Journal of Korea Multimedia Society Vol. 16, No. 11, November 2013(pp. 1305-1314)
http://dx.doi.org/10.9717/kmms.2013.16.11.1305

태그쌍의 의미유사도 기반 태그 랭킹 시스템

이시화†, 황대훈††

요 약

기존의 태그 기반 시스템들은 콘텐츠에 태깅된 태그들을 활용한 단일 태그 매칭을 통해 검색결과를 제공함에 따라 정확도가 낮은 검색결과를 제공하고 있으며, 또한 사용자들이 콘텐츠에 태깅 시 태그간의 연관관계 및 우선순위는 고려하지 않아 태그가 가지고 있는 콘텐츠와 관련된 정보들을 효율적으로 제공하지 못하고 있다. 이에 본 논문에서는 위의 문제점을 해결하기 위해 태그 기반 시스템에 적합한 태그간 의미 유사도를 추출하여 콘텐츠에 태깅된 태그들을 재 랭킹하기 위한 태그 랭킹 시스템을 제안하였다. 제안 시스템의 성능 평가는 이미지에 태깅된 태그(baseline)와 태그 동시출현 빈도수 기법을 적용한 랭킹(frequency) 결과를 본 논문에서 제안한 태그 랭킹 시스템에 의해 추출된 랭킹 결과와 비교 실험하였다.

Tag Ranking System based on Semantic Similarity of Tag-pair

Si-Hwa Lee†, Dae-Hoon Hwang††

ABSTRACT

The existing tag based system deducts a retrieval result with low accuracy through the usage of a single tag matching by using tags tagged in contents. And the system doesn't provide effectively contents related information which the tags have, as the users place tags on contents without considering the priority and associative relation between tags. For a solve of above problems, this paper suggests a tag ranking system which extracts semantic similarity between tags and re-ranks the tags tagged in contents. In order to evaluate the performance of suggested system, this paper experiments and compares the ranking result of this paper's tag ranking system with the result of baseline method using tags tagged in images and frequency method adapting tag co-appearance frequency.

Key words: Tag(태그), Tag Similarity(태그 유사도), WordNet(워드넷), Tag Ranking(태그 랭킹)

1. 서 론

태깅은 현재 많은 인터넷 사용자들로부터 큰 호응을 얻고 있으며, 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 적용되고 있다[1]. 그러나 태그 기반 검색 시스템[2-4]들은 콘텐츠에 태깅된 단일 태그 매칭을 통해 검색결과를 제공함에 따라 정확도가 낮은 검색결과를 제공하고 있다. 또한 사용자는 콘텐츠에 태깅 시 태그간의 연관관계 및 우선순위는 고려하지 않아 사용자 주관적인 태그들이 상당수 포함되며, 이로 인해 태그가 가지고 있는 콘텐츠와 관련된 정보들을 효율적으로 사용자들에게 제공하지 못한다.

이에 본 논문에서는 위의 문제점을 해결하기 위하여 태그 유사도를 이용한 태그 랭킹 시스템을 제안하였다. 시스템은 크게 기존 태그쌍 빈도수 추출 모듈,

※ 교신저자(Corresponding Author) : 황대훈, 주소 : 경기도 성남시 수정구 복정동 산 65번지 가천대학교 새롬관 5-14호(461-200), 전화 : 010) 5458-8111, FAX : 031) 757-6715, E-mail : hwangdh@gachon.ac.kr
접수일 : 2013년 9월 10일, 수정일 : 2013년 9월 27일
완료일 : 2013년 10월 8일
† 메디오피아테크(주), leesihwaman@gmail.com
†† 가천대학교 IT대학
※ "이 논문은 2013년도 가천대학교 교내연구비 지원에 의한 결과임."(GCU-2013-R301)

WordNet을 이용한 태그쌍 의미 유사도 추출 모듈과 태그간 유사도 가중치 값을 기반으로 콘텐츠에 태깅된 태그들을 재 랭킹하여 효율적인 정보 내비게이션 및 검색에 활용하기 위한 모듈로 구성되어있다.

제안한 시스템의 성능평가를 위해 이미지 공유 사이트인 Flickr[2]의 Open API를 이용하여 키워드 'lion', 'tomato' 태그를 포함하고 있는 상위 인기 이미지 각 500개와 그에 태깅된 태그 정보들을 수집하였으며, 기존 연관 태그 추출 기법들과의 랭킹 결과를 실험 분석하였다.

2. 관련 연구

2.1 태그 기반 시스템의 문제

태그는 매우 유연하고 역동적인 분류체계를 제공한다. 하지만 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계 또한 가지고 있는 있는데, 태그 기반 검색 시스템은 검색에 있어서 정확도가 떨어진다는 문제점이 있다. 태그는 어떤 정보를 넓은 범주의 카테고리에 위치시키는 데에는 매우 유용하지만, 사용자가 원하는 정확한 정보를 찾아내는 데에는 비효율적인 결과를 나타내고 있다. 이와 같은 원인으로는 부정확한 태그로 인한 낮은 검색결과와 비정렬된 태그로 인한 비효율적인 정보 내비게이션을 들 수 있다.

기존 태그 기반 검색시스템[2-4]들은 다음과 같은 문제점을 가지고 있다.

첫째, 부정확한 태그로 인하여 낮은 정확도를 가지는 검색결과의 문제로써, 현재 태그 기반 검색 시스템들은 웹 콘텐츠에 태깅된 태그를 기반으로 단일 태그 매칭을 통해 검색결과를 제공한다. 이 경우 콘텐츠에 부정확한 태그가 태깅되어 있을 경우 검색결과 또한 부정확하다는 문제점을 가진다.

둘째, 비정렬된 태그로 인한 비효율적인 정보 내비게이션의 문제로써, 사용자는 콘텐츠에 태깅 시 태그간의 연관관계 및 우선순위는 고려하지 않는다. 이 경우 태깅된 태그는 콘텐츠와 연관된 태그 외에도 사용자 주관적인 태그들이 상당수 포함되며, 이로 인해 태그가 가지고 있는 콘텐츠와 관련된 정보들을 효율적으로 사용자들에게 제공하지 못한다.

2.1 태그 기반 연구의 문제

기존 태그 기반 검색 시스템이 가지는 문제점을 해결하기 위해 협업 태깅[5], 태그 클러스터링[6,7], 태그의 계층구조 생성[8], 태그 기반 검색[9,10], 태그 기반 추천[11] 등의 많은 연구들이 진행되었다.

Jin[5]은 태그 분포에 따른 통계적 규칙 및 패턴을 추출하여 데이터 마이닝을 통한 협업적 태깅 시스템에 대한 연구를 진행하였으며, Leginus[6]과 Lu[7]는 태그의 동시출현 빈도수를 관계의 척도로 사용하여 태그 클러스터링에 관한 연구를 진행하였다. Li[8]는 태그 동시출현 빈도수를 기반으로 태그 계층구조를 생성하기 위한 계층 프레임워크 연구를 진행하였다.

이만형[9]은 태그의 동시출현 빈도수를 기반으로 연관도가 높은 태그들을 클러스터링하고, 클러스터 내의 태그들을 기반으로 다중태그 검색에 관한 연구를 진행하였으며, 임영섭[10]은 사용자들의 참여 정도를 고려하여 검색성능을 개선하기 위해 소셜 태깅 서비스인 del.icio.us의 사용자들을 활동성 관점에서 관찰하고 모델링화하여 검색에 활용함으로써 검색을 개선하는 방법을 제안하였다.

김현우[11]은 추천 시스템에서 새롭게 등장한 사용자에게 정확한 추천을 제공하지 못하는 Cold-start 문제를 해결하기 위해 사용자의 태그 셋으로 사용자 프로파일을 구성하고 자연언어처리에 사용되는 n-gram 모델을 이용하여 태그 셋을 확장하였으며 아이템의 인기도와 시간 정보를 추가적으로 활용하였다

그러나 기존 태그 기반 연구[5-11]들은 태깅된 태그간 유사도 산출을 위해 태그 동시출현 빈도수만을 이용한다.

사용자들이 태깅한 태그에는 사용자의 직관적 판단에 의한 주관적 성향의 태그들이 상당수 존재하며, 이중 일부 태그들은 높은 빈도수를 가진다. 이로 인해 높은 빈도수의 태그를 추출하여 활용하는 기존 연구들에서는 사용자 주관에 의한 태그들이 태그간 연관관계가 높은 태그로 정의되며, 또한 태그 동시출현 빈도수는 낮지만 태그간 의미관계가 높은 태그들에 대해서는 낮은 빈도수로 인해 태그간 연관관계가 없는 태그로 정의되는 문제점을 나타내고 있다.

3. 태그 랭킹 시스템

그림 1은 본 논문에서 제안한 태그 유사도를 이용한 태그 랭킹 시스템의 구성도를 나타내었으며, 크게

4개의 모듈로 구성되어 있다.

태그 빈도수 추출 모듈(Tag Frequency Extraction Module)은 웹상에 산제되어 있는 콘텐츠에 태깅된 태그들을 Open API를 활용하여 수집하고 태그들 간의 연관성에 따라 맵핑하여 연관 태그쌍(tag-pair)을 추출하는 역할을 수행한다.

태그 유사도 추출 모듈(Tag Similarity Extraction Module)은 태그 빈도수 추출 모듈에서 추출된 연관 태그쌍을 기반으로 태그쌍 빈도수 추출(Tag-pair Frequency Extraction)과 워드넷(Wordnet)을 이용한 태그쌍 의미 유사도 추출(Tag-pair Semantic Similarity Extraction)을위한 역할을 수행한다.

태그쌍 가중치 행렬 생성 모듈(Tag-pair Weight Matrix Creation Module)은 위의 두가지 방법에 의해 추출된 태그쌍 의미 유사도를 기반으로 가중치 행렬을 구성하여, 본 논문에서 제안하는 최종 태그 랭킹을 산출하기 위한 준비 역할을 수행한다.

마지막으로 태그 랭킹 모듈(Tag Ranking Module)은 추출된 태그쌍 가중치 행렬(TWM: Tag-pair Weight Matrix)의 태그간 유사도를 기반으로 각 이미지들에 태깅된 태그들을 랭킹하는 역할을 수행한다.

이중 태그 빈도수 추출 모듈과 태그 유사도 추출 모듈의 태그쌍 빈도수 추출은 선행 연구 [12,13]에서 진행하였으며, 본 논문에서는 WordeNet을 이용한 태그쌍 의미 유사도 추출을 통한 TWM의 생성 및 태그 랭킹을 위한 연구를 중심으로 진행하였다.

3.1 태그쌍 의미 유사도 추출(TSSE)

본 논문에서 제안한 WordNet을 이용한 태그쌍 의미 유사도 추출(TSSE : Tag-pair Semantic Similarity Extraction) 알고리즘은 기존의 태그 기반 연구들의 문제점을 극복하고자 한 것으로서, 정보량 기반의 Resnik[14] 방식과 Lin[15] 방식은 단어간 공통성뿐만 아니라 차별성도 포함시킴으로써 단어 간의 유사도를 측정하려 하였으나, 단어의 모호성에 의해 특정 단어들 간에는 상반된 유사도 결과가 만들어지는 문제점을 가지고 있다.

이에 본 논문에서는 Resnik 방식과 Lin 방식을 고려한 그림 2의 TSSE 알고리즘을 제안한다.

알고리즘의 진행 과정은 태그쌍을 개념적으로 포함하는 concept c에 속하는 모든 단어들의 집합인 $word(c)$를 찾는 것으로부터 시작한다. $word(c)$는 concept c에 개념적으로 속하는 단어들의 집합으로

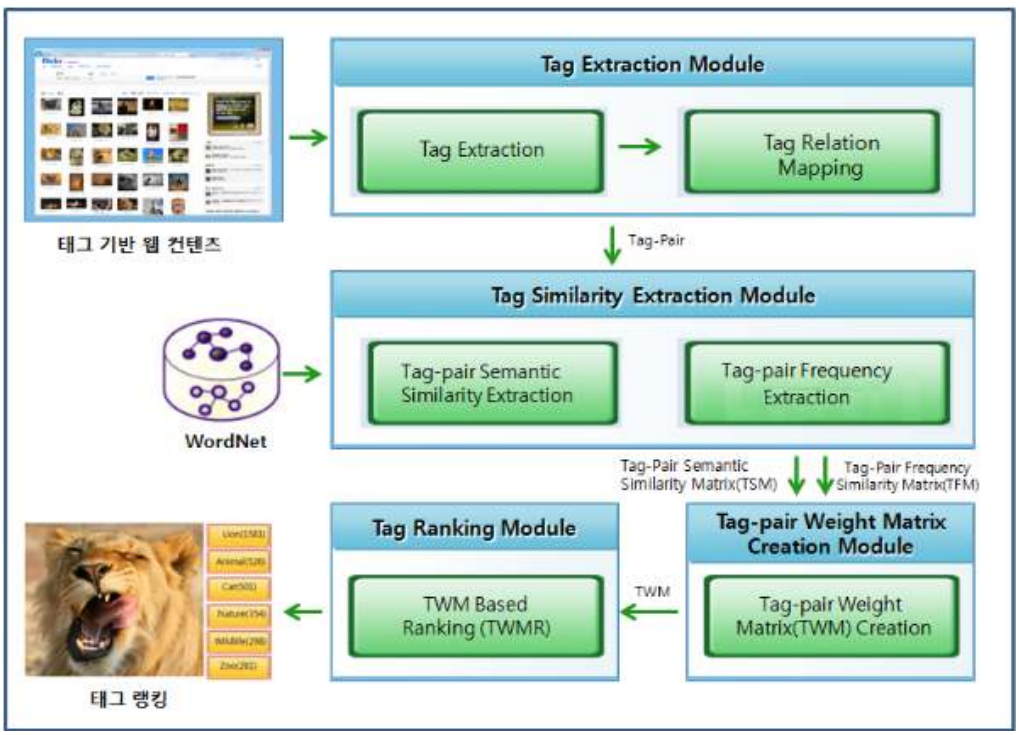

그림 1. 태그 랭킹 시스템

// c : 특정 태그쌍의 개념적으로 포함하는 concept
// $word(c)$: concept c의 information content file에 있는 모든 synset
// $count(n)$: $word(c)$ 에 속하는 각 synset 요소들의 빈도수
// $freg(c)$: 모든 synset 요소들의 빈도수의 합
// $\Pr(c)$: Concept Probability
// N : $word(c)$의 단어들 중 명사의 총 개수
// IC : Information Content
// $S(c_1,c_2)$: concept c_1과 c_2를 모두 개념적으로 포함하는 concept의 집합
// $Weight$: 가중치($0 < Weight < 1,\ Weight_1 + Weight_2 = 1$)
// TS : Tag Similaity 즉, Semantic Similarity between Tags

Find $word(c)$ by Concept c

for $(i = 1 \text{ to } |word(c)|)$ {

$freg(c) = freg(c) + count(i)$

}

$\Pr(c) = freg(c)/N$

$IC(c) = 1/\log Pr(c)$

$Sim_1(c_1,c_2) = \max_{c \in S(c_1,c_2)} IC(c)$

Compute $IC(c_1)$ and $IC(c_2)$

$Sim_2(c1,c2) = 2 * IC(S(c_1,c_2))/IC(c_1) + IC(c_2)$

$TS = \{ Sim_{1X} Weight_1 + Sim_{2X} Weight_2 \}/2$

그림 2. TSSE 알고리즘

써, WordNet 라이브러리[16]의 information content file에 있는 모든 synset을 의미한다.

그 후, $word(c)$에 속하는 각 synset들의 빈도수 $count(n)$을 합하여 모든 synset 요소들의 빈도수의 합 $freg(c)$을 추출한다. 이는 $word(c)$에 개념적으로 속하는 단어들의 수만큼 반복 진행된다.

다음으로 모든 synset 요소들의 빈도수의 합 $freg(c)$에 $word(c)$의 단어들 중 명사의 총 개수로 나누어 c의 개념 확률값인 $\Pr(c)$을 구한다. 그 후 추출된 $\Pr(c)$에 log 값을 취해 IC을 계산하는데, concept이 보다 포괄적(추상적)이 되면 IC 값은 보다 작아진다. 따라서 WordNet의 top concept의 IC는 0이며 IC가 1이면 같은 의미의 단어이다. 두 태그쌍과 가장 공통으로 연결된 LCS를 구하고 이중에서 가장 큰 값을 선택하여 두 태그간의 초기 유사도 값을 구한다.

이와 같은 과정을 반복하여 concept c_1과 c_2에 대한 $IC(c_1)$와 $IC(c_2)$를 구하고 이를 기반으로 2차 유사도 값을 산출한다. 그 후에는 마지막으로 이 두 유사도 값에 각 유사도의 반영 정도에 따르는 가중치($0 < Weight < 1,\ Weight_1 + Weight_2 = 1$)를 각각 곱하여 합한 결과를 2로 나누어 태그간 의미 유사도 값 TS를 산출한다.

3.2 태그쌍 가중치 행렬(TWM) 생성

기존 태그 유사도 추출 방법인 태그 동시출현 빈도수는 사용자들이 태깅한 태그를 기반으로 특정 키워드와 연관된 태그들을 쉽게 추출할 수 있다는 장점을 가지고 있으나, 단점으로는 사용자 주관적 태그가 높은 빈도수를 차지하며, 태그간 의미는 높지만 빈도수가 낮은 태그는 활용되지 못한다는 문제점을 가지고 있다. 그에 반해 WordNet을 이용하여 태그들 간의 사전적 관계를 기반으로 하는 TSSE 알고리즘에 의해 추출된 TWM의 유사도 값은 기존의 태그 빈도수 방식의 단점들을 해결하는 결과를 얻을 수 있다. 그러나 단점으로는 WordNet을 이용하기 위해서는 유사도를 비교하기 위한 키워드들을 알아야 된다는 문제점을 가지고 있다.

이에 본 논문에서는 두 태그간 유사도 문제점을

해결하기 위해 식 1을 적용하여 최종 태그간 유사도 척도인 태그쌍 가중치 행렬(TWM : Tag-pair Weight Matrix)을 생성하였다.

$$TWM(i,j) = TSM(i,j) \times TFM(i,j) \quad (1)$$

식 1에서 $TSM(i,j)$는 태그쌍 의미 유사도 행렬의 태그 i행 j열의 의미 유사도 값이며, $TFM(i,j)$는 태그쌍 동시출현 빈도수 행렬의 태그 i행 j열의 빈도수이다.

이와 같이 두 유사도 값을 제안하는 식 1을 적용하게 되면 태그 동시출현 빈도수의 문제점인 사용자 주관의 태그 및 빈도수가 낮지만 높은 의미를 가지는 태그의 문제점을 해결 가능하며, 또한 WordNet의 문제점인 유사도 비교를 위한 태그쌍 추출의 문제를 해결가능하다.

3.3 TWM 기반 랭킹 알고리즘

본 절에서는 기존 태그 기반 시스템이 가지는 두 번째 문제점인 비정렬된 태그로 인한 비효율적인 정보 내비게이션의 문제를 해결하기 위하여, 기존 콘텐츠에 태깅된 태그들을 3.1절에서 제안한 TWM을 기반으로 재랭킹하여 효율적인 정보 내비게이션을 위한 TWM 기반 랭킹(TWMR : TWM based Ranking) 알고리즘을 제안하였다.

제안하는 TWMR 알고리즘의 진행과정을 살펴보면 먼저, TWM에 포함된 태그의 개수만큼 반복 진행하여 TWM의 원소를 행 단위로 더하여 태그 i의 가중치의 합 $STWM(i)$을 구한다. 그 후 TWM을 참조하여 각 이미지 별로 태깅된 태그의 가중치 값을 구하기 위해 먼저 t번째 이미지 I_t의 $q \times 2$ 행렬 TM_t을 초기화

```
// p : TWM의 태그의 개수
// STWM(i) : TWM의 i번째 행의 원소들의 합, 즉 tag i와 연관된 태그들 간의 가중치의 합
// t : 이미지 갯수
// I_t : t번째 이미지
// q : t번째 이미지의 태그의 개수
// TM_t : t번째 이미지의 q×2 태그 행렬

// STWM의 원소를 행단위로 더하여, tag i의 가중치의 합을 구한다
for(i=1 to p) { // TWM의 태그의 개수만큼 반복
        STWM(i) = 0
        for(j=1 to p) {
            STWM(i) = STWM(i) + TWM(i,j)
        }
}
// STWM을 참조하여 각 이미지 별로 태그 랭킹을 구한다
for(i=1 to t) { // 이미지 개수만큼 반복
        for(j=1 to q) { // t번째 이미지의 태그의 개수만큼 반복
                TM_t(j,2) = 0  // TM_t을 초기화
                for(k=1 to p) { // TWM의 태그의 개수 만큰 반복
                        if(TM_t(j,1) == TWM's k-th tag) {
                                TM_t(j,2) = STWM(k)
                                break;
                        }
                }
        }
        // TM_t을 2번째 열, 즉 가중치를 기준으로 내림차순 정렬
        Descending sort TM_t by 2nd column
}
```

그림 3. TWMR 알고리즘

표 1 실험 평가 항목

키워드	이미지 수	전체 태그 수
lion	500개	12,588개
tomato	500개	11,088개

한다.

그 후 TWM의 태그의 개수 p만큼 반복하며 TM_t의 j행 1열의 태그가 TWM의 k번째 태그와 같은지 비교한다. 같으면, TM_t의 j행 2열에 TWM의 i번째 행의 원소들의 합 즉, tag i와 연관된 태그들 간의 가중치의 합 $STWM(i)$을 추가한다. 이러한 과정은 t번째 이미지의 태그의 개수 q만큼 반복되어지며, 그 후 t번째 이미지의 태그의 $q\times2$ 태그 행렬 TM_t의 2번째 열, 즉 가중치를 기준으로 내림차순 정렬한다. 이러한 과정은 모든 이미지의 개수 t만큼 반복 진행된다.

4. 실험 및 고찰

4.1 실험 데이터

본 논문에서 제안한 알고리즘을 평가하기 위하여 이미지 공유 사이트인 Flickr의 Open API[2]를 이용하여 키워드 'lion', 'tomato' 태그를 포함하고 있는 상위 인기 이미지 500개와 그에 태깅된 태그 정보들을 수집하였으며, 실험 평가 항목으로 수집한 실험 데이터는 표 1과 같다.

4.2 태그쌍 의미 유사도 추출(TSSE)

태그 동시출현 빈도수 기법의 높은 빈도수를 가지는 사용자의 주관에 의한 태그 및 의미적으로 연관도가 높지만 낮은 빈도수를 가지는 태그들의 문제를 해결하기 위해 제안한 TSSE 알고리즘을 적용하여 태그쌍들의 의미 유사도를 측정하였다.

그림 4(a)는 'lion' 데이터의 24,662의 태그쌍 중 유사도 기준 상위 20개의 태그쌍에 대한 의미 유사도와 태그 동시출현 빈도수의 유사도 값을 나타내고 있다. 사용자의 주관적 판단에 의한 태그인 'lion-specanimal', 'lion-animalkingdomelite', 'lion-pantheraler' 등의 태그 쌍은 본 논문에서 제안한 TSSE 알고리즘에 의해 추출된 의미 유사도에서는 연관관계가 없는 0의 값으로 산출되는 결과가 나타났다. 그러나 태그 동시출현 빈도수에서는 낮은 빈도수를 가지지만 태그간 의미관계가 높은 'lion-cat'의 유사도는 92로 가장 높은 유사도로 산출되는 결과로 나타났다.

'tomato' 데이터 또한 'tomato-vegetable', 'tomato-tomatoes' 등과 같은 태그들은 태그 동시출현 빈도수는 낮지만 의미 유사도가 높은 태그들로써 기존 연구들이 공통적으로 사용한 태그 동시출현 빈도수의 문제점을 나타내고 있다.

이와 같이 태그 동시출현 빈도수와 TSSE 알고리즘을 이용한 태그간 유사도 산출의 실험 및 평가를 통해, 본 논문에서 제안한 TSSE 알고리즘은 기존의 태그 동시출현 빈도수가 가지는 단점인 사용자 주관적 태그들을 제거할 수 있음을 알 수 있다. 또한 태그간 의미 연관관계는 높지만 낮은 빈도수를 가지는 태그들에 대해서는 높은 유사도 값이 산출되는 결과가 나타났다.

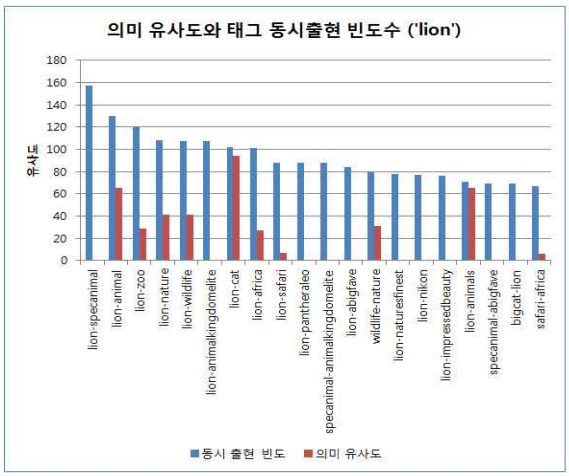

그림 4(a)

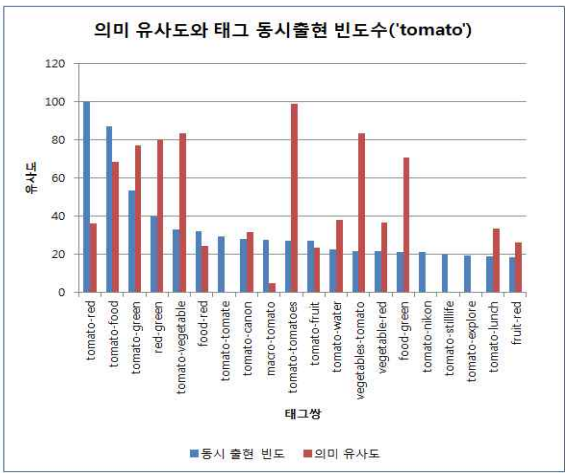

그림 4(b)

이러한 실험 결과를 통해 태그쌍 동시출현 빈도수와 제안한 TSSE 알고리즘은 단일 방법론으로 사용하기 힘든 반면, 두 방법론을 혼합하여 사용하면 서로의 단점들을 상호 보완할 수 있음을 알 수 있다. 이에 본 논문에서는 두 방법론에 의해 얻어진 유사도 값을 조합하여 새로운 태그간 유사도 척도인 가중치를 산출하였다.

4.3 태그쌍 가중치 행렬(TWM) 생성

태그쌍 동시출현 빈도수는 사용자들이 태깅한 태그를 기반으로 특정 키워드와 연관된 태그들을 쉽게 추출할 수 있다는 장점을 가지고 있다. 그러나 단점으로는 사용자의 주관적 태그가 높은 빈도수를 차지하며, 또한 태그간 의미는 높지만 빈도수가 낮은 태그는 활용되지 못하는 문제를 가지고 있다. 제안한 TSSE 알고리즘은 태그쌍 동시출현 빈도수가 가지는 두 가지 문제점을 해결할 수 있는 장점이 있으나, 의미 유사도를 비교하기 위한 연관 태그가 필요하다는 단점을 가지고 있다.

위의 문제를 해결하기 위해 3.1절에서 제안한 TWM 생성 방법론에 실험 데이터를 적용한 결과 데이터 'lion'은 3,346개, 'monitor'는 5,815개의 태그간 가중치로 구성된 TWM이 생성되었다.

그림 5는 실험 및 검증을 통해 최종 산출된 'lion' 데이터의 생성된 TWM의 일부를 나타낸 것이다.

	lion	cat	animal	animals	nature	wildlife	zoo	lions	leo	carnivore	bravo	africa	mammal	feline	predator	cub
lion	0	100	88	48	47	46	36	35	34	30	29	29	28	28	23	22
cat	100	0	25	12	10	9	18	9	10	0	9	4	9	23	6	11
animal	88	25	0	19	15	18	20	10	8	6	12	0	0	0	0	0
animals	48	12	19	0	12	17	7	10	2	10	4	8	2	1	7	4
nature	47	10	15	12	0	0	6	5	11	10	9	18	4	1	10	3
wildlife	46	9	18	17	0	0	3	6	9	14	8	19	6	1	15	3
zoo	36	18	20	7	6	3	0	2	8	0	6	1	0	4	1	6
lions	35	9	10	10	5	6	2	0	0	0	2	3	0	0	1	2
leo	34	10	8	2	11	9	8	0	0	8	4	6	2	6	8	4
carnivore	30	0	6	10	10	14	0	0	8	0	2	9	6	0	16	2
bravo	29	9	12	4	9	8	6	2	4	2	0	4	4	0	6	3
africa	29	4	0	8	18	19	1	3	6	9	4	0	2	0	8	3
mammal	28	9	0	2	4	6	0	0	2	6	4	2	0	2	5	3
feline	28	23	0	1	1	1	4	0	6	0	0	0	2	0	0	4
predator	23	6	0	7	10	15	1	1	8	16	6	8	5	0	0	1
cub	22	11	0	4	3	3	6	2	4	2	3	3	3	4	1	0
canon	21	6	0	7	8	5	0	3	1	0	6	6	1	0	3	0
male	17	12	0	2	2	2	4	0	6	0	1	2	0	7	2	3
king	15	5	6	2	2	3	2	1	1	0	2	1	1	0	2	0
tiger	14	7	0	3	1	1	2	0	0	0	0	0	0	2	0	0

그림 5 'lion' 데이터의 생성된 TWM 일부

4.4 TWM 기반 랭킹

본 절에서는 기존 태그 기반 시스템이 가지는 문제점을 해결하기 위해 제안한 TWM 기반의 태그 랭킹 결과를 기존 기법들과 비교 평가한다.

비교 실험 대상은 기존 이미지에 태깅된 태그(Baseline)와 태그 동시출현 빈도수를 이용한 태그 랭킹(Frequency) 결과를 본 논문에서 제안한 TWMR 결과와 비교 분석하였다.

표 2는 'lion' 데이터를 TWMR에 적용한 결과 중 일부를 나타낸 것이다. Baseline은 사용자들이 태그간 상관관계 및 태그간 중요성을 고려하지 않고 이미지들에 태깅한 결과로써, 기존 태그 기반 시스템이 가지는 두 번째 문제점인 비정렬된 태그의 결과를 나타내고 있다. 5개의 이미지의 상위에 태깅된 태그들은 이미지와 태그가 어떠한 연관관계를 가지는지, 혹은 사용자가 태깅한 태그들의 목적을 파악하기 어렵다.

첫 번째 이미지의 경우 태깅된 'pencilvscamera' 태그는 'Pencil', 'VS', 'Camera' 세 단어를 결합한 "그림과 사진의 대결"을 의미하는 단어로써, 다른 사용자들에게는 모호한 태그이며, 또한 그 외에 4번째 이미지를 제외한 모든 이미지들의 상위에 태깅된 태그들이 비정렬된 태그의 문제점 및 사용자 주관적 태그들로 구성된 결과임을 확인할 수 있다.

Frequency 랭킹 결과는 태그쌍 동시출현 빈도수를 기반으로 생성된 TFM의 태그간 빈도수를 이용하여 이미지에 태깅된 태그들을 랭킹한 결과로써, 첫 번째 이미지의 태그 랭킹결과 'lion'과 연관된 태그 'animal', 'wild'로 랭킹되었다. 다음으로 이미지의 다른 주제를 설명하는 태그들인 'art', 'photography'태그 순으로 태그 랭킹되는 좋은 결과가 나타났다.

그러나 나머지 이미지들의 태그 랭킹 결과에서는 사용자 주관의 태그들이 상위에 랭킹되는 부정확한 태그 랭킹 결과가 나타났다. 2번째, 3번째 이미지는 'hdr' 태그가 4번째 이미지는 'specanimal' 태그가 각각 이에 해당한다. 특히 5번째 이미지의 경우는 상위 태그 랭킹된 태그 중 'lion' 태그를 제외한 'specanimal', 'impressedbeauty', 'flickrbigcats', 'vosplusbellesphotos' 등의 태그들이 모두 사용자의 주관적인 태그들로 랭킹된 부정확한 결과가 나타났다.

부정확한 태그 랭킹 결과의 원인은 높은 태그 동시출현 빈도수를 가지는 태그들만을 이용할 경우 특정 주제들에서는 사용자의 주관적 태그들이 상위 빈도수를 가지며, 이로 인해 5번째 이미지와 같은 사용자의 주관적인 태그들이 상위 결과에 반영된다.

TWMR 방법은 TWM에 정의된 태그들 간의 가

표 2 'lion' 이미지의 태그 랭킹 결과 중 상위 5개의 이미지

Image Photo ID	Rank	Baseline	Frequency	The Proposed Method TWMR
 6967423963	1	pencilvscamera	lion	lion
	2	betaversion	animal	animal
	3	art	wild	wild
	4	photography	art	art
	5	painting	photography	light
 5142525802	1	anto	lion	lion
	2	xiii	nikon	blue
	3	antoxiii	hdr	sky
	4	france	sky	bridge
	5	français	macro	france
 5122477577	1	anto	lion	lion
	2	xiii	nikon	king
	3	antoxiii	hdr	sky
	4	france	sky	paris
	5	français	king	france
 4058859734	1	lion	lion	lion
	2	theunforgettablepictures	specanimal	animal
	3	saariysqualitypictures	impressedbeauty	wildlife
	4	impressedbeauty	flickrbigcats	-
	5	animal	vosplusbellesphotos	-

중치 값을 기반으로 랭킹에 적용한 결과로서, 기존 Baseline과 Frequency의 방식에 비해 매우 효율적인 태그 랭킹 결과가 나타났다. 실험에 사용된 이미지들은 여러 주제들 혹은 하나의 주제를 표현한 데이터들로써 이미지와 연관관계가 높은 태그들은 상위에 랭킹되었으며, 연관관계가 작은 태그들은 하위에 랭킹되는 좋은 결과가 나타났다.

표 3은 'tomato' 데이터의 태그 랭킹 결과를 나태낸 것이며, 기존 태깅된 Baseline의 태그들의 경우는 'lion' 데이터의 분석 결과와 동일한 기존 태그 기반 시스템의 두 가지 문제점을 포함하는 결과를 나타내었다.

Frequency 랭킹 결과는 2번째 이미지를 제외한 나머지 4개의 이미지들에서 사용자의 주관적인 태그들 및 사용자들의 태깅 의도를 알 수 없는 태그들이 상위에 랭킹되는 부정확한 랭킹 결과가 나타났는데, 원인은 'lion' 데이터의 분석결과와 동일하다.

TWMR 결과는 Frequency 랭킹 결과에 비해 좋은 결과가 나타났는데, 5번째 이미지의 경우는 TWMR 방법의 문제점인 낮은 가중치를 가지는 'canon', 'brown', 'girl', 'box' 등의 태그들도 랭킹 결과에 포함되는 등 'lion' 데이터와 동일한 문제점이 나타났다. 그러나 이미지와 태그간 연관관계를 고려한다면 해당 이미지의 주제에 따른 연관태그들이 효율적으로 랭킹된 결과라고 할 수 있다.

5. 결 론

본 논문에서는 기존의 태그 기반 시스템 및 연구들의 문제점을 해결하기 위해 태그 유사도를 이용한 클러스터 기반 태그 랭킹 시스템을 제안하였으며, 제안 시스템은 다음과 같은 2가지 알고리즘으로 구성된다.

첫째, 사용자의 주관적 태그 및 빈도수는 낮지만

표 3 'tomato' 이미지의 태그 랭킹 결과 중 상위 5개의 이미지

Image Photo ID	Rank	Baseline	Frequency	The Proposed Method TWMR
 2357390325	1	tomato	tomato	tomato
	2	macro	tomate	color
	3	bokeh	macro	macro
	4	color	color	square
	5	slice	bokeh	slice
 3886158299	1	food	tomato	tomato
	2	foodphotography	food	food
	3	tomato	red	red
	4	tomatoes	green	green
	5	mozzarella	tomatoes	tomatoes
 3228797170	1	tomato	tomato	tomato
	2	tomatoes	red	red
	3	corsica	tomatoes	tomatoes
	4	korsika	nikon	-
	5	corse	abigfave	-
 3029452838	1	tomato	tomato	tomato
	2	studio	canon	canon
	3	50d	brown	brown
	4	canon	girl	girl
	5	girl	box	box

태그간 의미관계가 높은 태그들로 인한 문제를 해결하기 위해 TSSE 알고리즘을 제안하였다. 사용자의 주관적 태그들은 제거하였으며, 빈도수는 낮지만 사전적으로 의미가 높은 태그들에 대해서는 가중치를 부여하여 태그 랭킹에 활용하였다.

둘째, 비정렬된 태그로 인한 비효율적인 정보 내비게이션의 문제를 해결하기 위해 TWMR알고리즘을 제안하였다. 기존 비정렬된 태그 들을 TSSE 알고리즘에 의해 추출된 태그간 의미 유사도를 기반으로 태그들 간의 상관관계를 고려하여 태그 랭킹하였다.

본 논문에서 제안한 시스템의 실험 및 성능평가를 위해 이미지 공유 사이트인 Flickr에서 'lion', 'tomato'의 상위 인기 이미지 각 500개와 그에 태깅된 태그 정보들을 수집하여 총 4개의 부분에 대해 실험 및 평가를 수행하였다.

첫 번째로, 태그간 유사도 추출 실험을 진행하였으며, 태그 동시출현 빈도수와 TSSE 알고리즘을 혼합하여 사용하면 서로의 단점을 보완할 수 있다는 실험 결과가 나타났다.

두 번째로, 태그 랭킹 기법의 성능 실험을 진행하였으며, TWMR 알고리즘을 적용한 랭킹 결과에서 사용자의 주관적 태그들은 제거되고, 사전적으로 의미가 높은 태그들은 상위에 랭킹되는 향상된 태그 랭킹 결과가 나타났다.

향후 연구 과제로는 본 실험에서는 단일 객체의 이미지들을 대상으로 실험을 수행하였으나, 다양한 객체가 존재하는 복잡한 이미지에 대한 태그 랭킹 기법과 대용량 이미지 공유 및 검색을 위하여 다양한 이미지에 대한 추가 실험이 필요하다.

참 고 문 헌

[1] 이시형, 노용만, "소셜 네트워크 활성화에 따른

이미지 태깅 기술 발전 동향," 한국정보진흥원, 2010.
[2] Flickr, http://www.flickr.com
[3] del.icio.us, http://del.icio.us/
[4] Technorati. http://technorati.com/
[5] Jin Ma "The Sustainability and Stabilization of Tag Vocabulary in CiteULike: An Empirical Study of Collaborative Tagging," *Online Information Review,* Vol. 36, No. 5, pp 655-674, 2012.
[6] Leginus M. and Durao, F., "Methodologies for Improved Tag Cloud Generation with Clustering," *Lecture Notes in Computer Science,* Vol. 11, No.7387, pp. 61-75, 2012.
[7] Lu C and Park J, "Exploiting the Social Tagging Network for Web Clustering," *IEEE Transactions on Systems, Man and Cybernetics,* Vol. 41, No. 5, pp. 840-852, 2011.
[8] Li X and Tang D, "nducing Taxonomy from Tags: An Agglomerative Hierarchical Clustering Framework," *Lecture Notes in Computer Science,* Vol, No. 7713, pp. 64-77, 2012.
[9] 이만형, Web2.0 환경에서의 효율적인 웹 콘텐츠 검색을 위한 다중 태그 기반 검색 기법, 경원대학교 박사학위 논문, 2012.
[10] 임영석, 김형주, "사용자 활동 점수에 기반한 태그 검색 개선," 정보과학회논문지, 제17권, 제3호, pp. 155-158, 2011.
[11] 김현우, 김형주, "태그 확장과 시간 정보를 이용한 아이템 추천 방법," 정보과학회논문지, 제18권, 제7호, pp. 521-527, 2012.
[12] 이시화, 황대훈, "3-태그 기반의 웹 이미지 검색 기법," 멀티미디어학회논문지, 제15권, 제9호, pp. 1169-1173, 2012.
[13] 이시화, 태그 유사도를 이용한 클러스터 기반 태그 랭킹 및 검색 시스템, 가천대학교 박사학위 논문, 2012.
[14] Philip Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *International Joint Conference on Artificial Intelligence,* Vol. 14, No. 1, pp. 448-453, 1995.
[15] Dekang Lin, "An Information-theoretic Definition of Similarity," *International Conference on Machine Learning,* Vol. 7, No. 7713, Vol, No, pp. 296-304, 1998.
[16] WordNet 3.1, "WordNet, a lexical database for the English language," http://wordnet.princeton. edu.

이 시 화

2005년 서울보건대학 컴퓨터정보과 졸업
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과(석사)
2012년 가천대학교 전자계산학과(박사)
2013년～현재 메디오피아테크(주) 과장
관심분야 : e-Learning, Context-Aware, Semantic Web, Web2.0, Tag mg′nt

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 computing