결과

평가 기준은 검색된 이미지들 중 이미지 내에 컴퓨터와 관련된 이미지를 포함하고 있으면, "정확", 그렇지 않으면, "부정확"으로 정의하여, 정확성(precision), 재현율(recall)을 평가하였다[12].

표 3은 3가지 방법론을 적용하여 실험데이터에 따른 정확도와 재현율을 평가한 결과를 나타내고 있다. 키워드 'apple'의 경우는 과일을 의미하는 'apple'과 회사관점에서 'apple'을 의미하는 2가지 속성을 가지고 있다. 기존 방식에 경우는 'apple'이라는 태그를 포함하는 모든 이미지를 검색결과로 제공함에 따라 그림 1과 같이 두 가지 의미를 포함하는 모든 이미지를 출력하는 문제점을 보이고 있다. 그에 반해 본 연구에서 제안한 2-태그와 3-태그를 기반으로 하는 검색의 경우는 태그 클러스터링을 통해 클러스터별로

표 3. 이미지 검색에 따른 정확도와 재현율

| 키 워 드 | 검색 방법 | 정확도 | 재현율 |
|---|---|---|---|
| 키워드 'computer'의 평가 | 단일 태그 | 49.1% | 49.1% |
| | 2-태그 쌍 | 82.2% | 42.5% |
| | 3-태그 | 90.0% | 36.3% |
| 'fruit' 관점에서의 'apple' | 단일 태그 | 34.1% | 34.1% |
| | 2-태그 쌍 | 91.3% | 34.1% |
| | 3-태그 | 96.2% | 31.6% |
| 'company' 관점에서의 'apple' | 단일 태그 | 53.3% | 53.3% |
| | 2-태그 쌍 | 92.9% | 43.3% |
| | 3-태그 | 95.9% | 37.3% |

검색에 적용함에 따라 여러 속성을 가지는 이미지들의 분류 및 정확성이 매우 향상된 결과를 도출하였다.

표 3에서의 단일 태그 검색 결과 평균 45%의 낮은 정확도 및 45%의 재현율을 보여주고 있다. 이는 키워드인 'computer'와 'apple'을 포함하는 모든 이미지들을 출력하는 단순 태그 매칭에 따른 태그가 가지는 첫 번째 문제점인 부정확한 태그로 인한 낮은 검색 결과의 원인이다.

선행 연구로 진행된 2-태그 쌍 검색 결과 평균 82.2%의 정확도 및 42.5%의 재현율을 도출하였다. 재현율이 다소 떨어진 이유는 그림 1의 8번째 이미지에 'city', 'computer', 'sky', 'mac', 'newyork', 'urban', 'tall', 'apple'과 같이 부정확한 태깅의 경우로써, 제안한 다중 태그 검색을 이용한 리소스 추출에 의한 검색 결과의 측면에서는 정확한 결과이다. 그러나 2-태그 쌍 검색을 적용한 결과 2 page의 1개, 4 page의 2개의 이미지들의 경우 키워드 'computer', 'apple'을 포함하고 있지만 추출하지 못한 경우이다. 이는 2 page 23번째 이미지에 태깅된 태그는 'computer', 'graphic'으로 본 논문에서 제안한 태그 쌍에 의해 추출된 태그들 중 'computer', 'apple'의 태그만을 포함하고 있음에 따라서 추출되지 못한 경우로써, 이러한 제안 시스템의 문제점은 향후 해결해야하는 문제점이다.

본 논문에서 제안한 3-태그 검색 결과는 평균 94%의 정확도 및 35%의 재현율을 나타내었다. 기존 단일 태그 매칭 기법에 비해 정확도는 49% 향상되었으며, 재현율 면에서는 10% 다소 떨어진 검색결과를 도출하였다. 선행 연구로 진행된 2-태그 쌍 검색 기법과의 비교에서는 정확도는 5% 향상되었으며, 재현율은 4% 떨어진 결과를 도출하였다. 재현율이 4% 감소한 이유는 여러 태그를 이용하여 검색에 활용할 경우 정확도면에서는 향상되는 결과를 도출하지만, 다수의 태그를 포함하는 이미지들은 적어짐에 따라 재현율 면에서는 감소하는 결과를 도출하였다. 또한 2-태그 기반 검색 기법의 문제점인 부정확한 단일태그를 포함하는 이미지의 문제점을 3-태그 기반에서도 동일한 문제점이다.

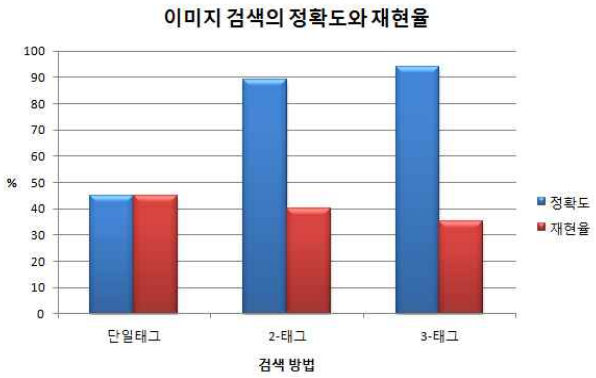


그림 8. 정확도와 재현율 평균

다음 그림 7은 3가지 방법론들의 이미지 검색 결과에 따른 정확도와 재현율 평균을 나타내었다.

## 5. 결 론

본 연구에서는 대표적인 태그 기반 사이트인 Flickr의 콘텐츠를 이용하여 리소스와 그에 따른 태그 정보들을 수집하여 태그들 간의 연관성에 따라 맵핑하고, 이를 클러스터링하기 위한 시스템을 설계 및 구현하였으며[7,8], 본 논문에서는 이를 기반으로 향상된 검색 서비스 제공을 위한 3-태그 기반 검색 알고리즘을 제안하였다. 또한 기존 단일 태그 매칭 기법 및 선행연구에서 제안한 2-태그 기법과의 이미지 검색 결과에 따른 비교·평가를 수행하였다.

평가 결과 기존 단일 태그 매칭 기법에 비해 정확도는 49% 향상되었으며, 재현율 면에서는 10% 다소 떨어진 검색결과를 도출하였다. 선행 연구로 진행된 2-태그 쌍 검색 기법과의 비교에서는 정확도는 5% 향상되었으며, 재현율은 4% 떨어진 검색 결과를 도출하였다.

향후 연구 과제로는 연관태그 맵핑 과정에서 연관관계가 정의되지 않은 이미지들로 인한 재현율 저하의 문제를 해결하기 위해 이미지의 특징 정보를 추출

하여 태깅된 태그와 맵핑하기 위한 연구가 필요하다.

## 참 고 문 헌


[1] 이석용, 정이상, "웹2.0 시대의 SNS(Social Network Service)," 경영정보학회지. 제29권, 제5호, pp. 143-166, 2010.

[2] Time O'Reilly, "*What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software,*" Communications & Strategies, pp. 17-37, 2007.

[3] 이시형, 노용만, "자동 이미지 태깅 기술 동향," 한국정보통신산업진흥원, 주간기술동향, 제1427호, pp. 1-8, 2010.

[4] 한성국, 정영식, 유재규, "웹2.0과 시맨틱웹, 그리고 진화의 방향," 정보과학회지, 제25권, 제10호, pp. 57-66, 2007.

[5] Lin Chen, "Tag-based Web Photo Retrieval Improved by Batch Mode Re-tagging," IEEE International Conference on Computer Vision and Pattern Recognition, pp.3440-3446, 2010.

[6] Martha Larson and Mohammand Soleymani, "Automatic Tagging and Geotagging in Video Collections and Communities," *Proceedings of the 4th ACM Conference*, pp. 51:1 - 51:8, 2011.

[7] 이시화, 이만형, 김용수, 황대훈, "Web 2.0에서의 효율적인 Tag Clustering을 위한 Threshold 선정에 관한 연구," 멀티미디어학회 추계 학술대회, pp. 22-25, 2007.

[8] 이시화, 이만형, 황대훈, "web2.0 환경에서의 효율적인 이미지 검색을 위한 태그 클러스터링 시스템의 설계 및 구현," 멀티미디어학회논문지, 제11권, 제8호, pp. 1169-1178, 2008.

[9] 윤기상, 윤광호, 김재광, 이지영, "태그를 이용한 개인화 검색 시스템," 한국과학회 추계 학술대회, pp. 320-323, 2009.

[10] 권대현, 홍준혁, 조수선, "워드넷 의미정보로 선별된 우선 태그와 이를 이용한 웹 이미지의 검색," 멀티미디어학회논문지, 제12권, 제7호, pp. 1032- 1042, 2009

[11] 강필구, 김남중, 이예슬, 채진석, "웹2.0을 위한 효율적인 태그 관리 시스템의 설계 및 구축," 한국정보과학회 추계 학술대회, pp. 472-476, 2006.

[12] Christopher H. Brooks and Nancy Montanez, "Improved Annotation of the Blogosphere Via Autotagging and Hierarchical Clustering," *International Conference on World Wide Web*, pp. 625-632, 2006.

[13] Nachiketa Sahoo and Jamie Callan, "Incremental Hierarchical Clustering of Text Documents," *ACM Int. Conf. on Information and Knowledge Management,* pp. 357-366, 2006.

[14] S. M. Shafi and Rafiq A. Rather "Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology," *Webology,* Vol. 2, No. 2, 2005.



**이 시 화**

2005년 서울보건대학 컴퓨터정보과 졸업
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과(석사)
2012년 가천대학교 전자계산학과(박사)
관심분야 : e-Learning, Context- Aware, Semantic Web, Web2.0, Tag



**황 대 훈**

1997년 동국대학교 수학과(학사)
1983년 중앙대학교 전자계산학과(석사)
1991년 중앙대학교 전자계산학과 (박사)
1983년~1985년 한국산업경제기술연구원(KIET) 연구원
2009년~2010년 한국멀티미디어학회 회장
1987년~현재 가천대학교 교수
관심분야 : e-러닝, Semantic Web, Web2.0, Cloud