Journal of Korea Multimedia Society Vol. 12, No. 7, July 2009(pp. 1032-1042)

워드넷 의미정보로 선별된 우선 태그와 이를 이용한 웹 이미지의 검색

권대현[†], 홍준혁[††], 조수선[†††]

요 약

본 연구는 태깅된 웹 이미지의 검색에서 태그들의 의미정보를 미리 추출하여 검색 시에 이용하고자 하는 것이다. 일반적으로 웹 이미지의 태그들은 사용자들에 의해 순서 구분 없이 무작위로 매겨지며 많게는 그 수가 100여개에 이른다. 본 논문에서는 이 태그들 간에 의미정보가 많이 공유된 것일수록 해당 이미지를 설명하는 중요 태그가 될 것임에 착안하여 이미지와 태그 정보가 업로드되는 시점에 중요도에 따른 우선 태그를 결정하고 이를 검색에 활용하는 방법을 소개한다. 제안된 방법은 워드넷에 기반하여 태그의 연관성점수를 계산하고 이를 이용하여 다단계 검색으로 태깅된 웹 이미지를 검색한다. 평가를 위하여 제안된 방법으로 검색된 결과와 검색어와 태그의 단순 비교방식인 기존의 검색을 비교하였으며 실험 결과, 정확도와 재현율에서 본 시스템의 우수함을 확인할 수 있었다.

Web Image Retrieval using Prior Tags based on WordNet Semantic Information

Daehyeon Kweon[†], Junhyeok Hong[††], Soosun Cho[†††]

ABSTRACT

This research is for early extraction and utilization of semantic information from the tags in tagged Web image retrieval. Generally, users attach a tag to a Web image with little thought of the order, up to over 100 ones. In this paper, we suggest a method of selecting prior tags based on their importance when tagged images are uploaded, and using them in image retrieval. Ideas came from the recognition of the important tags which give a better description of the image as the tags sharing more semantic information with other tags of the same image. This method includes calculation of relation scores between tags based on WordNet and multilevel search of tagged images with the scores. For evaluation, we compared the suggested method and other retrieval methods searching images with simple matching of tags to a given keyword. As the results, we found the superiority of our method in precision and recall rate.

Key words: Web Image Retrieval(웹이미지 검색), Prior Tags(우선 태그), Semantic Information(의미정보), WordNet(워드넷)

※ 교신저자(Corresponding Author) : 조수선, 주소 : 충북 충주시 대학로 72번지(380-702), 전화 : 043)841-5262, FAX : 043)841-5260, E-mail : sscho@cjnu.ac.kr
접수일 : 2009년 2월 2일, 완료일 : 2009년 6월 4일
[†] 준회원, 충주대학교 대학원 전자계산학과 석사과정
(E-mail : kfsura@nate.com)
[††] 준회원, 충주대학교 대학원 전자계산학과 석사과정
(E-mail : sunghyeon@gmail.com)
[†††] 종신회원, 충주대학교 컴퓨터과학과 조교수
※ 이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2008-531-D00027)

1. 서 론

웹 2.0 환경에서 태그기반의 데이터 분류는 새로운 인터넷 기술의 대표로 자리 잡고 있다. 이것은 컨텐트 분류를 위해 전통적인 분류 기준인 '카테고리' 대신 사용자들이 '태그'를 합동으로 생성하고 관리하는 방식인데 현재 인터넷 사용자들로부터 큰 호응을 얻고 있으며 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 활용되고 있다[1]. 웹 이미지 정보에서는 야후의 플리커(flickr)[2]와 같은 대규모 사진정보 공유사이트가 대표적인 예가 되고 있다. 플리커 사용자들은 본인의 사진 관리뿐만 아니라 다른 사용자들에게도 제공할 목적으로 사진을 업로드하고, 자신의 사진이 다양한 검색어에 대해 검색될 수 있도록 많으면 수십개 또는 백여개에 이르는 태그를 붙인다. 그러다 보니 정작 검색을 통해 적당한 사진 이미지를 찾고자 하는 사용자 입장에서는 원하는 것과는 상이한 결과를 얻게 되는 경우가 허다하다. 사용자에 의해 입력되는 태그가 매우 사적이고 체계화되지 못한 정보이며 실제로 이미지의 내용을 표현하는데 부적절한 것도 다수 포함되어 있기 때문이다.

본 논문에서는 이와 같은 태깅된 웹 이미지 검색에서 태그정보의 태생적인 한계를 극복하고 검색 결과의 만족도를 높이기 위한 시도를 소개한다. 그것은 프린스턴대학에서 개발된 워드넷(WordNet)[3]을 이용하여 의미정보를 많이 포함하고 있는 순으로 태그들을 순서화하여 검색에 이용하는 것이다. 특히 워드넷기반의 연관성점수를 이용한 이전 연구[4]의 결과를 향상시키기 위해 정보 검색 시점이 아닌 정보가 업로드되는 시점에 태그들 간의 연관성점수를 계산하여 이를 반영한 이미지 검색 방법을 제안한다.

본 연구가 기여하는 점은 다음과 같다. 첫째, 태그들 간의 연관성을 점수로 정량화하기 위해 어휘들의 의미 정보를 나타내는 워드넷을 이용함으로써 태깅된 이미지 검색에 워드넷 기반의 의미적 연관성(semantic relation)의 개념을 도입한다. 둘째, 다수의 태그들 중 이미지를 대표할만한 의미있는 태그들을 선순위로 두고 그렇지 않은 것들을 후순위로 넘김으로써 우선 태그(prior tags)를 선별하고 이를 검색에 활용함으로써 정확도 및 재현율에서 검색 성능을 향상시킨다. 셋째, 워드넷 의미정보를 이용한 우선 태그의 선별은 데이터 업로드 시에 일어나므로 검색 시의 계산량에 전혀 영향을 미치지 않는다. 따라서 검색 속도 면에서 다른 방법에 비해 유리하다.

본 논문에서는 제안된 시도의 효과를 확인하기 위해 현재 세계적으로 인기있는 태그기반의 사진 이미지 공유 사이트 중 하나인 야후의 플리커를 대상으로 실험하고 검색 결과를 평가하였다. 논문의 구성은 다음과 같다. 이어지는 2장에서는 조사된 관련연구를 설명하고 3장에서는 워드넷에 기반한 연관성점수 계산과 우선 태그의 결정, 그리고 우선 태그를 이용한 검색 알고리즘을 설명한다. 4장에서는 실험 및 평가를 다루며, 5장에서 결론을 맺는다.

2. 관련연구

2.1 태깅된 데이터의 검색

웹2.0의 대표적 기술 중 하나인 태깅은 매우 유연하고 역동적인 분류체계를 제공하지만 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계 또한 안고 있는 것이 사실이다. 연구[5]에서는 태깅이 지니고 있는 한계들에 대해 요약하고 있다. 그 내용을 보면, 첫째로, 태깅은 동의어나 유의어에 대한 관리를 제공해주지 않는다. '해변'과 '바닷가'등과 같은 태그들은 실제로는 서로 의미가 동일하거나 유사함에도 불구하고 서로 다른 태그로 분류되는 것이다. 둘째로, 태깅은 정보의 검색에 있어서 정확도(precision)가 떨어진다. 따라서 태그는 어떤 정보를 넓은 범주의 카테고리에 위치시키는 데에는 매우 유용하지만, 사용자가 원하는 정확한 정보를 찾아내는 데에는 효율적이지 못한 것으로 지적하고 있다. 두 번째 문제로 지적한 정확도가 떨어지는 현상은 많은 경우, 첫 번째 문제인 동의어나 유의어에 대한 고려가 없다는 데에서 기인한다고 볼 수 있다.

따라서 단순한 태그기반의 검색은 태그들이 동일한 이미지에 달려있다 하더라도 그들이 어떻게 연관되어 있는지를 찾아낼 수 없으므로 많은 한계점을 드러낸다. 이와 같은 한계를 극복하고자 태그들의 동시 출현(co-occurrence)에서 서로 연관된 태그들(inter-related tags)을 찾아내어 검색에 활용하려는 다양한 시도들이 있었다[6,7]. 더불어 태그의 의미 정보를 활용하기 위해 씨멘틱 웹을 이용하여 태그들의 의미적 연관성(semantic relation)을 찾아내어 온톨

로지 매칭으로 검색하고자 하는 시도들도 있었다[8,9]. 연관된 태그들을 이용하여 클러스터링하는 방법은 현재 플리커에서 클러스터 보기 기능으로 제공되고 있으나 동일 클러스터로 분류된 태그들 간에 어떠한 공통 의미를 가진 것이 아니므로 서로 다른 클러스터로 분류된 이미지들 사이에 뚜렷한 차이를 보이지 않는 경우가 많다. 또한, 온톨로지를 접목하여 태그 의미 정보를 찾아내고 이를 클러스터링하고자 할 때에는 태그들의 특성상 특정 온톨로지 상에 나타나는 의미적 연관성들을 찾아내기 힘들어 실험결과 클러스터링에 포함되는 범위가 매우 제한적인 것으로 알려졌다[8].

본 연구에서는 이와 같은 태그기반 검색의 한계를 해결하기 위해 워드넷 상의 동의어(synonym) 및 상위어(hypernym) 집합을 이용하여 태그들 간의 연관성을 계산한 후, 이를 이용하여 우선 태그를 추출하고 검색에 적용하였다. 즉, 동시 출현만으로 획득되는 연관성(relation) 정보대신 워드넷 온라인 어휘사전의 적용을 통해 의미적 연관성(semantic relation) 정보를 획득하여 검색에 사용하는 것이다. 또, 의미정보를 얻기 위해 엄격한 온톨로지를 적용하는 대신 간편한 온라인 어휘사전을 적용함으로써 그 효과를 높이고자 하였다. 즉, 본 연구는 두 가지 방향으로 진행되는 기존 연구들 중 일부를 확장하면서 동시에 틈새를 공략한 연구 방법이라고 할 수 있다.

한편, 연구[10]에서는 태그기반 검색의 정확도가 떨어지는 문제를 해결하고자 하나의 이미지에 태깅된 여러 태그들 사이의 연관 태그 매핑을 통해 태그 가중치 매트릭스를 생성하고 이를 이용한 클러스터링에 기반하여 검색시스템을 개발하였다. 시스템 성능 평가 결과 정확도는 매우 향상되었지만 재현율이 다소 떨어지는 것으로 확인되었다. 그 이유는 태그들이 검색어를 포함하더라도 연관성이 높은 태그쌍을 포함하고 있지 않으면 클러스터에 들어가지 않기 때문이다. 본 연구에서는 태그들 사이의 연관성을 고려함에 있어서 태그 본래의 대표적 의미가 나머지 태그들에 얼마나 잘 표현되어 있는가를 계산함으로써 동의의나 유의어에 대한 적극적인 고려를 하였으며, 그 결과, 검색 화면의 상위 4개, 8개, 12개 페이지 등에서 정확도와 재현율의 현격한 향상을 확인할 수 있었다.

특히 본 논문에서 제안하는 방법은 태그연관성점수로 우선 태그를 선별하는 것이 검색 시점이 아닌 데이터 업로드 시점에 이루어지므로 이를 위한 계산량은 검색 시간에 아무런 영향을 미치지 않는다는 것이 장점이다. 즉 검색어와 상관없이 태그들끼리의 연관성을 미리 계산하여 우선태그들을 저장하고, 검색 시점에는 키워드와 이 우선 태그들 간의 매칭만 확인하므로 키워드와 대량의 태그들 간의 연관성을 계산하는 기존의 연구에 비해 속도면에서 우위를 차지하게 된다.

2.2 워드넷을 이용한 어휘간 유사성 또는 연관성 계산

워드넷은 인간의 어휘 체계를 표현하기 위해 1990년대 프린스턴대학에서 개발된 영어 어휘 데이터베이스로서 현재 유닉스 버전으로 WordNet 3.0이 공개되어 있다[3]. 어휘들 간의 관계로 동의어(synonym), 반의어(antonym), 하위어(hyponym), 상위어(hypernym) 등을 표현함으로써 어휘의 의미에 대한 카테고리 분류가 잘 정의되어 있으며, 어휘의 계층구조와 연관관계가 잘 표현되어져 있다. 특히 워드넷은 명사, 동사, 형용사, 부사들이 동의어 집합인 Synset으로 나뉘어져 있고, Synset과 Synset 사이의 의미적인 관계를 표현한 링크로 이루어져 있으며[11], 이 Synset을 이용하여 개념간의 유사성을 측정하는 방법들이 에지기반 측정방법, 노드기반 측정 방법, 의미기반 측정방법 등 다양하게 발표되고 있다[12,13].

연구[14]에서는 워드넷 명사사전을 활용하여 검색엔진 질의어의 모호성을 해결하고자 하였다. 이 연구에서는 워드넷 명사사전에서 동의어, 상위어, 주석 등을 추출하여 사용자에게 나타내고, 사용자로 하여금 질의어의 의미를 선택하게 하여 재조합 질의어를 생성하였으며, 이렇게 완성된 재조합 질의어를 이용하여 검색엔진의 성능을 향상시켰다. 하지만 재조합 질의어를 이용하기 위해 사용자 인터페이스 측면에서 한 단계 복잡도가 증가하는 방식이며, 검색시점에 워드넷을 통한 동의어, 상위어 등의 추출과 후보 어휘들이 결정되는 방식이므로 본 연구와는 차이가 있다. 기타 의미적 데이터 검색을 위해 워드넷을 활용한 여러 연구들[15-17]도 워드넷 기반의 단어간 의미 유사성을 측정하기 위한 다양한 방법을 소개하고 있지만 결국 모든 계산이 검색 시점에 이루어지므로 본 논문에서 제안하는 업로드 시점의 우선 태그 선별 방법과는 차이가 있다.

3. 연관성점수에 기반한 우선 태그 선별과 검색

3.1 워드넷을 이용한 의미단어집합의 구성과 연관성 점수 계산

연관 태그를 찾아내는 기존의 방법은 하나의 오브젝트에 동시에 출현하는 태그들(co-occurrence tags)끼리의 포함 관계를 이용하여 클러스터링하거나[18], 태그들을 동시 출현하는 횟수를 가중치로 가지는 에지로 연결함으로써 태그 공간을 그래프로 구성한 후, 그 결과를 클러스터링하는 것[6] 등이 있다. 그러나 이와 같은 연관 태그 구성만으로는 태그들 간의 의미적 차이는 여전히 구분하기가 어렵다.

연구[4]에서는 워드넷을 이용하여 검색어의 동의어와 상위어들로 이루어진 '의미단어집합(set of word meanings)'을 구성하고 이것과 태그들 간의 일치 여부를 판단하여 '연관성 점수(relation score)'를 계산한 후, 이를 이용하여 검색 이미지를 재정렬하였다. 본 연구에서는 연구[4]의 의미단어집합 구성 방법을 그대로 사용하되 이를 검색어가 아닌 태그들 각각에 적용하여 태그연관성점수를 계산하고자 한다.

의미단어집합의 구성 방법은 먼저 해당 어휘의 첫 번째 동의어와 상위어를 구해 집합에 포함시키고, 구성된 집합 내 각각의 단어들에 대해 다시 첫 번째 동의어와 상위어를 구해 의미단어집합에 포함시키는 것이다. 첫 번째 동의어만을 이용하는 이유는 워드넷이 단어의 여러 가지 의미 중에서 의미상의 사용빈도에 따라 동의어의 순위를 매기고 있으므로 해당 어휘를 가장 대표적으로 표현하는 동의어를 사용하기 위해서이다. 한편, 첫 번째 동의어만 포함함으로써 빈약해지기 쉬운 의미단어집합은 상위어를 포함시킴으로써 적당히 확장되는 효과를 가지게 된다.

이상의 의미단어집합을 이용하여 이미지별로 검색어와의 연관성점수를 계산하는 방법은 검색어의 의미단어집합에 몇 개의 태그가 포함되어있는지 그 개수를 세고, 또, 입력된 태그의 순위를 살펴서 계산한다. 구체적인 의미단어집합 구성 예와 의미단어집합을 이용하여 연관성점수를 계산하는 알고리즘은 연구[4]를 참조하기 바란다.

3.2 태그들의 상호 연관성점수 계산

앞 절에서 검색어와 태그들의 연관성점수를 계산하기 위해서는 검색어의 의미단어집합에 이미지의 각 태그들이 얼마나 많이 포함되어 있는지를 살펴서 점수화한다고 하였다, 즉 검색어의 의미를 제대로 내포한 태그들을 많이 가지고 있는 이미지일수록 검색순위에서 우위를 차지하게 된다. 본 논문에서는 이와 같은 아이디어를 태그들 끼리에도 적용한다면 보다 우수한 검색 효과를 얻을 수 있을 것이라 예상하여 태그들 간의 연관성점수를 계산하여 검색에 이용하고자 한다. 태그들 간의 연관성점수가 뜻하는 바는 다음과 같다, 이미지의 어떤 태그가 다른 태그들의 의미단어집합에 여러 번 포함될수록 그 태그는 다른 태그들과 의미가 중복되는 것이므로 다른 태그들과의 연관성이 높다고 보는 것이다. 따라서 태그연관성점수가 높은 태그가 그 이미지를 표현하는 데 있어서 중요도가 높은 우선 태그(prior tags)가 되므로 이를 검색에서 활용하고자 하는 것이다. 앞 절에서 설명한 검색어의 연관성점수를 이용하는 방법과 비교하여 살펴보면 표 1과 같다.

표 1. 연관성점수를 이용한 검색방법 비교표

	검색어연관성 이용방법	태그연관성 이용방법
목적	검색어의 본래의미를 잘 표현하는 태그를 많이 가진 이미지 순으로 검색	태그연관성점수로 태그들의 중요도를 계산하여 우선 태그를 정한 다음 이를 이미지 검색에 활용
검색방법	각 이미지에 대한 검색어연관성점수를 구하고 이 점수가 높은 순으로 결과 배치	각 이미지별로 우선 태그들만을 검색어와 차례로 비교하여 일치하면, 해당 이미지를 그 우선 태그의 태그연관성점수가 높은 순으로 배치
연관성 점수 계산방법	검색어의 의미단어집합을 구성하고 각 이미지의 태그들이 이 집합에 포함되는 횟수가 많을수록 점수 증가, 이미지 당 하나의 검색어연관성점수를 가짐	하나의 이미지에 딸린 태그집합에서 어떤 태그가 다른 태그의 의미단어집합에 포함되는 횟수가 많을수록 점수 증가, 태그별로 하나씩 태그연관성점수를 가짐
의미단어 집합	검색어의 의미단어집합 구성	이미지의 각 태그별로 의미단어집합 구성

어떤 사진 이미지 P의 태그집합을 TagSet(P)라고 하고, 그 원소인 각각의 태그들을 T_1, T_2, T_3, ..., 라고 할 때, 태그 T_i의 태그연관성점수(TagVal(T_i))를 구하는 알고리즘은 그림 1과 같다. 여기서 MeanSet(T_i)는 태그 T_i의 의미단어집합을 나타내며, 함수 WordNet()은 워드넷을 이용하여 매개변수로 받은 어휘의 의미단어집합을 구하는 함수이다. WordNet(T_j)는 앞 절에서 설명한 바와 같이 워드넷 상에서 어휘 T_j의 Synset과 T_j의 상위어에 대한 Synset의 합집합이다. WordNet()을 의미단어집합 MeanSet()에 대응시킨 후, 어떤 태그 T_i가 다른 태그들의 의미단어집합 MeanSet()에 한번 포함될 때마다 T_i의 태그연관성점수에는 상수 a가 더해지도록 하였다. 또한 각 태그들은 등록자의 입장에서 입력한 순서에 따라 이미지와 더 밀접한 관계를 가진 것이므로 이를 감안하여 등록자의 입력순서를 고려한 가산점을 더하도록 하였다. 가산점은 태그순서가 첫 번째일 때 0.99a가 되고(의미단어집합에 한번 포함되는 것과 거의 같은 효과), 99번째일 때 0.01a가 되며, 100번째 이상일 때는 점수가 없도록 하였다. 이는 최대 100개까지의 태그들을 고려한 것이다.

이렇게 이미지 P에 대한 각 태그들의 연관성점수를 구한 후 점수가 높은 순으로 태그들을 정렬하여 저장하게 된다. 이 때, 점수가 높은 태그가 다른 태그들과 의미가 많이 중복되어 그 중요성이 높으므로 이미지를 잘 표현하는 태그가 된다고 보는 것이다. 실험에서는 연관성점수가 높은 순으로 1stTag, 2ndTag, ..., 5thTag 등, 상위 5개의 우선 태그들을 따로 분류하여 이들의 태그연관성점수를 저장하였다. 사전 실험을 통해 여섯 번째부터는 해당 이미지와 거리가 먼 태그로 판단하였기 때문에 다섯 번째까지만 활용하고자 한다. 그림 2는 본 시스템의 전체적인 구성을 순서도로 나타낸 것이다.

```
Length = length(TagSet(P));
for( i = 0 ; i < Length ; i++) {
      TagVal(Ti) = 0;
      for( j = 0 ; j < Length ; j++) {
            MeanSet(Tj) = WordNet(Tj);
            if(Ti ∈ MeanSet(Tj))
                  TagVal(Ti) = a+(1-i/100)*a;
      }
}
```

그림 1. 태그연관성점수 계산 알고리즘

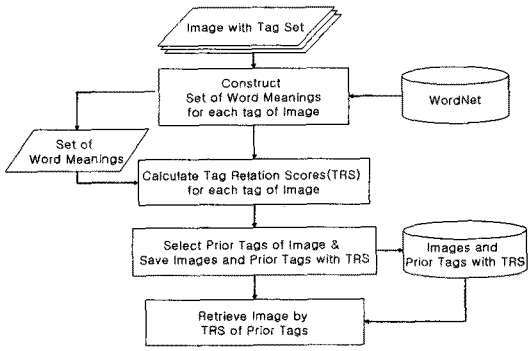


그림 2. 시스템 구성도

3.3 우선 태그를 이용한 다단계 검색

앞 절에서 설명한 방법대로 각 이미지마다 태그연관성점수가 높은 순으로 상위 5개 우선 태그들과 그 태그연관성점수가 저장되었을 때, 검색은 다음과 같은 다단계 방법으로 이루어진다. 먼저 1차 검색으로, 입력된 검색어가 우선 태그들 중 1stTag와 일치하는지 살펴 일치하는 이미지들을 그 태그연관성점수, TagVal(1stTag)가 높은 순으로 배치한다. 이를 검색 쿼리문으로 나타내면 "SELECT * FROM $db WHERE ('1stTag' like '$keyword') ORDER BY 'TagVal(1stTag)' DESC"와 같다. 제4장에서 설명할 실험에서는 검색어를 'house'로 하여 테스트하는데, 결과화면의 첫 번째 페이지에 처음으로 검색된 이미지의 우선 태그들은 {house, reflections, tower, clouds, building}으로 나왔다. 이 이미지가 첫 번째로 검색된 것은 검색어인 'house'가 1stTag이면서 그 이미지의 TagVal('house')가 'house'를 1stTag로 가지는 이미지들 중 가장 큰 값이기 때문이다.

다음으로 1차 검색 결과에 포함되지 않으면서 검색어가 2ndTag와 일치하는 이미지들을 역시 태그연관성점수 TagVal(2ndTag)가 높은 순으로 배치하고, 1차 검색결과에 이어서 결과화면에 나열한다. 이를 검색 쿼리문으로 나타내면 "SELECT * FROM $db WHERE ('1stTag' not like '$keyword' AND '2ndTag' like '$keyword') ORDER BY 'TagVal (2ndTag)' DESC"와 같으며, 실험에서 2차 검색의 첫 번째 이미지로 검색 된 것은 10페이지 19번째, 즉 전체 259번째 이미지인데 이것의 우선 태그들은 {condominium, house, vancouver, ocean, sea}로 나타났다. 검색어인 'house'가 2ndTag인 것을 알 수

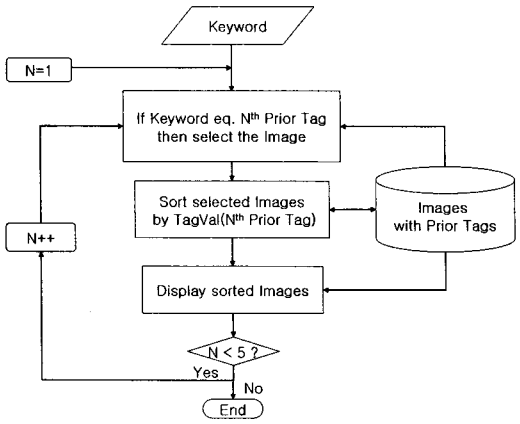


그림 3. 우선 태그를 이용한 다단계 검색

있다.

동일한 방법으로 5차 검색까지 수행하고, 마지막으로 6차 검색에서는 5차까지 검색되지 않은 자료들 중에서 검색어가 우선 태그가 아닌 나머지 태그들 중 하나와 일치하는 것들을 찾아낸다. 하지만 이와 같은 검색에서는 검색어와는 거리가 먼 자료들이 검색되어 실제 테스트에는 5차 검색까지만 수행하였다. 그림 3은 검색 알고리즘을 순서도로 표현한 것이다.

4. 실험 및 평가

4.1 실험 데이터의 구성과 검색 첫 화면 비교

실험을 위해 먼저 검색어 'house'를 이용하여 플리커에서 검색한 결과를 각 이미지와 그에 붙은 태그들로 데이터베이스에 저장하였다. 또한 실시간으로 변하는 플리커에서, 이후 실험에 사용한 것과 똑같은 데이터를 사용한 검색 결과를 그대로 보존하기 위해 플리커의 모든 결과 페이지는 화면을 그대로 캡처하여 저장하였다. 한 페이지당 24개씩, 모두 50개 페이지의 1,200개 이미지를 데이터베이스에 저장한 후, 제3장의 표 1에서 설명한 검색어연관성 이용방법과 태그연관성 이용방법으로 각각의 검색을 수행하였다. 시스템 구현을 위해 Linux환경에서 MySql Server 5.0와 Apache Server 2.2를 사용하였으며 PHP 5.2.4를 활용하여 웹프로그래밍으로 구현하였다.

검색 결과, 그림 4, 그림 5, 그림 6에서 보는 것과

그림 4. 플리커 검색 결과

같이 플리커에서는 첫 번째 화면의 24개 이미지 중 정확한 자료가 7개였으나 검색어연관성 이용방법과 태그연관성 이용방법에서는 각각 19개, 18개로 나타나 원하는 결과를 앞 페이지에서 바로 찾고자 하는 사용자의 만족도를 월등히 향상시켜주고 있다. 검색어연관성 이용방법에서는 그 알고리즘에서 알 수 있듯이 검색어 'house' 의미단어집합에 포함되는 가장 많은 수의 태그를 가진 이미지가 가장 먼저 검색될 것이다. 실험에서는 이에 해당하는 첫 번째 이미지의 태그 수가 5개였으며, 내용은 {building, dwelling, home, house, structure}였다. 태그연관성 이용방법에서도 같은 이미지가 첫 번째로 검색되었으며 그 이미지의 우선 태그들은 {house, reflections, tower, clouds, building}으로 구해졌다.

한편, 부적합한 검색결과의 예로서 플리커 검색 첫 페이지의 4번째로 나온 자료인 'Noble Jake'이란 제목의 사진은 'house' 포함하여 'dog', 'grass', 'lawn', 'backyard'.. 등 27개의 태그를 가지고 있으나 이중 검색어인 'house'의 의미단어집합에 포함되는 태그는 'house' 1개뿐이었다. 이 사진은 검정색 개를 찍은 사진에서 배경으로 약간의 건물이 보이는 것으

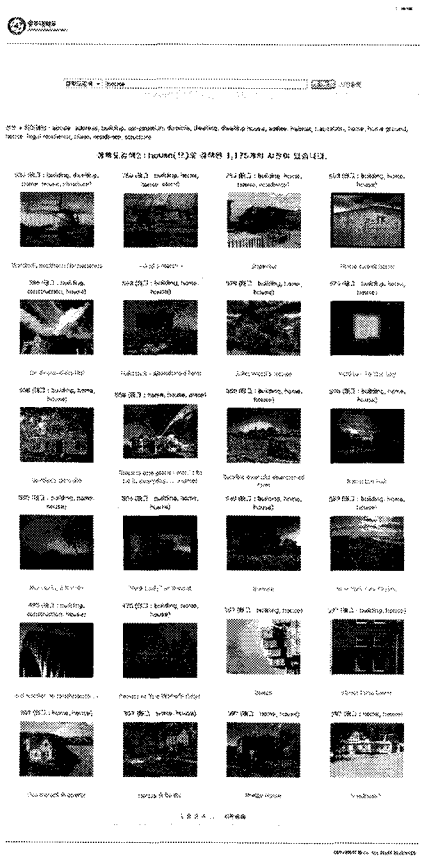


그림 5. 검색어연관성 이용방법 검색 결과

그림 6. 태그연관성 이용방법 검색 결과

로 검색어인 'house'의 검색 결과로는 부적합하다. 플리커에서는 태그 중 검색어인 'house' 있으면 최근 업로드된 순으로 보여주어 첫 페이지에 나타났으나 검색어연관성 이용방법에서는 40페이지(총50페이지 중)의 첫 번째 즉, 937번째로 출력되었으며, 태그연관성 이용방법에서는 38페이지 17번째 즉, 929번째로 검색 결과에 출력되었다. 검색어연관성 이용방법에서는 단 하나의 태그만 검색어와 일치하여 검색어연관성 점수가 낮았으며, 태그연관성 이용방법에서는 'house' 아닌 다른 태그, 예를 들면 'dog'와 같은 태그가 태그연관성 점수가 높아 우선 태그로 되고 'house'는 우선 태그에 들지 못하게 된다. 따라서 5차까지의 검색 중 태그 'house'가 검색어와 일치하지 않아서 앞 페이지를 차지하지 못하게 된다.

4.2 제안된 시스템의 성능 평가

본 연구에서의 시도를 평가하기 위해 50개 페이지 중 25개 페이지에 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. 또한 'house'는 그 의미단어집합에서 알 수 있듯이 'home'과 동의어로도 사용될 수 있으므로 집으로 보이는 건축물의 내부, 즉 거실, 침실, 주방과 같은 공간을 나타내면 역시 '적합'으로 판정하였다. 다만, 앞서 'Noble Jake'의 예에서 보았듯이 집으로 보이는 건축물이 배경의 일부이면 '부적합'으로 판정하였다. 플리커, 검색어연관성 이용방법, 태그연관성 이용방법 각각의 검색 결과를 비교한 결과는 표 2와 같다. 쉽게 예상할 수 있듯이 플리커의 적합한 이미지 비율은 검색된 결과 페이지들의 순위에 상관없이 대략 30%에서 50% 정도로 고르게 나타난 반면, 워드넷 기반의 의미정보를 이용한 방법들은 앞쪽에 위치한 페이지들이 적합한 이미지들을 많이 포함하고 있

표 2. 검색 결과 비교표

구분 / 페이지	플리커				검색어연관성 이용				태그연관성 이용			
	적합	부적합	합계	비율(%)	적합	부적합	합계	비율(%)	적합	부적합	합계	비율(%)
1	7	17	24	29.17	19	5	24	79.17	18	6	24	75.00
2	9	15	24	37.50	17	7	24	70.83	16	8	24	66.67
3	10	14	24	41.67	16	8	24	66.67	13	11	24	54.17
4	11	13	24	45.83	10	14	24	41.67	13	11	24	54.17
5	5	19	24	20.83	15	9	24	62.50	8	16	24	33.33
6	8	16	24	33.33	6	18	24	25.00	13	11	24	54.17
7	7	17	24	29.17	12	12	24	50.00	13	11	24	54.17
8	12	12	24	50.00	9	15	24	37.50	16	8	24	66.67
중 략												
23	10	14	24	41.67	6	18	24	25.00	16	8	24	66.67
24	12	12	24	50.00	6	18	24	25.00	13	11	24	54.17
25	9	15	24	37.50	13	11	24	54.17	11	13	24	45.83
합계	237	363	600	39.50	273	327	600	45.50	329	271	600	54.83

표 3. 검색 결과 정확성 및 재현율 비교표

구분 / 페이지	플리커					검색어연관성 이용					태그연관성 이용				
	적합	부적합	합계	정확성	재현율	적합	부적합	합계	정확성	재현율	적합	부적합	합계	정확성	재현율
TOP 4	37	59	96	0.39	0.08	62	34	96	0.65	0.14	60	36	96	0.63	0.13
TOP 8	69	123	192	0.36	0.15	104	88	192	0.54	0.23	110	82	192	0.57	0.24
TOP 12	103	185	288	0.36	0.23	149	139	288	0.52	0.33	158	130	288	0.55	0.35
TOP 16	144	240	384	0.38	0.32	193	191	384	0.50	0.43	208	176	384	0.54	0.46
TOP 20	183	297	480	0.38	0.41	234	246	480	0.49	0.52	262	218	480	0.55	0.58

어 처음 한, 두 페이지에서 적합 비율이 월등히 높게 나타났다. 따라서 처음 한, 두 페이지에서 원하는 결과를 빨리 찾고자하는 사용자의 요구를 보다 빠르고 정확하게 충족시킬 수 있다.

3가지 검색의 성능을 보다 정확히 비교 평가하기 위해 총 25개 페이지 중에서 각각 첫 4개, 8개, 12개, 16개, 20개 페이지에 속한 적합한 데이터의 비율을 조사하여 정확성(precision)과 재현율(recall)을 계산하였다. 정확성은 해당 페이지들에서 검색된 전체 이미지 수에 대한 적합한 이미지의 수로 계산하고, 재현율은 전체 50개 페이지에서 검색된 적합한 이미지 수(총 451개)에 대한 해당 페이지들의 적합한 이미지 수로 계산한다.

먼저 정확도를 살펴보면, 그림 7에 나타난 것과 같이 첫 4개 페이지에서 플리커의 0.39에 비해 검색어연관성 이용방법이 0.65, 태그연관성 이용방법이 0.63의 결과를 보여 플리커보다 워드넷기반의 의미정보를 이용한 방법들이 정확도에서 월등히 향상된 결과를 나타내었다. 특히 본 논문에서 새로이 시도된 태그연관성 이용방법은 상위 8, 12, 16, 20개 페이지에서 기존의 검색어연관성 이용방법보다 더 높은 정확도를 보였다. 재현율의 비교에서도 그림 8에서와 같이 태그연관성 이용방법이 가장 우수하게 나타났

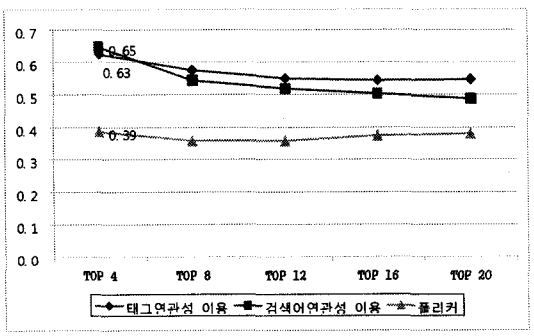


그림 7. 정확성 비교 그래프

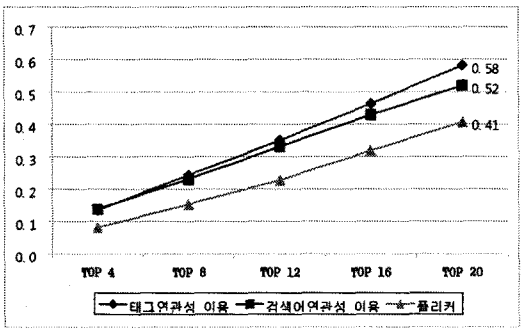


그림 8. 재현율 비교 그래프

으며 전반적으로 플리커에 비해 워드넷기반의 의미정보를 이용한 방법들이 우수한 것으로 드러났다. 재현율은 50개 페이지, 1,200개 자료 중 적합한 이미지 451개에 대한 비율이므로 페이지 누적수가 많을수록 증가하는 그래프로 그려진다. Top20을 보면, 50개 페이지 중 20개 페이지, 즉, 상위 40%에 해당하는 페이지의 재현율이 플리커 0.41, 검색어연관성 이용방법이 0.52, 태그연관성 이용방법이 0.58인 것을 알 수 있다.

태그기반의 웹 이미지 검색시스템에서 사용자들은 원하는 이미지들을 검색 결과의 앞쪽 페이지에서 발견하기를 선호한다. 즉, 찾고자 하는 이미지가 검색어인 태그와 공통의 의미를 지닌 이미지들로서 정확성을 만족하면서도, Top20 등 앞쪽 페이지에 정확한 이미지들을 더 많이 보여줄 수 있다면 사용자의 만족도를 높일 수 있을 것이다. 따라서 본 실험의 결과는 제안된 방법이 웹 이미지 검색시스템 사용자의 만족도를 높이는 데 크게 기여할 수 있음을 보여준다.

5. 결 론

본 논문에서는 태그기반의 웹 이미지 검색 시스템에서, 정확도 및 재현율의 향상으로 사용자 만족도를 높이면서도 검색 속도 면에서는 유리한 새로운 시도를 소개하였다. 그것은 이미지 태그들의 워드넷 의미정보를 이용한 태그연관성점수를 기반으로 해당 이미지에서 우선 태그들을 선별하고 이를 이용하여 태깅된 웹 이미지를 검색하는 방법이다.

제안된 방법은 복잡한 알고리즘을 사용하는 기계학습기반의 데이터 분류(classification) 또는 군집화(clustering) 방법에 비해 매우 빠르고 간편하게 적용될 수 있으므로 실용성이 높은 방법이라 할 수 있다. 그 성능은 플리커뿐만 아니라 워드넷 의미정보로 검색어연관성점수를 계산하여 검색에 적용한 기존의 방법[4]과 비교했을 때도 전체 정확도와 재현율 면에서 더 우수한 것으로 드러났다. 특히 본 논문에서 제안하는 방법은 우선 태그의 선별이 검색 시점이 아닌 데이터 업로드 시점에 이루어지므로 검색 시간에 전혀 영향을 미치지 않는다는 장점이 있다. 이는 이미지검색에서 검색 시점에 키워드의 워드넷 의미정보를 이용하는 여타의 연구들[4,15,16]과도 차별되는 장점이 된다. 이와 같은 장점은 제안된 방법으로 웹 이미지 검색시스템을 구현하였을 때 검색의 정확도와 함께 검색의 속도 면에서 시스템의 성능을 향상시키는 데 기여할 수 있으므로 본 연구 결과의 활용성을 높여준다.

다만 제안된 방법은 워드넷을 이용한 의미단어 집합의 구성과 연관성점수의 계산을 기반으로 하기 때문에 워드넷의 한계가 바로 본 시스템의 한계로 작용한다는 것이 단점이다. 즉, 워드넷 자체가 정확하고 풍부한 정보를 포함하고 있지 못하면 시스템의 기능이 제한적일 수밖에 없다. 만약, 태그기반의 데이터 검색을 위해 워드넷을 대신할만한 대용량의 의미정보 집합을 자동 생성하여 이용할 수 있다면 본 연구의 시도는 더욱 높은 효과를 보일 수 있을 것이다. 따라서 향후 연구과제로는 방대한 온라인 어휘 사전인 위키피디아(Wikipedia)를 활용하여 문서들이 속하는 카테고리와 그 카테고리 트리를 뽑아내어 페이지 내용과 트리의 경로를 이용하여 유사성(similarity)을 계산하는 알고리즘을 연구할 계획이다.

참 고 문 헌

[1] 홍성태, 임일, "웹 2.0 환경에서 정보 분류와 필터링, 그리고 협업을 위한 기술의 동향 및 발전 방향," Telecommunications Review, 제17권, 제4호, 2007.

[2] http://www.flickr.com/

[3] "WordNet, a lexical database for the English language," http://wordnet.princeton.edu/

[4] 권대현, 홍준혁, 조수선, "태그기반의 웹 이미지

검색에서 워드넷을 이용한 검색 순위 조정," 한국멀티미디어학회 추계학술발표대회논문집, pp. 685-688, 2008. 11.

[5] 이강표, 김두남, 김형주, "웹 2.0 환경에서의 태깅기술 동향," 한국정보과학회지, 제25권, 제10호, 2007.

[6] 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.

[7] A. Hotho, et al., "Information Retrieval in Folksonomies: Search and Ranking," In Proc. of ESWC'06, 2006.

[8] S. Angeletou, et al., "Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report," In Proc. of Workshop: Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference, 2007.

[9] V.D. Celine, et al., "Folksontology: An integrated approach for turning folksonomies into ontologies," In Proc. of Workshop: Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference, 2007.

[10] 이시화, 이만형, 황대훈, "Web2.0 환경에서의 효율적인 이미지 검색을 위한 태그 클러스터링 시스템의 설계 및 구현," 멀티미디어학회 논문지, 제11권, 제8호, 2008.

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

[12] Philip Resnik, "Using information content to evaluate semantic similarity in a taxonomy," Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.

[13] C. Leacock, M. Chodorow, "WordNet: An Electronic Lexical Database," C Fellbaum (Ed.), Combining local context and WordNet similarity for word sense identification, The MIT Press, pp. 265-283, 1998.

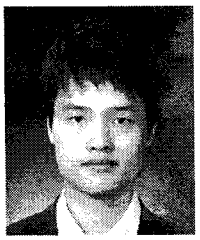
[14] 김형일, 김준태, "워드넷 기반 협동적 평가와 하이퍼링크를 이용한 검색엔진의 성능 향상," 정보처리학회논문지B, 제11-B권, 제3호, 2004.

[15] 최준호, 조미영, 김판구, "컬러 분포와 WordNet 상의 유사도 측정을 이용한 의미적 이미지 검색," 정보처리학회논문지B, 제11-B권, 제4호, 2004.

[16] Y. Alp Aslandogan. et al., "Using Semantic contents and WordNet in image retrieval," In Proc. of ACM SIGIR97, Philadelphia PA, USA, pp. 286-295, 1997.

[17] G. Varelas, et al., "Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web," In Proc. of the 7th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany, pp. 10-16, 2005.

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

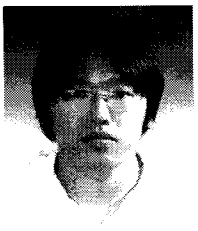


권 대 현

2009년 2월 충주대학교 컴퓨터과학과 공학사

2009년 3월~현재 충주대학교 산업대학원 전자계산학과 석사과정

관심분야 : 정보검색, 데이터베이스, 센서 데이터처리



홍 준 혁

2009년 2월 충주대학교 컴퓨터과학과 공학사

2009년 3월~현재 충주대학교 산업대학원 전자계산학과 석사과정

관심분야 : 정보검색, 임베디드시스템

조 수 선

1987년 서울대학교 계산통계학과 이학사
1989년 서울대학교 대학원 계산통계학과 이학석사
2004년 충남대학교 대학원 컴퓨터과학과 이학박사
1989년~1994년 (주)웅진미디어 연구원
1994년~2004년 한국전자통신연구원 선임연구원
2004년~현재 충주대학교 컴퓨터과학과 조교수
2006년~2007년 미시간대학교(앤아버) 통계학과 방문연구원
관심분야 : 데이터마이닝, 웹이미지 검색, 영상처리