Journal of Korea Multimedia Society Vol. 11, No. 8, August 2008(pp. 1169-1178)

# Web2.0 환경에서의 효율적인 이미지 검색을 위한 태그 클러스터링 시스템의 설계 및 구현

이시화[†], 이만형[††], 황대훈[†††]


## 요 약

웹 2.0에서 대부분의 정보는 사용자에 의해 생산되고, 사용자가 붙인 태그에 의해 분류되어진다. 현재 태그와 연관된 서비스 및 연구들은 자동 태깅 기법이나 태그 클라우드 구성 기술에 초점이 맞춰 진행되어짐에 따라, 태그에 의해 분류되어진 정보 및 리소스들을 효율적으로 분류하여 사용자에게 제공하는 연구는 미흡한 실정이다. 이에 본 논문에서는 웹상에 산재되어있는 리소스 및 그에 따른 태그 정보들을 수집하여 태그들 간의 연관성에 따라 맵핑하고, 이를 클러스터링하여 검색에 적용하기 위한 시스템을 설계 및 구현하였다. 또한 제안 시스템의 성능평가를 위해 태그 기반 대표사이트인 플리커 사이트의 이미지 검색 결과와의 정확성과 재현율을 비교 평가함으로서 향상된 검색결과를 제시하였다.


# Design and Implementation of Tag Clustering System for Efficient Image Retrieval in Web2.0 Environment

Si-Hwa Lee[†], Man-Hyoung Lee[††], Dae-Hoon Hwang[†††]


## ABSTRACT

Most of information in Web2.0 is constructed by users and can be classified by tags which are also constructed and added by users. However, as we known, referring by the related works such as automatic tagging techniques and tag cloud's construction techniques, the research to be classified information and resources by tags effectively is to be given users which is still up to the mark. In this paper, we propose and implement a clustering system that does mapping each other according to relationships of the resource's tags collected from Web and then makes the mapping result into clusters to retrieve images. In addition, we analyze our system's efficiency by comparing our proposed system's image retrieval result with the image retrieval results searched by Flickr website.


**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 웹 2.0이다[1]. 웹 2.0에서 대부분의 정보는 사용자에 의해 생산되고, 웹 페이지, 사진, 웹 링크와 같은 다양한 콘텐츠들에 태그를 이용하여 정보들을 체계화시킨다. 또한 이를 공유함으로써 다양한 정보자원간의 체계와 연결관계를 만들 수 있도록 하는 것이다. 태그(tag)는 학

※ 교신저자(Corresponding Author) : 황대훈, 주소 : 경기도 성남시 수정구 복정동 산 65번지(461-701), 전화 : 031)750-5327, FAX : 031)757-6715, E-mail : hwangdh@kyungwon.ac.kr
접수일 : 2008년 3월 3일, 완료일 : 2008년 5월 27일

[†] 준회원, 경원대학교 전자계산학과 박사과정
(E-mail : leesihwaman@gmail.com)
[††] 준회원, 현대전문학교 교수
(E-mail : atomv@nate.com)
[†††] 종신회원, 경원대학교 교수

생들의 이름표, 수하물의 딱지, 제품의 상표를 뜻하며[1], 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 적용되고 있다[2]. 현재 태그에 관한 서비스 및 연구들은 자동 태깅 기술과 효과적인 태깅 방법에 대한 연구를 비롯해 tag cloud 구성 기술, 다중 응용에서의 협업태깅 기술에 많은 연구들이 진행되고 있다[1]. 그러나 태깅에 사용된 태그가 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 태그에 의해 분류되어진 정보 및 리소스들을 효율적으로 분류하여 활용하지 못하고 있는 실정이다.

이에 본 논문에서는 웹상에 산재되어있는 리소스 및 그에 따른 태그 정보들을 수집하여 태그들 간의 연관성에 따라 맵핑하고, 이를 클러스터링하여 검색에 활용하기 위한 시스템을 설계 및 구현하였다. 또한 클러스터링된 결과를 기반으로 태그기반 대표 사이트 중 하나인 플리커 사이트[3]의 이미지검색 결과와 본 논문에서 제안하는 클러스터링을 통한 이미지 검색 결과와의 정확성과 재현율의 비교평가를 통해 향상된 검색결과를 제시한다.

## 2. 관련 연구

### 2.1 웹2.0환경에서의 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[1].

이러한 웹2.0 환경에서의 핵심 기술은 태그이며, 태그(tag)는 어떠한 정보에 대하여 사용자가 직접 만드는 메타데이터(metadata)를 의미한다. 여기서 정보란 웹상에 존재하는 모든 형태의 정보들을 가리킨다[4]. 태그는 현재 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 활용되고 있다[2].

태깅은 웹상에 사용자가 원하는 아이템들을 그들이 원하는 방식대로 묘사하도록 하는 과정으로, Flickr, Del.icio.us, Technorati 등과 같은 웹 사이트는 사용자들이 특정 아이템을 묘사하기 위해 사용한 태그를 메타데이터로 검색에 활용하는 태깅 서비스를 성공적으로 구현하여 서비스하고 있다[5]. 하지만 태깅에 사용된 태그가 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 실제로는 태그가 가지는 한계로 인해 만족스럽지 못한 검색결과를 도출하고 있다[4,6]. 이러한 문제들을 개선하려는 노력의 일환으로 태그 구름, 협업 태깅, 연관 태그 추출, 온톨로지를 태깅에 접목하는 등의 연구가 진행되어 오고 있으나, 이렇다 할 결과를 도출하지 못하고 있는 실정이다.

### 2.2 태그 기반 시스템의 한계

2.1절에서와 같이 태그는 매우 유연하고 역동적인 분류체계를 제공한다. 하지만 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계를 가지고 있는 것이 사실이다. 태그기반 시스템에는 다음과 같은 근본적인 두 가지 문제점을 가지고 있다.

첫째, 태그 기반 검색 시스템은 검색에 있어서 정확도(precision)가 떨어진다. 이강표 등[4]에 따르면, 태그는 어떤 정보를 넓은 범주의 카테고리에 위치시키는 데에는 매우 유용하지만, 사용자가 원하는 정확한 정보를 찾아내는 데에는 효율적이지 못함을 보이고 있다. 이는 리소스에 태깅된 태그들 중 부정확한 태그들이 많이 존재하기 때문이며, 이로 인해 낮은 검색결과의 원인이 된다.

둘째, 태깅된 태그는 비구조화된 데이터이다. 그 예로서 어떠한 이미지에 태깅된 mac, imac, desktop, setup 태그들의 경우 그림을 설명하기 위해 태깅된 태그로서 서로 어떠한 관계가 있는지를 판단할 수 없다. 이는 태깅의 주 목적 중의 하나인 태그를 통한 정보 네비게이션을 방해하는 원인이 되며, 정보 네비게이션에 대한 검색결과 또한 첫 번째 문제에 따라 낮은 검색결과를 제공할 수 밖에 없다.

이러한 태그가 가지는 근본적인 두 가지 문제점 중 본 연구에서는 첫 번째 문제점인 부정확하게 태깅된 태그로 인한 낮은 검색의 효율성을 향상시키기 위한 연구를 중심으로 진행하였다.

### 2.3 태그 클러스터링 관련 연구

태그가 가지는 첫 번째 문제점을 해결하기 위한 연구의 일환으로 클러스터링 기법들을 적용한 연구들이 진행되고 있으며, 현재까지 이루어진 연구들을 중심으로 장단점을 분석하였다.

Christopher 등[7]은 블로그 상에 존재하는 뉴스 문서들을 수집하여 TF와 IDF의 평가방법론을 이용하여 유사문서 추출 및 그 과정에서 추출된 유사한 키워드들을 기반으로 자동 태깅을 위한 연구를 진행하였다. 또한 자동 태깅된 태그 정보들을 기반으로 계층적 클러스터링 알고리즘[8]를 적용하여 중요도가 높은 태그부터 낮은 태그 순으로 구조화된 클러스터로 생성하는 방법론을 제안하였다. 그러나 이 연구에서 제안한 방법론은 특정 콘텐츠들에 대해서만 효율적인 것으로 분석되었으며[7], 또한 클러스터링 과정 중 어느 시점이 높은 태그이고 어느 시점부터가 낮은 태그인지에 대한 명확한 제시는 못하고 있다. 이와 같이 계층적 클러스터링 알고리즘을 적용할 경우, 비구조화된 태그들을 계층적인 구조로 구조화시키기에는 적합하지만, 웹의 특성상 태깅된 태그들에는 부정확한 태그들이 많이 존재하며, 이를 처리하기 위한 별도의 방법론이 필요하다.

Begelman 등[9]은 RSS를 이용하여 태그들을 수집하고 이를 기반으로 태그들을 좌표평면 상에 표현하였다는 가정 하에 연관 태그들을 클러스터링 하기 위해 Spectral Bisection 알고리즘[10]을 이용하는 방법론을 제안하였다. 이 방법은 태그 그래프를 양분하는 과정과 양분된 그래프가 또다시 양분될 필요가 있는지 즉, 양분 과정을 마칠 것인지를 검증하는 과정을 통해 클러스터 단위로 충분히 클러스터링될 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.

이와 같이 현재 태그 클러스터링을 통해 연관 태그를 추출하기 위한 연구들이 진행되어 지고 있지만 태그가 가지는 근본적인 두 가지 문제점을 해결하지 못하고 있는 실정이다.

## 2.4 CAST 알고리즘

CAST(Complexity Analysis of Sequence Tracts) 알고리즘은 그래프 기반의 알고리즘으로서, 바이오 인포메틱스 분야에서 유전자들 사이의 평균 거리를 기반으로 연관관계가 높은 유전자들을 클러스터링하기 위한 검증된 알고리즘이다[11].

CAST 알고리즘은 클러스터 구분의 근거가 되는 $\theta$값을 기준으로 가까운 유전자는 클러스터에 포함시키고, 멀리 있는 유전자는 제외시키는 일련의 동작 과정으로 연관관계가 높은 유전자들로 구성된 클러스터를 생성하는 기법으로서, 클러스터에 포함된 유전자들은 단일 연결 구조를 가진다.

본 논문에서는 검증된 CAST 알고리즘의 $\theta$값을 기반으로 클러스터링하는 방법론을 응용하여 3.2절에서 클러스터링 알고리즘을 제안한다.

## 3. 제안 시스템

본 장에서는 관련연구를 통해 2.3절에서 제시한 태그가 가지는 두 가지 문제점 중 첫 번째 문제점인 부정확한 태그들로 인한 낮은 검색의 정확도를 향상시키기 위해 그림 1의 태그 클러스터링을 통한 이미지 검색 시스템을 제안한다.

### 3.1 가중치 매트릭스 생성 방법론

연관성이 높은 태그 그룹으로 클러스터링 하기 위한 첫 번째 단계로 연관 태그들 간의 맵핑을 수행한다. 여기에서 맵핑이란 동일한 리소스에 대하여 서로 다른 사용자에 의해 부여된 태그들을 서로 관련이 있는 태그로 연관관계를 부여하는 것이다. 예를 들어 어떠한 이미지 1에 대해 태깅된 태그(teacher, school, classroom, me, female, woman)들은 서로 다른 사용자들이 이미지 1에 대하여 보고 느낀 것을 태그 정보로 표현한 것이다. 따라서 동일한 리소스에 대한 상이한 태그들은 서로 연관성이 있다고 정의할 수 있다. 이러한 연관성을 이용하여 이미지들에 태깅된 태그들을 서로 맵핑하게 되며, 연관태그 맵핑 모듈에 의해 수행된다[12,13].

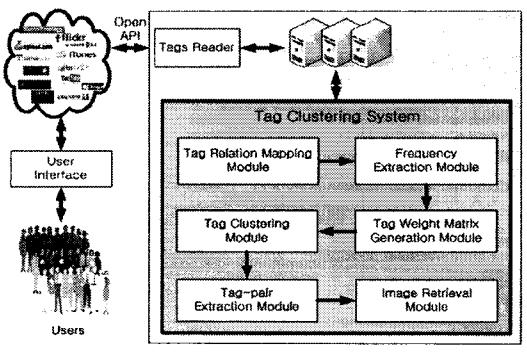


그림 1. 태그 클러스터링을 통한 이미지 검색 시스템

또한 빈도수 추출 모듈은 태그 맵핑 과정에서 태그들 간의 연관관계와 동일 태그의 출현 빈도를 이용하여 태그 빈도수를 추출하며, 이를 기반으로 태그 가중치 행렬 생성 모듈은 가중치 행렬(weight matrix)을 생성하게 된다. 여기서 사용되는 빈도수 추출 방식은 [12]와 같다.

### 3.2 태그 클러스터링 방법론

3.1에서 생성된 가중치 행렬을 기반으로 연관도가 높은 태그들을 클러스터링 하기 위해 그림 2의 클러스터링 알고리즘을 제안한다. 알고리즘은 CAST 알고리즘의 $\theta$값을 통해 클러스터링 해나가는 방법론을 응용하여 제안하였으며, 또한 CAST 알고리즘의 단일 연결 관계로 구성되는 방법론을, 연관 태그 추출을 통해 검색에 활용하기 위해 다중 연결 관계로 클러스터링하는 방법론을 제안한다.

```
// 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

클러스터링 알고리즘을 크게 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.

```
//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
//T(k) : Max(A(i,j))의 tag i와 tag j에 incident한 모든 tag들의 집합
//B(l,m) : T(k) 에 포함된 태그들의 가중치 행렬
//Max(B(l,m)) : 가중치 행렬 B(l,m)의 원소 중 최대값을 가지는 원소로서, tag l과 tag m의 가중치
//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)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)
        //TagDB에서 Max(A(i,j))의 tag i와 tag j를 가지는 모든 이미지 검색 Search all images including tag i and tag j in TagDB
        //Max(A(i,j))의 tag i와 tag j에 incident한 모든 tag를 C(num)에서 탐색하여 T(k)를 구성
        Find T(k)
        //B(l,m)이 empty가 될 때까지 반복
        Repeat {
            //가중치 행렬 B(l,m)의 원소 중 최대값을 가지는 Max(B(l,m)) 선택
            Find Max(B(l,m))
            //TagDB에서 Max(B(l,m))의 tag l과 tag m를 가지는 모든 이미지 검색
            Search all images including tag l and tag j in TagDB
            //B(l,m)에서 Max(B(l,m))는 삭제
            Remove Max(B(l,m)) from B(l,m)
        } until (B(l,m)==empty)
        /A(i,j)에서 B(l,m)는 삭제
        Remove B(l,m) from A(i,j)
    } until (A(i,j)==empty)
}
```

그림 3. 클러스터 기반 검색 알고리즘

이러한 진행과정은 $\theta$(threshold)보다 큰 가중치를 가지는 모든 tag들이 클러스터에 포함될 때까지 태그 클러스터링 모듈에 의해 반복 수행하며, 이를 통해 연관관계가 높은 태그들로 클러스터링하게 된다.

### 3.3 태그 쌍 기반 검색 방법론

3.2절의 태그 클러스터링 방법론에 의해 행렬로 생성된 클러스터를 검색시스템에 적용하기 위해서는 그에 적합한 검색 기법이 필요하며, 이를 위해 클러스터 기반 검색 알고리즘을 다음 그림 3과 같이 제안한다.

알고리즘의 진행과정을 살펴보면, 클러스터링된 클러스터 *C(num)*에 포함된 태그들의 가중치 행렬 *A(i,j)*의 원소 중 최대 가중치를 가지는 *Max(A(i,j))*를 선택하여 선택된 tag *i*와 tag *j*를 가지는 모든 이미지를 *TagDB*에서 추출한다. 그 후, 선택된 *Max(A(i,j))*의 tag *i*와 tag *j*에 incident한 모든 tag를 *C(num)*에서 탐색하여 *T(k)*를 구성하며, *T(k)*에 포함된 가중치행렬인 *B(l,m)*의 원소중 최대값을 가지는 원소 *Max(B(l,m))*를 선택하여 *tag l*과 *tag m*를 가지는 모든 이미지를 *TagDB*에서 추출한다. 그 후, *B(l,m)*에서 *Max(B(l,m))*는 삭제되며, *B(l,m)*이 *empty*가 될 때까지 반복한다.

이러한 과정은 클러스터 *C*의 가중치 행렬 *A(i,j)*가 empty가 될 때까지 반복하며, 클러스터 *C(num)*의 개수만큼 반복 진행한다.

## 4. 구현 및 실험

### 4.1 시스템 환경

본 논문에서 제안한 시스템의 각 모듈별 방법론을 구현 및 실험하기 위해 Windows 2000 Server 및 MySql Server 5.0 환경에서 J2SDK 1.5를 활용하여 구현하였으며, 또한 실험에 사용한 데이터는 웹2.0의 선두주자인 Flickr의 Open API[14]를 이용하여 이미지 및 그에 태깅된 태그 정보들을 수집하였다. 수집된 데이터는 키워드 computer를 통해 검색된 상위 1~5 page에 해당하는 리소스 120개 및 태깅된 836개의 태그들을 실험 데이터로 선정하였다.

### 4.2 가중치 매트릭스의 생성

수집된 실험 데이터를 기반으로 높은 연관관계로 구성된 태그 클러스터로 생성하기 위한 첫 번째 단계는 연관 태그 추출을 통한 가중치 매트릭스 생성이다. 이를 위해 3.2에서 제안한 연관 태그 맵핑 방법론을 적용하여 연관 태그들을 추출하였다[12].

그 결과 실험데이터 836개의 태그 중 연관 태그 맵핑을 통해 276개의 태그가 추출되었다. 또한 태그 맵핑 과정에서 태그들 간의 연관관계와 동일 태그의 출현 빈도를 이용하여 빈도수를 추출하며, 이를 위해 [12]에서 제안한 방법론을 적용하여 빈도수를 추출하였다. 이와 같이 제안한 연관 태그 맵핑 및 빈도수 추출 과정을 통해 태그 클러스터링을 위한 기초데이터인 가중치 매트릭스를 생성하였으며, 그 결과 276개의 행과 열로 구성된 가중치 매트릭스가 생성되었다[12].

### 4.3 클러스터 생성을 위한 실험 및 분석

그림 2에서 제안한 클러스터링 알고리즘의 $\theta$값은 생성된 가중치 매트릭스에 포함된 태그들을 클러스터에 포함할지 여부를 결정하는 중요한 임계치로써, 실험 및 분석을 통해 실험 데이터에 적합한 $\theta$값을 선정한다.

실험은 2~12의 $\theta$값을 순차적으로 시스템에 적용하여 생성된 클러스터 및 태그 수를 측정하였으며, 이를 기반으로 $\theta$값에 따라 생성된 클러스터들의 응집도를 분석하였다.

2~12의 $\theta$값만을 적용한 이유는 $\theta$값 1은 연관 태그 맵핑에 의해서 생성되는 기본 가중치이며, 12까지로 한정지은 이유는 12이후의 $\theta$값에 의해 생성되는 클러스터의 경우 단일 태그 쌍으로 생성되는 클러스터가 대다수이기 때문에 응집도 평가가 무의미하다.

$\theta$값을 시스템에 적용한 결과 그림 4는 $\theta$값에 따라 생성된 클러스터 내의 태그 수 평균을 산출한 결과를 보여주고 있다. 이 실험 결과를 통해 $\theta$값이 커질수록

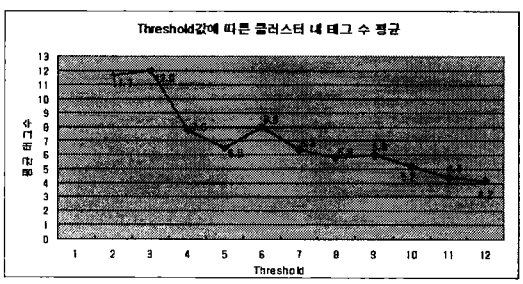


그림 4. $\theta$값에 따른 클러스터 내 태그 수 평균

클러스터 내에 태그 수는 점차 줄어드는 반면, 반대로 $\theta$값이 적어질수록 태그 수는 증가하는 것을 볼 수 있다. 이는 $\theta$값이 증가할수록 가중치가 높은 태그들이 클러스터에 포함되기 때문에 태그 수는 줄어들며, 반대로 $\theta$값이 적어질수록 태그 수가 증가하는 이유는 가중치가 낮은 모든 태그들이 클러스터에 포함되기 때문이다.

또한 그림 4의 결과를 기반으로 실험데이터에 적합한 $\theta$값 선정을 위해 전체 클러스터의 응집도 평균을 계산하였으며, 그 결과는 그림 5와 같다. 여기에서 응집도란 클러스터 내의 가중치로 연결된 태그들 간의 군집 정도를 의미하며, 본 논문에서 사용한 응집도 계산은 식 (1)과 같다.

식 (1)에서의 $A_i(j,k)$는 클러스터 $C(i)$ 내에 가중치로 연결된 tag $j$와 tag $k$을 의미하며, $n$은 $C(i)$에 속해 있는 tag 수를, $m$은 $\theta$값에 따라 생성된 클러스터 $C(i)$의 수를 의미한다.

$$\text{전체 클러스터 } C(i)\text{의 응집도 평균} = \frac{\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n}\sum_{k=1}^{n}A_i(j,k)}{n(n-1)}\right)}{m} \quad (1)$$

그림 5의 $\theta$값에 따른 전체 클러스터의 응집도 평균을 산출한 결과 $\theta$값이 커질수록 전체 클러스터의 응집도는 증가함을 알 수 있으며, $\theta$값이 적어질수록 응집도는 낮아지는 결과를 볼 수 있다. 이는 $\theta$값이 커질수록 가중치가 높은 태그들이 클러스터 내에 포함되기 때문에 응집도는 증가하는 것이며, 반대로 $\theta$값이 적어질수록 가중치가 낮은 태그들이 클러스터에 다수 포함되기 때문에 응집도는 낮아지는 결과이다. 또한 이 결과를 통해 $\theta$값이 커질수록 응집도는 증가하다가, $\theta$값 9 이후에는 응집도의 증가 추세가 일정하게 증가하는 결과를 도출하였다.

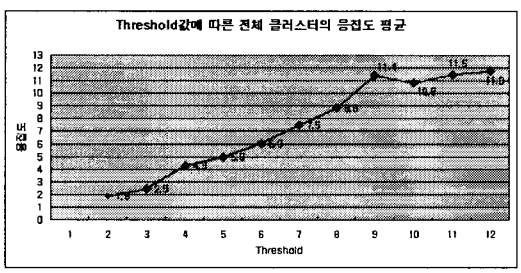


그림 5. $\theta$값에 따른 전체 클러스터의 응집도 평균

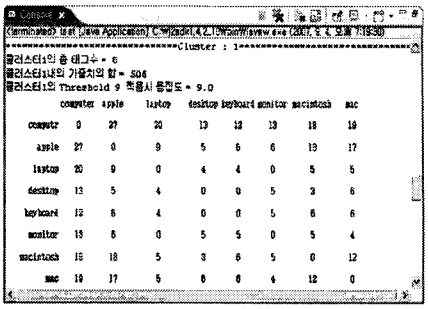

그림 6. 생성된 클러스터 1

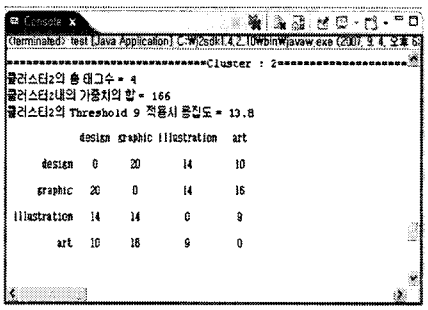

그림 7. 생성된 클러스터 2

위와 같은 $\theta$값 선정을 위한 실험 및 분석을 통해 본 연구에서 사용된 실험 데이터를 클러스터링하기 위한 $\theta$값은, 응집도의 변화가 일정해지는 시점인 $\theta$=9를 본 연구에서 제안하는 클러스터링 시스템에 적용하였다.

선정된 $\theta$=9를 적용한 결과 총 2개의 클러스터가 생성되었으며, computer와 관련된 8개의 태그로 구성된 클러스터 1[그림 6]과 design 및 graphic과 관련된 4개의 태그로 구성된 클러스터 2[그림 7]가 생성되었다.

## 4.4 연관 토픽쌍 및 이미지 추출 결과

실험 및 분석을 통해 4.3에서 생성된 클러스터를 검색시스템에 적용하기 위해 그림 3에서 제안한 태그 쌍 추출 알고리즘을 적용하였다. 추출결과 클러스터 1은 26쌍의 연관 태그 쌍과 클러스터 2는 6쌍의 연관 태그 쌍을 추출하였으며, 클러스터 1에 해당하는 26쌍의 연관 태그 쌍은 다음 그림 8과 같다.

그림 9, 10은 그림 8에서 추출된 태그 쌍을 기반으로 그림 3의 태그 쌍 기반 이미지 검색 알고리즘을 적용하여 클러스터 내의 각각의 토픽들에 이미지 리

| Relation Tag List | | |
|---|---|---|
| Tag1 | Tag2 | Weight |
| computer | apple | 27 |
| computer | desktop | 23 |
| computer | laptop | 20 |
| computer | mac | 19 |
| computer | macintosh | 18 |
| apple | desktop | 18 |
| apple | mac | 17 |
| computer | monitor | 13 |
| computer | keyboard | 12 |
| apple | laptop | 9 |
| apple | keyboard | 6 |
| apple | monitor | 6 |
| apple | desktop | 5 |

| Relation Tag List | | |
|---|---|---|
| Tag1 | Tag2 | Weight |
| mac | macintosh | 12 |
| mac | desktop | 6 |
| macintosh | keyboard | 6 |
| mac | keyboard | 6 |
| macintosh | laptop | 5 |
| mac | laptop | 5 |
| mac | monitor | 5 |
| macintosh | monitor | 4 |
| macintosh | desktop | 3 |
| desktop | monitor | 5 |
| desktop | laptop | 4 |
| monitor | keyboard | 5 |
| keyboard | laptop | 4 |

그림 8. 클러스터 1의 연관 태그쌍

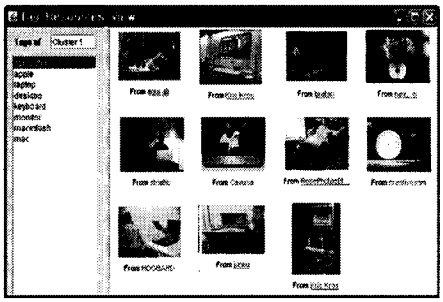

그림 9. 클러스터 1의 computer 태그 리소스

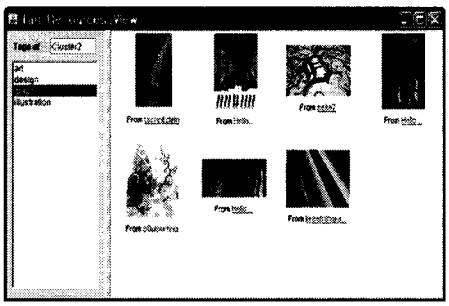

그림 10. 클러스터 2의 graphic 태그 리소스

소스를 부여한 결과를 보여주고 있으며, 본 논문의 실험 데이터인 키워드 computer를 통해 수집된 1~5page의 이미지들 중 1page에 해당하는 이미지들을 추출한 결과이다.

이 결과를 통해 기존 Flickr 사이트와 같은 태그 기반 사이트들은 단순히 키워드인 computer를 포함하는 태그들을 가진 이미지를 모두 보여줌에 따라서 사용자가 의도한 키워드와 다른 이미지들을 많이 내포한 결과를 보여준다. 그에 반해 본 논문에서 제안한 시스템에 의해 추출된 그림 9의 결과는 사용자가 요청한 키워드 computer와 관련성이 높은 이미지들을 추출 가능할 뿐만 아니라 생성된 클러스터 내의 각각의 태그들에 적합한 이미지 리소스가 부여된 결과를 도출하였다. 또한 Flickr 사이트의 부정확하게 태깅된 태그로 인한 잘못된 검색 결과 또한 연관도가 높은 이미지들로 분류시키는 향상된 결과를 그림 10과 같이 도출하였다.

## 5. 시스템 비교 평가

본 절에서는 태그 기반 시스템이 가지는 근본적인 두 가지 문제점 중 첫 번째 문제점을 해결하기 위한 방안으로 본 논문에서 제시한 태그 클러스터링을 통한 이미지 검색결과와 태그 기반 대표 사이트인 Flickr 검색결과와의 비교평가를 통해 제안 시스템의 향상된 검색결과를 제시한다.

비교평가 방법으로는 Flickr Open API를 이용하여 수집한 표 1에 해당하는 이미지들과 데이터를 제안한 시스템에 적용하여 생성된 표 2의 computer, apple과 관련된 클러스터 내의 각각의 태그들에 부여된 이미지들의 검색결과를 비교 평가하였다. 비교평가 기법으로는 정보검색 시스템의 성능측정 지표로 활용되고 있는 정확성(precision)과 재현율(recall) 평가기법을 적용하였다.

여기에서 정확한 이미지란, 추출된 이미지들 중 이미지 내에 키워드 computer, apple와 관련된 이미지를 포함하고 있으면, "정확", 그렇지 않으면, "부정확"으로 정의하였다.

그림 11의 Flickr 검색결과는 키워드 computer를 통해 수집한 1~5 page에 해당하는 120개의 이미지들의 정확성을 보여주고 있다. 이 중 정확한 이미지는 59개, 부정확 이미지는 61개로 평가됨에 따라서, 49%의 정확성 및 49% 재현율을 보여주고 있다. 이는 키워드인 computer를 포함하는 모든 이미지들을 출력하는 단순 태그 매칭에 따른 태그가 가지는 첫 번째 문제점인 부정확한 태그로 인한 낮은 검색결과의 원인이다.

그림 11의 제안 시스템 결과는 제안한 시스템에 의해 생성된 표 2의 computer 관련 클러스터에서

표 1. 평가 항목

| 키워드 | 상위 1~5Page | 태그 수 |
|---|---|---|
| computer | 120개의 이미지 | 836 |
| apple | 120개의 이미지 | 1017 |

표 2. 표 1의 평가항목을 제안 시스템에 적용한 결과

| 키워드 | threshold | 클러스터 수 | 생성된 클러스터 | 클러스터 내의 태그 |
|---|---|---|---|---|
| computer | 9 | 2 | computer 관련 | **computer**, apple, laptop, desktop, keyboard, mac, monitor, macintosh |
| | | | design 관련 | design, graphic, art, illustration |
| apple | 7 | 2 | apple (fruit 관점) | **apple**, fruit, green, red |
| | | | apple (company 관점) | **apple**, mac, macintosh, powerbook, ipod, imac |

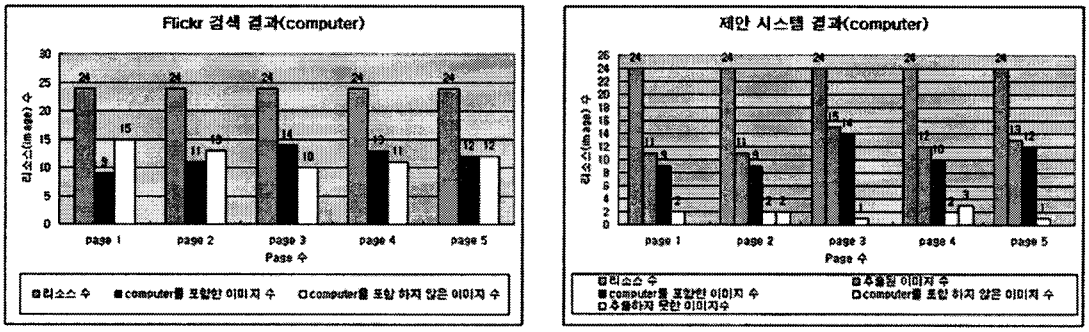


그림 11. 키워드 computer의 평가 결과

computer 태그에 부여된 이미지들의 정확성을 평가한 결과이다. computer 태그에 부여된 이미지는 62개였으며, 이 중 정확한 이미지는 54개, 부정확한 이미지는 8개로 평가됨에 따라, 87%의 정확성 및 45% 재현율을 확인할 수 있다. 부정확한 이미지 8개가 추출된 이유는 사용자가 잘못 태깅함으로서 추출된 경우로서, 제안한 토픽쌍을 이용한 리소스 추출에 의한 검색결과의 측면에서는 정확한 결과이다. 그러나 2 page의 2개, 4 page의 3개의 이미지들의 경우 태그 computer를 포함하고 있지만 추출하지 못한 경우이다. 이는 2 page 23번째 이미지에 태깅된 태그는 computer, window로서 본 논문에서 제안한 태그 쌍에 의해 추출된 태그들 중 computer 하나만 포함하고 있음에 따라서 추출되지 못한 경우로서, 이러한 제안 시스템의 문제점은 향후 해결해야하는 문제점이다.

그림 12, 13은 키워드 apple를 통해 수집한 데이터를 평가한 결과이다. Flickr에서 수집한 1~5 page의 이미지들의 경우 fruit 관점에서 apple를 의미하는 41개의 이미지와 apple 회사명 관점에서 관련된 64개의 이미지 및 전혀 상관없는 15개의 이미지들로 구성되어 있다. 이는 사용자 관점에서 fruit를 의미하는 apple을 검색결과로 원했다면, 그림 12의 평가 결과인 120개의 이미지 중 41개가 정확한 34%의 정확성 및 34%재현율을, apple 회사명 관점에서의 검색결과를 원했다면, 그림 13의 평가 결과인 120개의 이미지 중 64개가 53%의 정확성 및 53% 재현율을 검색결과라 할 수 있다.

그에 반해 제안한 시스템에 적용한 결과, 태그 클러스터링 과정을 통해 fruit 관점에서의 apple과 회사명 관점에서 apple 관련 태그들로 구성된 2개의 클러스터가 생성되었으며, 평가 결과, fruit 관점의 클러스터에서의 apple 태그에 부여된 이미지는 43개였으며, 이중 정확한 이미지는 38개, 부정확한 이미지는 5개, 추출하지 못한 이미지 3개로 평가됨에 따라, 88%의 정확성 및 32% 재현율을 도출하였다.

또한 생성된 company 관점에서의 apple 클러스터에서의 apple 태그에 부여된 이미지는 60개였으며, 이 중 정확한 이미지는 58개, 부정확한 이미지는 2개, 추출하지 못한 이미지 수 6개로 평가됨에 따라, 97%의 매우 높은 정확성 및 48% 재현율을 도출할 수 있었다. 부정확한 이미지 및 추출하지 못한 이미지는 computer 비교평가의 원인과 동일하다.

이 결과를 통해 태그기반 대표 사이트 중 하나인 Flickr 사이트의 정확성 및 재현율은 평균 45.3%로 평가되었다. 그에 반해 본 논문에서 제안한 시스템은

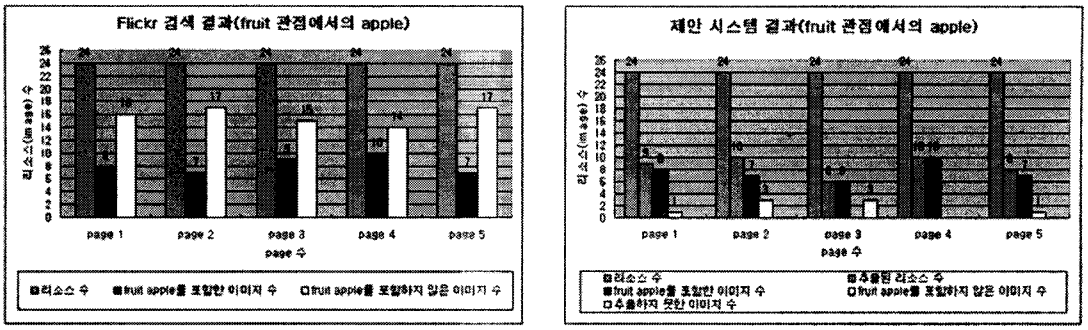


그림 12. fruit 관점에서의 apple 평가 결과

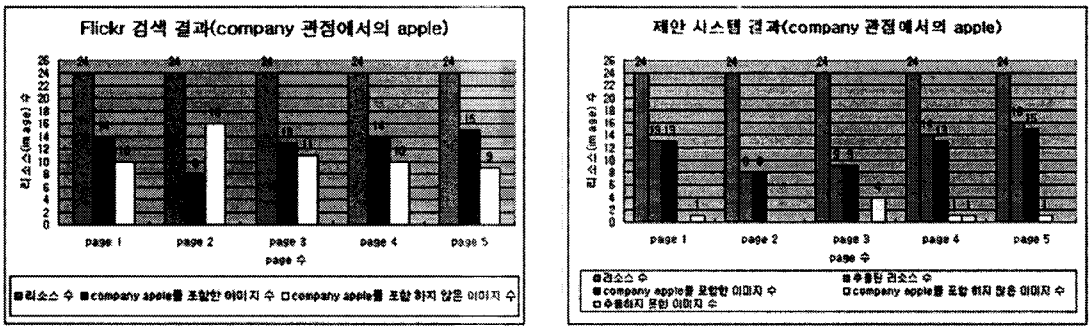


그림 13. company 관점에서의 apple 평가 결과

평균 90.7%의 정확성과 41.7%의 재현율로 평가됨에 따라, 재현율 면에서는 평균 3.6% 떨어지는 결과를 도출하였지만, 정확성 면에서는 평균 45.4%의 향상된 정확성을 도출하였다.

## 6. 결 론

본 논문에서는 태그 기반 시스템이 가지는 두 가지 문제점 중 첫 번째 문제점인 부정확한 태그로 인한 낮은 검색결과의 문제점을 해결하기 위해 태그 기반 검색 시스템을 제안하였다. 부정확한 태그를 제거하기 위해 본 연구에서는 연관 태그들 간의 맵핑을 통해 태그 가중치 매트릭스를 생성하고 이를 기반으로 높은 연관 태그로 구성된 클러스터로 생성하기 위한 알고리즘을 제안 하였다. 또한 향상된 검색을 위해 생성된 클러스터를 기반으로 검색 시스템에 적용하기 위한 태그 쌍 추출 및 검색 알고리즘을 제안하였다.

이와 같이 제안한 방법론들을 통해 태그들 간의 연관도가 매우 높은 클러스터를 생성하였으며, 클러스터 기반 검색 알고리즘에 적용한 결과 향상된 검색결과를 도출 할 수 있었다.

또한 제안한 시스템의 성능평가를 위해 태그 기반 대표 사이트인 Flickr와의 이미지 검색결과의 정확성과 재현율을 비교평가 하였으며, 그 결과 재현율 면에서는 평균 3.6% 떨어지는 결과를 도출하였지만, 정확성 면에서는 평균 45.4%의 매우 향상된 검색결과를 도출할 수 있었다.

향후 연구 과제로는 태그 기반시스템의 두 번째 문제점을 해결하기 위한 연구 및 성능평가에서 도출된 부정확한 이미지와 추출하지 못한 이미지를 해결하기 위한 연구를 진행할 예정이다.

## 참 고 문 헌

[1] 정부연, "2006년 인터넷 화두 웹2.0(Web2.0)," 기술동향, 2006.

[2] 홍성태, 임일, "웹2.0 환경에서 정보 분류와 필터링, 그리고 협업을 위한 기술의 동향 및 발전 방향," *Telecommunications Review*, 제17권,

제4호, 2007.
[3] Time O'Reilly, "What is Web2.0," http://www.oreilly.net.com/pub/a/oreilly/time/news/20-05/09/30/what-is-web-20.html, 2005.
[4] 이강표, 김두남, 김형주 "웹2.0 환경에서의 태깅 기술 동향," *한국정보과학회지*, 제25권, 제10호, 2007.
[5] 박영진, 송길영, 김경서, 송성환, "웹2.0관 정보 검색," *ITFIND 주간기술동향*, 제12권, 제5호, 2006.
[6] Ellyssa Kroski, "The Hive Mind Folksonomies and User-Based Tagging," InfoTangle, http://infotangle.blogsome.com/2006/12/07/thehive-mind-folksonomies-and-user-based-tagging/, 2006.
[7] Christopher H. Brooks and Nancy Montanez, "Improved Annotation of the Blogosphere Via Autotagging and Hierarchical Clustering," *International Conference on World Wide Web*, 2006.
[8] Nachiketa Sahoo and Jamie Callan, "Incremental Hierarchical Clustering of Text Documents," *CIKM 2006*, 2006.
[9] Grigory Begelman, Philipp Keller, and Frank Smadja, "Automated Tag Clustering Improving Search and Exploration in the Tag Space," *International Conference on World Wide Web*, 2006.
[10] S. White and P. Smyth, "A Spectral Clustering Approach to Finding Communities in Graphs," *In SIAM International Conference on Data Mining*, 2005.
[11] Neil C. Jones and Pavel A. Pevzner, "An Introduction to Bio Informatics Algorithms," pp. 340-385, 2000.
[12] 이시화, 무효려, 이만형, 황대훈, "web2.0 환경에서의 Tag Clustering 시스템 설계 및 구현," *한국멀티미디어학회 춘계 학술대회*, Vol.10, No.1, 2007.
[13] 이시화, 이만형, 황대훈, "Tag Clustering을 통한 Web 이미지 검색의 효율성 분석," *한국멀티미디어학회 추계 학술대회*, Vol.10, No.2, 2007.
[14] Flickr, http://www.flickr.com/

**이 시 화**

2005년 서울보건대학 컴퓨터정보과 졸업
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과 석사과정 졸업
2008년 현재 경원대학교 전자계산학과 박사과정
관심분야 : e-Learning, Context- Aware, Semantic Web, Web2.0

**이 만 형**

1997년 건양대학교 정보통신학과(학사)
1999년 경원대학교 전자계산학과(석사)
2007년 경원대학교 전자계산학과(박사수료)
1999년~현재 현대전문학교 교수
관심분야 : e-러닝, Semantic Web, 보안

**황 대 훈**

1997년 동국대학교 수학과(학사)
1983년 중앙대학교 전자계산학과(석사)
1991년 중앙대학교 전자계산학과(박사)
1983년~1985년 한국산업경제기술연구원(KIET) 연구원
1987년~현재 경원대학교 교수
2004년~2006년 한국멀티미디어학회 부회장
관심분야 : e-러닝, Semantic Web, 유비쿼터스 컴퓨팅