Journal of Korea Multimedia Society Vol. 14, No. 11, November 2011(pp. 1491-1499)

위키피디아 기반의 의미 연관성을 이용한 태깅된 웹 이미지의 검색순위 조정

이성재[†], 조수선[††]

요 약

오늘날 이미지, 동영상과 같은 멀티미디어 데이터를 웹 공간에 저장하고 검색할 때, 태그를 이용하는 추세는 보편화되어 있다. 본 논문에서는 태깅된 웹 이미지의 검색에서 태그들의 의미적 중요도를 계산하고, 이를 이용하여 검색 순위를 조정하는 시도를 소개한다. 일반적으로 웹상에 저장된 대부분의 사진 이미지들은 실제로는 중요하지 않지만 사용자의 주관적인 판단으로 추가된 태그들을 다수 포함하고 있으며, 이들은 태그의 단순 비교방식으로 이미지를 검색할 때 정확도를 떨어트리는 주요 원인이 된다. 따라서 어떤 이미지에 붙은 수많은 태그들 중에서 의미적으로 보다 중요한 태그들을 찾아내어 검색에 이용한다면 더욱 만족스러운 검색 결과를 얻을 수 있다. 본 논문에서는 위키피디아 기반의 의미 연관성을 활용하여 검색어 또는 다른 태그들과의 의미 연관성이 높은 태그를 해당 이미지의 대표 태그로 판단하고 이를 이용하여 검색 순위를 조정하는 방법을 제안한다. 실험 결과, 방대한 온라인 백과사전인 위키피디아를 이용하여 계산된 의미적 연관성을 이용함으로써 기존의 연구에 비해 향상된 결과를 얻을 수 있었다.

Tagged Web Image Retrieval Re-ranking with Wikipedia-based Semantic Relatedness

Seongjae Lee[†], Soosun Cho[††]

ABSTRACT

Now a days, to make good use of tags is a general tendency when users need to upload or search some multimedia data such as images and videos on the Web. In this paper, we introduce an approach to calculate semantic importance of tags and to make re-ranking with them on tagged Web image retrieval. Generally, most photo images stored on the Web have lots of tags added with user's subjective judgements not by the importance of them. So they become the cause of precision rate decrease with simple matching of tags to a given query. Therefore, if we can select semantically important tags and employ them on the image search, the retrieval result would be enhanced. In this paper, we propose a method to make image retrieval re-ranking with the key tags which share more semantic information with a query or other tags based on Wikipedia-based semantic relatedness. With the semantic relatedness calculated by using huge on-line encyclopedia, Wikipedia, we found the superiority of our method in precision and recall rate as experimental results.

Key words: Web Image Retrieval(웹 이미지 검색), Tags(태그), Wikipedia(위키피디아), Semantic Relatedness(의미적 연관성)

※ 교신저자(Corresponding Author) : 조수선, 주소 : 충북 충주시 대학로 50 (380-702), 전화 : 043-841-5262, FAX :043-841-5260 , E-mail :sscho@cjnu.ac.kr
접수일 : 2011년 9월 20일, 수정일 : 2011년 11월 9일
완료일 : 2011년 11월 22일
[†] 준회원, 충주대학교 컴퓨터정보공학과 석사과정
(E-mail : lnew1004@gmail.com)
[††] 종신회원, 충주대학교 컴퓨터정보공학과 교수
(E-mail : sscho@cjnu.ac.kr)
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2010-0013307)

1. 서 론

이미지, 동영상 등과 같은 멀티미디어 데이터에 전통적인 분류 기준인 '카테고리' 대신 사용자들이 '태그'를 합동으로 생성하고 관리하는 폭소노미(folksonomy) 방식은 웹 상에서 이미 보편화되어 있다. 폭소노미는 종종 어떤 플랫폼에서 사용되는 공유 태그 집합으로 묘사되는데, 자유롭고 쉽게 접근하여 사용할 수 있으며 유연하고 빠르게 적용할 수 있다는 점이 장점이다. 하지만 태그가 사용자의 개인적 탐색 경험에 기반하기 때문에 일관성 없이 이용자에 따라 지나치게 다양해질 수 있다는 문제점 또한 노출하고 있다[1]. 검색 분야에서 본다면 태그들의 의미정보를 확인하기 어려우므로 동의어, 다의어 등에 대한 어떤 구분도 할 수 없게 한다. 또 수많은 태그들 중 어떤 것이 대표성을 가지는지 알 수 없기 때문에 종종 검색어와 관련이 없거나 내용상 매우 동 떨어진 검색 결과를 산출하기도 한다. 폭소노미 방식의 사진 이미지 공유 사이트인 야후의 플리커(flickr)[2]와 같은 곳에서는 사용자 본인의 사진 관리뿐만 아니라 다른 사용자들에게도 제공할 목적으로 사진을 업로드하고, 자신의 사진이 다양한 검색어에 의해 검색될 수 있도록 수십개의 태그를 붙이고 있다. 그러다보니 정작 검색을 통해 적당한 사진 이미지를 찾고자할 때에는 원하는 것과는 매우 상이한 결과를 얻게 되는 경우가 허다하다.

본 논문에서는 하나의 이미지에 붙은 수많은 태그들 중에서 검색어 또는 다른 태그들과의 의미 연관성이 높은 주요 태그들을 선별하고, 이들을 이용하여 태깅된 웹 이미지의 검색순위를 조정하는 방법을 소개한다. 주어진 검색어와 이미지 태그들의 일치 여부만을 단순 비교하여 검색하는 방식 대신, 검색어와 의미적 유사성 또는 연관성이 높은 태그들을 가진 이미지 중심으로 검색순위를 조정하고자 하는 것이다. 또 검색 대상이 되는 이미지의 많은 태그들 중 다른 태그들과의 의미적 유사성 또는 연관성이 높은 것들을 대표 태그로 보고 이를 이용한 검색 순위 조정 방법도 소개한다. 의미 연관성의 크기를 계산하기 위해 세계 최대의 온라인 백과사전인 위키피디아(Wikipedia)[3]를 이용한다. 최근의 영문 위키피디아에서 중복을 포함하여 천만 개에 이르는 문서들(articles)로부터 문서 제목(topic)들 사이의 유사성 값을 계산하고 이를 이용하여 검색어와 태그 사이, 또는 태그와 태그 사이의 의미 연관성을 측정하는 것이다. 위키피디아는 지금도 끊임없이 확장되고 있는 방대한 규모의 온라인 사전으로서, 집단지성으로 구현되고 다수의 사용자에 의해 일정 정도 검증도 된 양질의 정보를 제공하므로 의미 연관성을 계산하는 좋은 수단이 될 수 있음을 보이고자 한다.

제안하는 위키피디아 기반의 의미 연관성을 이용한 방법이 기존의 방법에 비해 얼마나 효과적인지 확인해보기 위해 플리커에서 검색된 이미지를 대상으로 실험하고 검색 순위 조정의 결과를 평가한다. 본 논문은 다음과 같이 구성된다. 이어지는 2장에서는 관련연구를 설명하고 3장에서는 위키피디아를 이용한 의미 연관성 및 이에 기반한 대표 태그의 결정 과정과 이를 이용한 검색 순위 조정 방법을 설명한다. 4장에서는 실험 및 평가를 다루며, 5장에서 결론을 맺는다.

2. 관련연구

2.1 연관 태그를 이용한 검색

웹2.0의 대표적 기술 중 하나인 태깅은 다수의 일반 사용자들에 의해 만들어지는 매우 유연하고 역동적인 분류체계를 제공하지만 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계 또한 안고 있는 것이 사실이다[1]. 검색 분야에 한정하여 본다면, 단순한 태그기반의 검색은 태그들이 동일한 오브젝트에 달려있다 하더라도 그들이 어떻게 연관되어 있는지를 찾아낼 수 없으므로 많은 한계점을 드러낸다. 이를 극복하기 위한 기존의 연구에서는 연관된 태그(inter-related tag)를 찾아서 그에 따른 분류를 클러스터로 나타내고 검색에 활용할 수 있도록 하였다. 연관태그를 찾아내는 기존 방법에는 어떤 오브젝트에 동시에 출현하는 태그들(co-occurrence tags)끼리의 포함 관계를 이용하여 클러스터링하거나[4], 태그들을 동시 출현하는 횟수를 가중치로 가지는 에지로 연결함으로써 태그 공간을 그래프로 구성한 후, 그 결과를 클러스터링하는 것[5] 등이 있다. 그러나 이와 같은 방법으로는 단순히 연관 태그를 파악할 뿐, 태그들 간의 의미적 연결성은 여전히 밝혀내기가 어렵다. 의미 정보를 이용한 것이 아니고 동시 출현한 횟수 등을 고려하였기 때문이다. 플리커에서 구현

된 '클러스터로 둘러보기' 기능은 연관태그들끼리의 클러스터링을 통해 그에 속하는 이미지를 분류하여 보여주는 것인데 이를 통해 기존 방법의 한계를 알 수 있다.

기존의 연관태그에 의미적 확장을 더하고자 하는 시도로서 워드넷(WordNet)[6] 상의 동의어(synonym) 및 상위어(hypernym) 집합을 이용하여 태그들 간의 연관성을 계산한 후, 이를 이용하여 우선 태그를 추출하고 검색에 적용하는 연구[7]가 있었다. 동시 출현만으로 획득되는 연관성 대신 워드넷 온라인 어휘사전의 적용을 통해 의미적 연관성정보를 획득하여 검색에 사용한 것이다. 또, 의미정보를 얻기 위해 엄격한 온톨로지를 적용하는 대신 간편한 온라인 어휘사전을 적용함으로써 그 효과를 높이고자 한 것이다. 하지만 어휘집합이 풍부하지 못한 워드넷 자체의 한계로 말미암아 의미 연관성 계산이 매우 제한적으로 적용될 수밖에 없는 것이 한계이다.

2.2 위키피디아를 이용한 어휘간 유사성 또는 연관성 계산

위키피디아는 2001년에 시작되어 전 세계 사람들이 공동으로 참여하는 웹 기반의 다언어 백과사전으로, 위키로 만들어져 있어 누구든지 '편집'을 눌러서 내용을 고칠 수 있으며, 목적에 관계없이 자유롭게 사용할 수 있는 자유 콘텐츠 프로젝트이다[3]. 위키피디아는 내용의 접근 및 편집에 대해서는 거의 제한이 없으나, 데이터의 품질은 비교적 엄격하게 관리하여 내용을 검사하고 상호비판을 통해 정리해 나가는 방식으로 평판과 동료평가에 의해 관리되고 있다. 방대한 데이터 량과 비교적 신뢰할만한 품질 등으로 말미암아 현재 인터넷 상에서 가장 활발하게 활용되고 있는 참고정보원 중의 하나이다. 이와 같은 장점 때문에 시맨틱 웹, 지식베이스, 자연어처리 등 다양한 분야에서 활발하게 위키피디아를 이용하고 있다.

위키피디아에 기반하여 어휘간 유사성 또는 연관성 정도를 계산하고자 하는 다수의 연구가 있었는데, Strube & Ponzetto의 WikiRelate[8]는 위키피디아를 이용하여 어휘들 사이의 의미적 연관성을 계산하고자 하는 대표적인 시도이다. 이들은 과거의 연구에서 워드넷에 적용한 path-length measure를 위키피디아에 적합하도록 변형하여 사용하였다. Gabrilovich & Markovitch의 연구[9]에서는 정보검색 분야에서 사용하는 벡터공간 모델을 활용하여 위키피디아 각 문서에 포함되어 있는 모든 텀들로 가중치 벡터를 구성하고 두 문서 사이의 유사성을 해당 벡터들을 비교함으로써 계산하였다. 이와 같은 초기의 방법들보다 효과적이면서 저비용의 측정 방법은 Milne & Witten의 연구[10]에서 발표되었다. 이들은 정보검색 분야에서 널리 사용되는 TF-IDF(Term Frequency-Inverse Document Frequency) 방법을 위키피디아 각 문서 내의 링크 가중치들로 구성된 벡터에 적용하였다. TF-IDF 방법이 문서내 단어들의 TF-IDF 가중치를 계산하여 벡터를 구성하고, 구성된 문서 벡터들의 코사인 거리를 이용하여 두 문서간 유사성을 측정하는 것처럼 위키피디아의 문서들을 링크 가중치로 구성된 벡터로 보고 이들간의 유사성을 두 벡터 사이의 코사인 값으로 계산한 것이다. 이 연구에서는 링크 가중치를 사용함으로써 텀 가중치 벡터를 사용하는 기존 방법에 비해 저비용으로 효과적인 결과를 얻을 수 있음을 입증해 보였다.

위키피디아를 이용하여 어휘의 의미를 추출하려는 연구는 최근 몇몇 국내 논문들[11,12]에서도 찾을 수 있다. 이강표 등의 연구 [11]은 del.icio.us 태그들 간의 상하위 관계를 산출하기 위해 위키피디아 본문을 이용한다. 어떤 태그에 매핑될 위키피디아 텍스트를 찾기 위해 해당 태그와 이웃하는 태그들과의 연관성을 상하위 관계 함수로부터 계산하여 해당 태그의 의미를 결정한다. 김성호 등의 연구 [12]는 한국어 위키피디아를 기반으로 검색 질의어의 중의성을 해소하려는 시도이다. 본 논문에서는 연구[10]에서 발표된 링크 가중치 방법을 적용하여 위키피디아의 두 문서(article) 사이의 유사성을 두 문서의 제목(topic)들 사이의 연관성으로 보고 이를 이용하여 태그들의 의미 연관성을 계산한다. 이는 위키피디아 본문이나 텍스트 텀들을 그대로 이용하는 방법들에 비해 계산이 빠르고 효과적이기 때문이다.

3. 위키피디아 기반의 의미 연관성 계산과 검색순위 조정

3.1 링크 가중치 벡터의 구성과 의미 연관성 계산

Milne & Witten의 연구 [10]에서 밝혔듯이 위키피디아의 각 문서를 그 문서가 포함하는 아웃링크들의 가중치 벡터로 표현하고 이를 이용하여 문서들 간의

유사성을 계산하는 방법은 매우 빠르고 효과적이다. 위키피디아의 문서들을 링크 가중치로 구성된 벡터로 보고 이들 간의 유사성을 두 벡터 사이의 코사인 값으로 계산한다. 링크 가중치들로 문서 벡터를 구성하기 위해 하나의 링크에 대해 살펴보면, 어떤 소스 문서 s로부터 타겟 문서 t로의 아웃링크가 있을 때 이 링크의 가중치 w는 다음 식과 같이 계산된다.

$$w(s \rightarrow t) = \log(\frac{|W|}{|T|}) \quad \text{if } s \in T, \quad 0 \, otherwise \tag{1}$$

여기서 T는 타겟 문서 t를 링크하는 모든 문서들의 집합이고, W는 위키피디아 전체 문서 집합이다. 식에서 보듯이 링크 가중치는 위키피디아 전체 문서 집합에서 타겟 문서를 링크하는 문서들이 차지하는 비율을 역수로 뒤집어 놓은 것이다. 즉, 타겟 문서를 링크하는 비율이 높을수록 해당 소스 문서 내에서 그 링크의 중요도는 떨어진다.

이와 같은 링크 가중치들로 문서 벡터를 구성한 후, 두 문서 사이의 연관성을 벡터들의 코사인 값으로 계산하기 위해서는 아래 식과 같은 두 벡터 A, B 사이의 코사인 유사도(cosine similarity)가 이용된다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{2}$$

이론적으로 위키피디아 전체 문서의 개수인 천만에 이르는 차원의 벡터가 구성되어야 하지만 그와 같은 방법으로 실제 계산은 매우 어려워진다. 본 연구에서는 위키피디아 전체 문서에 알파벳 순서로 고유번호를 부여하고 링크 가중치와 함께 해당 링크의 고유번호를 같이 저장하는 방법으로 벡터를 구성하였다. 벡터간 코사인 유사도는 고유번호가 같은 링크들끼리 가중치를 곱하여 합산함으로써 보다 간편하게 계산된다. 이렇게 계산된 코사인 값은 두 문서 제목 간의 의미 연관성 점수로 이용된다.

3.2 의미 연관성을 이용한 대표 태그의 선별과 검색순위 조정

본 논문에서는 선행 연구와의 비교를 위해 권대현 등의 연구 [7]에서 사용한 검색어 연관성 및 태그 연관성 이용방법을 그대로 적용한다. 검색어 연관성 이용방법은 주어진 검색어와 각 태그 사이의 의미 연관성을 이용하는 방법이고, 태그 연관성 이용방법은 검색어가 아닌 태그들 끼리의 의미 연관성을 이용하는 방법이다. 연구 [7]에서는 워드넷을 기반으로 한 연관성 계산 알고리즘이 사용된 반면, 본 연구에서는 연관성 계산을 위해 위키피디아 기반의 코사인 유사도가 사용된다.

검색어 연관성 이용방법에서는 주어진 검색어와 검색 대상 이미지 각 태그들 사이의 연관성을 위키피디아 기반 코사인 유사도로 계산한 후 그 값이 높은 순으로 대표 태그를 찾는다. 대표 태그는 기존 연구와의 성능 비교를 위해 동일한 개수인 5개로 정하였다. 어떤 사진 이미지 P의 태그집합을 TagSet(P)라고 하고, 그 원소인 각각의 태그들을 T_1, T_2, T_3, ..., 라고 할 때, 검색어 연관성 이용방법으로 태그 T_i의 중요도, TagVal(T_i)를 구하고 5개의 대표 태그를 찾는 알고리즘은 그림 1과 같다. 여기서 WikiCosSimilarity(Keyword, T_i)는 위키 문서 벡터들 간의 코사인 유사도를 이용한 검색어와 태그 사이의 의미 연관성 점수이며 식 2와 같이 계산된다.

```
Length = length(TagSet(P));
for( i = 0 ; i < Length ; i++) {
        TagVal(Ti) = WikiCosSimilarity(Keyword, Ti);
}
for( k = 1 ; k <= 5 ; k++) {
        TagVal(KeyTagk)= 0;
        for( i = 0 ; i < Length ; i++) {
            if (TagVal(Ti) > TagVal(KeyTagk))
            then {
                KeyTagk = Ti;
                selected = i;
            }
        }
    TagVal(Tselected) = 0;
}
```

그림 1. 검색어 연관성 이용방법에 의한 태그 중요도 계산 및 대표 태그의 선택

한편, 태그 연관성 이용방법은 검색어와 상관없이 태그들 사이의 의미 연관성 점수가 높은 태그들을 해당 이미지의 대표 태그들로 보고 이를 검색에서 활용하고자 하는 것이다. 워드넷을 기반으로 한 연관성 계산 알고리즘에서는 각 태그별로 의미단어집합을 구성하는 중간단계가 필요하지만 본 연구에서는 위키피디아 문서 벡터로부터 계산된 코사인 값이 태

그들 간 의미 연관성 점수로 바로 사용되므로 절차가 간단하다. 어떤 이미지에 붙은 전체 태그 집합에서 두 태그들 사이의 의미 연관성 점수를 모두 계산하고, 한 태그에 대하여 이를 합산함으로써 해당 태그의 중요도를 계산한 후, 그 값이 높은 순으로 대표 태그를 선택한다. 태그 T_i의 중요도, TagVal(T_i)를 태그 연관성 이용방법으로 구하는 알고리즘은 그림 2와 같다. 여기서 WikiCosSimilarity(T_i,T_j)는 위키문서 벡터들 간의 코사인 유사도를 이용한 두 태그 사이의 의미 연관성 점수이다. 5개의 대표 태그를 선별하는 방법은 그림 1과 동일하다.

```
Length = length(TagSet(P));
for( i = 0 ; i < Length ; i++) {
        TagVal(Ti) = 0;
        for( j = 0 ; j < Length ; j++)
             TagVal(Ti) += WikiCosSimilarity(Ti,Tj);
}
```

그림 2. 태그 연관성 이용방법에 의한 태그 중요도 계산

검색어 연관성 이용방법 및 태그 연관성 이용방법으로 구한 상위 5개의 대표 태그, $KeyTag_1$, $KeyTag_2$, ..., $KeyTag_5$와 그 중요도, TagVal($KeyTag_1$), TagVal($KeyTag_2$), ..., TagVal($KeyTag_5$)를 따로 저장하고, 이들을 이용하여 연구 [7]과 같은 다단계 방법으로 검색 순위 조정이 이루어진다. 먼저 1차 검색으로, 입력된 검색어가 $KeyTag_1$과 일치하는지 살펴서 일치하는 이미지들을 그 태그 중요도, TagVal($KeyTag_1$)가 높은 순으로 배치한다. 다음으로 1차 검색 결과에 포함되지 않으면서 검색어가 $KeyTag_2$와 일치하는 이미지들을 역시 태그 중요도, TagVal($KeyTag_2$)가 높은 순으로 배치하고, 1차 검색결과에 이어서 결과화면에 나열한다. 동일한 방법으로 5차 검색까지 수행한다. 그림 3은 검색 알고리즘을 순서도로 표현한 것이다.

4. 실험 및 평가

4.1 시스템 구현과 결과 화면의 예

위키피디아 기반의 태그 의미 연관성을 계산하기 위해 영문 위키피디아(2011.01.15.버전) 데이터를 사용하였으며 약 30GB의 pages-articles 타입의 XML

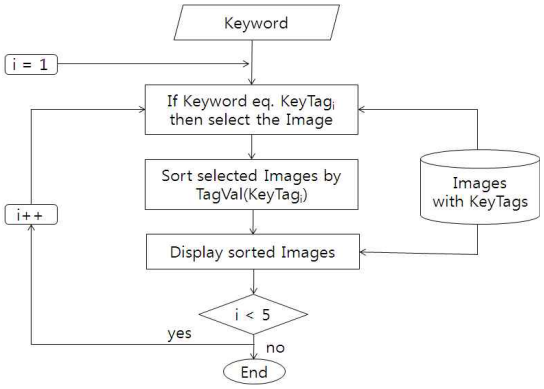


그림 3. 대표 태그를 이용한 다단계 검색

을 위키미디어 다운로드 사이트(http://dumps. wikimedia.org/)에서 내려 받아 사용하였다. 해당 XML 파일은 위키피디아 모든 영문 페이지를 통째로 가지고 있어서 파일 하나의 용량이 30GB가 넘기 때문에 스트림으로 처리하는 작업이 필요하였다. 이를 위해 JAVA 기반의 StAX 라이브러리를 사용하여 XML 파일을 스트림 처리하였으며 XML에서 MySQL을 이용하여 저장하였다. 수집된 문서(article)의 수는 10,861,570(중복 포함) 건이었으며 의미 있는 전체 아웃링크의 수는 75,261,480(중복 포함)개였다.

실험에 사용할 플리커 이미지 데이터를 수집하기 위해 공개된 플리커 API를 사용하였다. 플리커에서 검색어 'house'로 검색된 상위 이미지 1500개를 플리커 API의 XML로 출력되는 레코드를 파싱하여 자동 저장하였다. 1500개 이미지들의 총 태그 수는 중복 포함하여 153,993개였으며 워드넷과 위키피디아 적용 알고리즘을 이용하여 각각의 태그 순위를 결정하였다. 전반적인 시스템 개발 환경은 표 1과 같다.

전체 실험 결과 중 태그 연관성 이용방법으로 검색한 결과들과 플리커 검색 결과를 비교하기 위한 검색 첫 화면이 그림 4와 그림 5, 그리고 그림 6에 나타나 있다. 그림 6의 플리커와 그림 5의 워드넷 이용 방법에서는 첫 화면 24개 이미지 중 정확한 자료가 각각 14개, 17개였으나 그림 4의 위키피디아 이용 방법에서는 24개 모두 적합으로 나타나 원하는 결과를 앞 페이지에서 바로 찾고자 하는 사용자의 만족도를 월등히 향상시켜주고 있다.

4.2 시스템 성능 평가

본 연구에서는 기존 연구 결과와의 성능 비교를

표 1. 시스템 개발 환경

종 류		구 성	사 양
서버 시스템	H/W	CPU	제온 E5506 * 2
		RAM	DDR3 4GB ECC * 4
		HDD	1TB * 2 (RAID 0 구성)
	S/W	OS	CentOS 5.6
		DBMS	My-SQL 5
		Web Server	Apache Server 2.2
		Programming Language	PHP 5.2.4
		기타	StAX, FlikrAPI, WordNet 3.0
데이터 수집 시스템	H/W	CPU	코어2쿼드 Q9400
		RAM	DDR2 4GB
		HDD	500GB
	S/W	OS	Windows 7
		DBMS	My-SQL 5
		Programming Language	JAVA, PHP
	Data	Wikipedia	Database backup dumps (2011. 01. 15)
		Flikr Images	'house'와 관련된 상위 1500개의 검색결과

위해 1500개 'house' 이미지에 대해 검색 결과의 적합 여부를 수작업으로 판정하였다. 판정의 기준은 검색어 'house'를 집으로 보이는 건축물의 외형으로 판단하고 이에 해당되면 '적합' 해당되지 않으면 '부적합'으로 판정하였다. 플리커 검색, 워드넷 기반의 검색, 위키피디아 기반의 검색을 상위 20개 페이지에 대해 각각 비교한 결과는 표 2와 같다. 각 페이지에는 그림 4, 5, 6과 같이 24개의 이미지가 출력된다. 이전 연구 [7]에서 수행한 실험 결과에서는 플리커의 적합 이미지 비율이 검색된 결과 페이지들의 순위에 상관없이

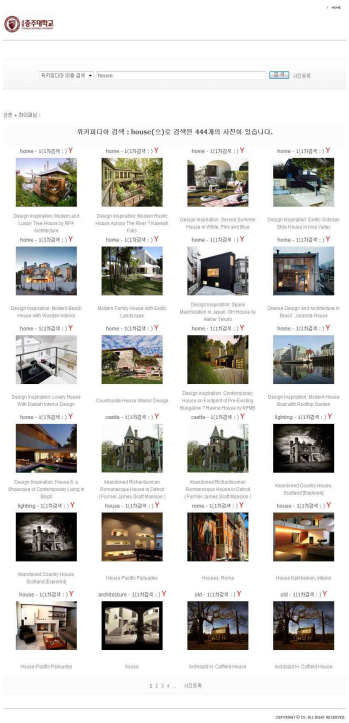


그림 4. 위키피디아 연관성 이용방법 결과

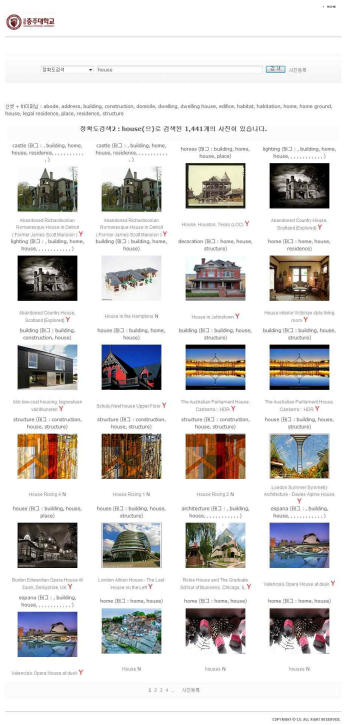


그림 5. 워드넷 연관성 이용방법 결과

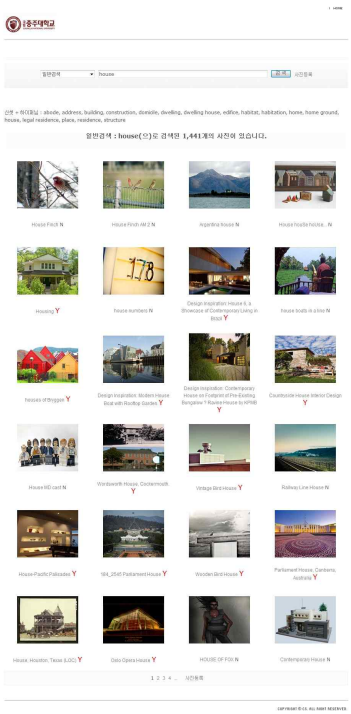


그림 6. 플리커 검색 결과

대략 30%에서 50% 정도로 나타난 반면, 본 실험에서는 대부분의 페이지에서 50%이상의 비율로 적합한 이미지들이 나타났고 20개 페이지 전체 적합 비율도 56%로 높아졌다. 2년 전에 비해 플리커 검색 엔진 자체의 성능이 향상된 것을 알 수 있다.

본 논문에서 제안하는 위키피디아 연관성 이용방법은 20개 페이지의 평균 적합비율이 태그 연관성 이용방법에서 66%, 검색어 연관성 이용방법에서 84%로 나타나 플리커 검색의 결과뿐만 아니라 워드넷 연관성 이용방법의 두 평균 59%, 58%에 비해서도 높은 것으로 나타났다. 특히 위키피디아 기반의 검색어 연관성 이용방법에서 평균 적합 비율이 84%로 가장 좋은 결과를 보였으며, 첫 두 페이지에서 적합 비율이 모두 100%로 기존 방법에 비해 월등히 높게 나타났다.

5가지 검색의 성능을 보다 정확히 비교 평가하기 위해 첫 4개, 8개, 12개, 16개, 20개 페이지에 속한 적합한 데이터의 비율을 조사하여 정확도(precision)와 재현율(recall)을 계산하였으며, 표3과 같은 결과를 얻었다. 정확도는 해당 페이지들의 전체 검색 이미지 수에 대한 적합 이미지의 수로 계산하고, 재현

표 2. 검색 결과 비교표

구분 / 페이지	일반검색(플리커)		워드넷 태그 연관성 이용		워드넷 검색어 연관성 이용		위키 태그 연관성 이용		위키 검색어 연관성 이용	
	적합	적합비율	적합	적합비율	적합	적합비율	적합	적합비율	적합	적합비율
1	14	0.58	17	0.71	14	0.58	24	1.00	24	1.00
2	16	0.67	21	0.88	21	0.88	20	0.83	24	1.00
3	19	0.79	18	0.75	19	0.79	17	0.71	20	0.83
4	17	0.71	19	0.79	19	0.79	15	0.63	20	0.83
5	13	0.54	22	0.92	8	0.33	12	0.50	17	0.71
6	15	0.63	12	0.50	16	0.67	9	0.38	18	0.75
7	17	0.71	11	0.46	13	0.54	16	0.67	13	0.54
8	15	0.63	13	0.54	14	0.58	17	0.71	18	0.75
9	14	0.58	16	0.67	11	0.46	13	0.54	19	0.79
10	7	0.29	8	0.33	15	0.63	19	0.79	20	0.83
11	13	0.54	14	0.58	17	0.71	19	0.79	19	0.79
12	13	0.54	16	0.67	11	0.46	11	0.46	20	0.83
13	13	0.54	15	0.63	11	0.46	11	0.46	23	0.96
14	10	0.42	13	0.54	11	0.46	16	0.67	24	1.00
15	10	0.42	10	0.42	13	0.54	12	0.50	22	0.92
16	12	0.50	14	0.58	10	0.42	18	0.75	20	0.83
17	12	0.50	10	0.42	16	0.67	14	0.58	19	0.79
18	13	0.54	10	0.42	12	0.50	19	0.79	24	1.00
19	10	0.42	11	0.46	18	0.75	9	0.75	20	0.83
20	15	0.63	14	0.58	10	0.42	0	0.00	21	0.88
전체적합비율	268	0.56	284	0.59	279	0.58	291	0.66	405	0.84

표 3. 검색 결과 정확도 및 재현율 비교표

구분 / 페이지	일반검색(플리커)			워드넷 태그 연관성			워드넷 검색어 연관성			위키 태그 연관성			위키 검색어 연관성		
	적합	정확도 (%)	재현율 (%)	적합	정확도 (%)	재현율 (%)	적합	정확도 (%)	재현율 (%)	적합	정확도 (%)	재현율 (%)	적합	정확도 (%)	재현율 (%)
TOP4	66	68.8	7.1	75	78.1	8.0	73	76.0	7.8	76	79.2	8.1	88	91.7	9.4
TOP8	126	65.6	13.5	133	69.3	14.3	124	64.6	13.3	130	67.7	13.9	154	80.2	16.5
TOP12	173	60.1	18.5	187	64.9	20.0	178	61.8	19.1	192	66.7	20.6	232	80.6	24.9
TOP16	218	56.8	23.4	239	62.2	25.6	223	58.1	23.9	249	64.8	26.7	321	83.6	34.4
TOP20	268	55.8	28.7	284	59.2	30.4	279	58.1	29.9	291	65.5	31.2	405	84.4	43.4

율은 전체 1500개 이미지 중 적합한 이미지의 수(총 933개)에 대한 검색된 적합 이미지 수로 계산한다.

그림 7과 그림 8에서 나타난 것과 같이 본 논문에서 제안하는 위키피디아 기반의 연관성을 이용했을 때, 검색어 연관성 이용방법이나 태그 연관성 이용방법 모두에서 우수한 결과를 보였다. 특히 위키피디아 기반의 검색어 연관성 이용방법에서는 정확도와 재현율에서 다른 방법들에 비해 월등히 우수한 결과를

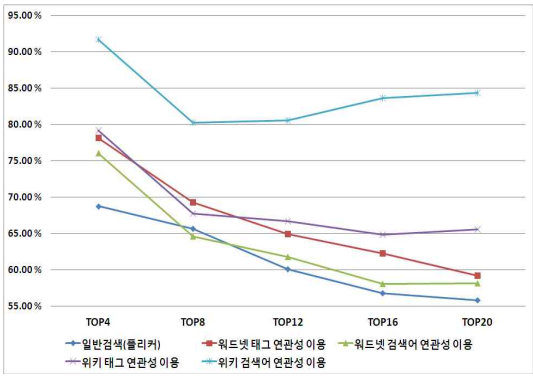


그림 7. 정확도 비교 그래프

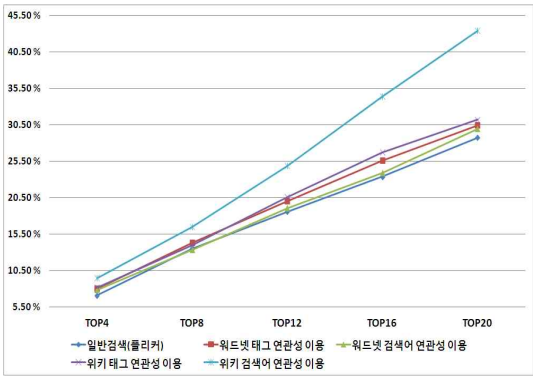


그림 8. 재현율 비교 그래프

나타내었다. 상위 20개 페이지에서 평균 84.4%의 정확도를 보였고, 전체 63개 페이지에 해당하는 데이터(1500개 이미지) 중 20개 페이지만으로 43.4%의 재현율을 나타내었다.

태깅된 웹 이미지의 검색시스템에서 사용자들은 원하는 이미지들을 검색 결과의 앞쪽 페이지에서 발견하기를 선호한다. 즉, 찾고자 하는 이미지가 검색어와 공통의 의미를 지닌 이미지들로서 정확성을 만족시키면서도, Top20 등 앞쪽 페이지에서 더 많이 보여줄 수 있다면 사용자의 만족도를 크게 높일 수 있을 것이다. 실험 결과로부터 태깅된 이미지 검색시스템에서 위키피디아 기반의 의미 연관성을 이용하여 검색 순위를 조정하는 것이 매우 효과적인 방법임을 알 수 있다.

5. 결 론

본 논문에서는 태깅된 웹 이미지 검색 시스템에서 태그들의 의미 연관성을 이용하여 검색순위를 조정하고자 할 때, 위키피디아를 이용하는 것이 기존 연구[7]의 워드넷을 이용하는 방법보다 월등히 효과적임을 입증하였다. 특히 위키피디아 본문을 사용하는 것이 아니라 링크 정보만을 사용함으로써 보다 간편하게 코사인 유사도 값을 계산할 수 있었고, 이 유사도 값만을 이용하면서도 정확도 및 재현율에서는 눈에 띄는 향상을 얻을 수 있었다. 이것은 세계 최대의 온라인 백과사전인 위키피디아의 내용적 풍부함을 실증하는 예라고 할 수 있다. 특히 위키피디아 본문의 내용 정보를 이용하지 않고 링크 정보만을 이용하면서도 높은 검색 성능을 얻을 수 있었다는 것이 중요한 연구 성과이다.

제안된 방법은 복잡한 알고리즘을 사용하는 기계

학습기반의 데이터 분류(classification) 또는 군집화(clustering) 방법에 비해 매우 빠르고 간편하게 적용될 수 있으므로 실용성이 높은 방법이라 할 수 있다. 특히 태그 연관성 이용방법에서는 태그의 선별이 검색 시점이 아닌 데이터 업로드 시점에 이루어지므로 검색 시간에 영향을 미치지 않는다는 장점이 있다. 검색어 연관성 이용방법에서도 이미지에 붙은 태그들 중 하나를 적합한 검색어로 보고 각 검색어 별로 대표 태그를 데이터 업로드 시점에 미리 선별해서 저장해 둘 수 있으므로 동일한 장점을 살릴 수 있다. 향후 연구에서는 구축된 위키피디아 DB를 활용하여 의미적 연관성을 이용할 수 있는 다양한 응용 분야에 적용할 계획이다.

참 고 문 헌

[1] 이정미, "폭소노미의 개념적 접근과 웹정보서비스에의 적용," 한국비블리아학회지, 제8권 제2호, pp. 141-159, 2007.

[2] 플리커 사이트, http://www.flickr.com/

[3] 위키피디아 사이트, http://www.wikipedia.org/

[4] P. Schmitz, "Inducing Ontology from Flickr Tags," In Proc. of the Collaborative Web Tagging Workshop at WWW '06, 2006.

[5] G. Begelman, et al., "Automated Tag Clustering: Improving Search and Exploration in the Tag Space," In Proc. of the Collaborative Web Tagging Workshop at WWW '06, 2006.

[6] G. A. Miller, "WordNet: An On-line Lexical Database," *International Journal of Lexicography*, Vol.3, No.4, 1990.

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

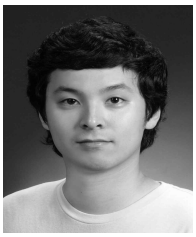
[8] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic Relatedness using Wikipedia," *In Proc. of Association for the Advancement of Artificial Intelligence (AAAI'06)*, pp. 1419-1424, 2006.

[9] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis," In Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), pp. 1606-1611, 2007.

[10] D. Milne and I. H. Witten, "An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links," In Proc. of Association for the Advancement of Artificial Intelligence : Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008), pp. 25-30, 2008.

[11] 이강표, 김현우, 장충수, 김형주, "FolksoViz: Wikipedia 본문을 이용한 상하위 관계 기반 폭소노미 시각화 기법," 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제14권 제4호, pp. 341-457, 2008.

[12] 김성호, 배상준, 고영중, " 한국어 위키피디아 기반 질의어 중의성 해소 및 확장 자질 추출 기법," 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제17권 제3호, pp. 141-204, 2011.



이 성 재

2007년 3월~2011년 2월 충주대학교 컴퓨터정보공학과 공학사

2011년 3월~현재 충주대학교 대학원 컴퓨터정보공학과 재학 중

관심분야 : 데이터마이닝, 모바일 웹

조 수 선

1983년 3월~1987년 2월 서울대학교 계산통계학과 이학사

1987년 3월~1989년 2월 서울대학교 대학원 계산통계학과 이학석사

2000년 3월~2004년 2월 충남대학교 대학원 컴퓨터과학과 이학박사

1994년 6월~2004년 3월 한국전자통신연구원 선임연구원

2006년 7월~2007년 7월 미시간대학교(앤아버) 통계학과 방문연구원

2004년 4월~현재 충주대학교 컴퓨터정보공학과 부교수

관심분야: 웹 컨텐츠 마이닝, 모바일 웹, 기계학습