제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

위키피디아 의미정보를 이용한 태깅된 웹 이미지 검색

이성재, 조수선
충주대학교 컴퓨터정보공학과
e-mail:lnew1004@gmail.com

Tagged Web Image Retrieval with Wikipedia Semantic Information

Sungjae Lee, Soosun Cho
Dept. of Computer Science & Information Engineering,
Chungju National University.

요 약

오늘날, 웹 공간에서는 사진과 같은 멀티미디어 자료를 공유하기 위하여 다양한 방법으로 문서의 정보를 표현하고 있다. 이러한 정보를 이용하기 위해 제목, 내용등에서 형태소 분석을 통해 의미가 있는 단어들을 이용하는 경우도 있지만 그 문서 혹은 자료와 관련있는 태그를 기입하고 활용하는 것이 보편화 되어 있다. 본 연구에서는 위키피디아 문서를 이용하여 이미지 태그들 사이의 연관성을 활용하여 이미지 검색 순위를 조정하였다. 약 1000만건의 문서로 이루어진 위키피디아를 이용하여 태그들의 연관성을 계산하였으며, 실험결과 태그 기반의 이미지를 검색 할 때 향상된 결과를 얻을 수 있었다.

1. 서론

이미지나 멀티미디어 자료를 웹문서로 표현함에 있어서 사용되는 폭소노미(folksonomy)는 사용자들이 협력태깅으로 생성하고 관리하는 웹 2.0을 이끌고 있는 대표적인 요소이다. 이러한 폭소노미는 쉽고 빠르며 자유롭게 이용할 수 있는 장점이 있지만 사람들이 기록하는 태그들은 문서 혹은 이미지의 의미를 표현하기 보다는 별다른 규칙 없이 의미 없는 태그를 기록하는 경우가 대부분이다. 이러한 태그를 이용한 검색은 원하는 이미지와 매우 상이한 이미지를 종종 출력하게 된다. 그렇기 때문에 본 연구에서는 이미지 검색의 효율을 높이기 위해 위키피디아[1] 기반의 의미 연관성 검색을 이용, 플리커[2] 이미지의 대표성을 가지는 태그를 추출하고 순위를 결정하고자 한다.

제안하는 웹 이미지의 검색순위 조정을 위한 위키피디아 의미 연관성 이용방법이 얼마나 효과적인지 확인하기 위해 플리커 이미지를 대상으로 실험하고 검색결과를 평가하였다. 본 논문에서는 2절에서 관련연구를 설명하고 3절에서는 위키피디아 기반의 태그 의미 연관성 계산 방법을 설명하며 4절에서는 실험 및 평가를 다룬다. 5절에서는 결론을 맺는다.

2. 관련연구

폭소노미를 이용한 웹 문서의 메타데이터 생성은 유연하고 다양한 분류체계를 제공할 수 있지만 근본적인 한계도 가지고 있다. 한 가지 예를 들면 입력한 태그들 간의 연관성을 찾아낼 수 없기 때문에 많은 문제점들이 나타나게 된다. 이러한 문제점을 해결하기 위해 사용자들이 넣은 태그의 연관성을 계산하고자 워드넷(WordNet)[3]을 이용한 동의어(synonym) 및 상위어(hypernym) 집합을 이용하여 태그들 간의 연관성을 계산, 우선 태그들을 추출하여 사용자가 요구하는 검색어에 대응하는 문서의 순위를 결정하는 연구[4]가 있었다.

본 논문에서 사용할 위키피디아는 약 1000만 건의 문서로 이루어진 웹 백과사전 서비스이다. 위키피디아의 자료는 전 세계 누구든지 접근, 수정이 자유롭지만 데이터 품질 관리를 위해 수정된 내용을 검토하여 적용되는 방식으로 서비스가 이루어지고 있다. 이렇기 때문에 위키피디아의 방대한 데이터베이스를 이용하여 다양한 분야에서 다수의 연구가 진행되어 왔다. 위키피디아를 이용한 어휘들 사이의 연관성을 계산하고자 했던 시도가 있었고 이들은 과거의 연구에서 적용하였던 path-length measure를 위키피디아에 맞게 변형하여 사용하였다[5]. 또한 위키피디아 각 문서에 포함되어 있는 모든 텀들로 가중치 벡터를 구성하여 두 문서 사이의 유사성을 해당 벡터들을 비교함으로써 계산하기도 하였다[6]. 하지만 이와 같은 초기 연구들 보다 효과적인 연구는 Milne & Witten의 연구[7]에서 발표 되었다. 이들은 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 문서 간 유사성을 측정하는 것처럼 위키

피디아의 문서들을 링크 가중치로 구성된 벡터로 보고 이들간의 유사성을 두 벡터 사이의 코사인 값으로 계산한 것이다.

최근 위키피디아를 이용하여 어휘의 의미를 추출하고자 하는 국내 연구[8,9]도 찾을 수 있다. 연구[8]은 태그들간의 상·하위 관계를 산출하기 위해 위키피디아 본문을 이용하고 연구[9]는 한국어 위키피디아를 기반으로, 검색질의어의 중의성을 해소하려는 시도이다. 본 논문에서는 연구[7]에서 발표된 링크 가중치 방법을 적용하여 위키피디아의 두 문서 사이의 유사성을 두 문서의 제목들 사이의 연관성으로 보고 이를 이용하여 검색어와 태그 사이의 의미 연관성을 계산한다. 이는 위키피디아 본문이나 텍스트 텀들을 그대로 이용하는 방법들에 비해 계산이 빠르고 효과적이기 때문이다.

3. 위키피디아 기반의 태그 의미 연관성 계산

위키피디아는 문서와 문서간의 링크로 이루어져 있다. 이러한 링크를 가지고 이를 벡터로 표현하고 문서들 간의 유사성을 계산하는 방식은 매우 효과적이다. 이는 연구[7]에서 밝혔듯 문서들의 링크 가중치로 구성된 벡터를 가지고 두 벡터 사이의 코사인 값으로 계산한다. 예를 들어 하나의 문서 s로부터 연결되어있는 문서 t로 아웃링크가 있을 때 이 링크의 가중치 w는 다음 식과 같이 계산된다.

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

여기서 T는 문서 t를 가리키는 모든 문서들의 집합이고 W는 위키피디아 전체 문서들의 집합이다. 위의 식에서 보듯 링크 가중치 w는 위키피디아 모든 문서 집합에서 타겟 문서를 링크하는 문서들이 차지하는 비율을 역수로 뒤집어 놓았기 때문에 문서를 가리키는 비율이 높을수록 해당 소스 문서 내에서 그 링크의 중요도는 떨어지게 된다.

링크 가중치를 이용하여 위키피디아 문서의 벡터를 구성한 후 이를 이용하여 사용자의 검색어와 검색대상의 태그들 간의 연관성을 계산하기 위해서는 아래의 식과 같이 두 벡터 검색어 S와 태그 T의 코사인 유사도(cosine similarity)가 이용된다.

$$similarity = \cos(\theta) = \frac{S \cdot T}{\|S\| \|T\|} = \frac{\sum_{i=1}^{n} S_i \times T_i}{\sqrt{\sum_{i=1}^{n} (S_i)^2} \times \sqrt{\sum_{i=1}^{n} (T_i)^2}}$$

약 1000만개의 위키피디아 전체문서에 대한 벡터가 구성되야 하지만 그와 같은 방법으로 실제 계산은 매우 어렵고 효율적이지 못하다. 그래서 본 연구에서는 전체 문서에 알파벳 순서의 고유번호를 부여하고 링크 가중치와 함께 해당 링크번호를 같이 저장하였다. 검색어와 태그에 부여된 벡터간 코사인 유사도는 고유번호가 같은 링크 가중치를 곱하여 합산함으로써 빠르고 효율적으로 계산하게 된다. 이렇게 계산하여 합산된 결과는 검색어와 태그간의 의미 연관성 점수로 이용되게 된다.

의미 연관성을 이용한 이미지의 검색에서는 검색어와 의미 연관성 점수가 높은 순으로 태그들을 정렬하고 이 순위에 기반하여 순차적으로 검색하는 방법을 적용한다. 즉, 우선순위가 높은 5개의 태그들을 이미지마다 선별하고 이들을 검색에 활용하는 것이다.

4. 실험 및 평가

위키피디아 기반의 연관성 점수를 계산하기 위해 본 논문에서는 영문 위키피디아 2011년 1월 15일자 데이터를 사용하였다. 이 문서는 pages-articles 타입의 약 30GB XML[10] 단일파일로 이루어져 있다. 해당 XML파일은 영문 위키피디아의 모든 페이지의 데이터를 가지고 있으며 문서간의 연결 관계나 특수한 설정들도 기록되어 있다. 하지만 이러한 XML 데이터를 파싱하기에는 너무 큰 용량을 가지고 있어서 JAVA기반의 StAX 라이브러리를 사용하여 MY-SQL로 변환하여 파싱하였다. 이렇게 수집된 문서(article)의 수는 10,861,570(중복포함) 건이며 의미 있는 아웃링크(outlink)의 수는 75,261,480(중복포함)건 이였다.

또한 실험에서 사용할 플리커 이미지를 사용하기 위하여 플리커 API[11]를 사용하였다. 플리커에서 추출한 이미지는 검색어 'house'를 이용하여 1500개의 이미지를 추출하였다. 이 1500개 이미지의 총 태그의 수는 153,993개였으며 위키피디아 적용 알고리즘을 이용하여 검색어와 태그간의 연관성을 계산하였다. 또 기존 연구[4]와의 비교를 위해 워드넷을 이용한 연관성 계산 알고리즘도 적용하여 실험하였다. 실험을 위해 사용한 시스템은 다음 <표 1>과 <표 2>와 같다.

<표 1> 서버용 시스템 개발 환경

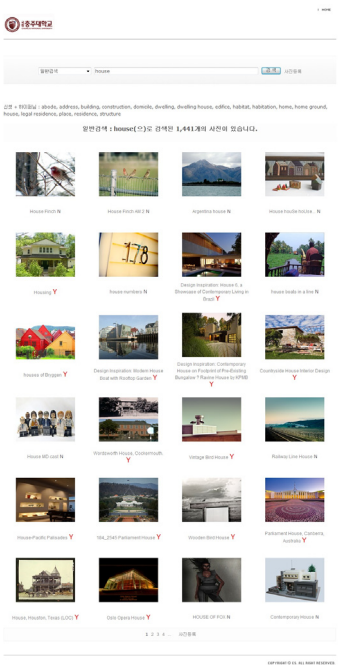
종류	구성	사양
H/W	CPU	제온 E5506 * 2
	RAM	DDR3 4GB ECC * 4
	HDD	1TB * 2 (RAID 0 구성)
S/W	OS	CentOS 5.6
	DBMS	My-SQL 5
	Web Server	Apache Server 2.2
	Programming Language	PHP 5.2.4

<표 2> 데이터 수집용 시스템 개발 환경

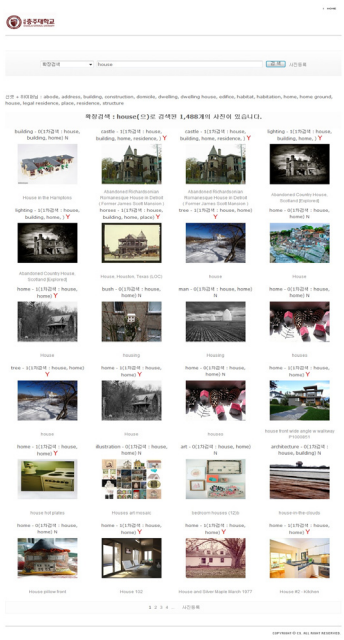
종류	구성	사양
H/W	CPU	코어2쿼드 Q9400
	RAM	DDR2 4GB
	HDD	500GB
S/W	OS	Windows 7
	DBMS	My-SQL 5
	Programming Language	JAVA, PHP
Data	Wikipedia	Database backup dumps (2011. 01. 15)
	Flikr	'house'와 관련된 상위 1500개의 검색결과

성능평가를 위해 1500개 플리커 이미지를 대상으로 한 각각의 검색 결과는 그림 1, 2, 3과 같다. 플리커에서 수집된 1500개의 'house'의 이미지를 건축물의 외형으로 판단하고 외형물에 해당하면 '적합'을 그렇지 않으면 '부적합'으로 판정하였다. 그림에서 보이는 것처럼 플리커 검색과 워드넷 기반의 연관성 검색에서 각각 14개 이미지가 적합으로 판정된 반면, 본 논문에서 제안하는 위키피디아 기반의 연관성 검색에서는 24개 모두 적합한 이미지로 판정되었다.

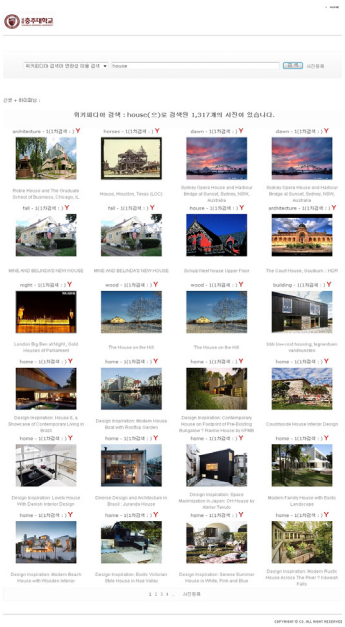
(그림 1) 플리커 검색 첫 페이지 결과



(그림 2) 워드넷 검색어 연관성 이용 검색 첫 페이지 결과



(그림 3) 위키 검색어 연관성 이용 검색 첫 페이지 결과



<표 3>은 각 검색엔진의 결과 상위 20개 페이지의 적합 비율을 보여준다. 상위 20페이지에서(페이지당 24개 이미지) 나오는 약 480개의 이미지 검색결과를 가지고 적합율을 측정해본 결과 기존의 플리커와 워드넷 검색어 연관성을 이용하였을 때에는 각각 56%, 58%의 적합성을 보였으나 본 논문서에 제안한 방법으로는 84%의 적합성을 보여줘 기존의 연구보다 월등한 적합비율을 보여주고 있다. 또한 사용자들은 원하는 결과를 앞페이지에서 바로 찾고자 하는데 본 연구에서 적용한 방법은 첫 두페이지에서 100%의 적합비율을 보여주고 있어서 사용자 만족도를 월등히 향상시키고 있다.

<표 3> 검색 결과 비교표

구분 / 페이지	플리커 검색		워드넷 검색어 연관성 이용		위키 검색어 연관성 이용	
	적합	적합 비율	적합	적합 비율	적합	적합 비율
1	14	0.58	14	0.58	24	**1.00**
2	16	0.67	21	0.88	24	**1.00**
3	19	0.79	19	0.79	20	0.83
4	17	0.71	19	0.79	20	0.83
5	13	0.54	8	0.33	17	0.71
6	15	0.63	16	0.67	18	0.75
7	17	0.71	13	0.54	13	0.54
8	15	0.63	14	0.58	18	0.75
9	14	0.58	11	0.46	19	0.79
10	7	0.29	15	0.63	20	0.83
11	13	0.54	17	0.71	19	0.79
12	13	0.54	11	0.46	20	0.83

13	13	0.54	11	0.46	23	0.96
14	10	0.42	11	0.46	24	1.00
15	10	0.42	13	0.54	22	0.92
16	12	0.50	10	0.42	20	0.83
17	12	0.50	16	0.67	19	0.79
18	13	0.54	12	0.50	24	1.00
19	10	0.42	18	0.75	20	0.83
20	15	0.63	10	0.42	21	0.88
전체 적합비율	268	0.56	279	0.58	405	**0.84**

또한 <표 4>는 3가지의 검색의 성능을 정확히 평가하기 위해서 첫 4개, 8개, 12개, 16개, 20개 페이지의 정확도(precision)와 재현율(recall)을 비교한 표이다. 정확도는 해당 페이지의 이미지 수에 대한 적합 이미지수로 계산하고 재현율은 전체 1500개 이미지의 적합한 이미지의 수(총 933개)에 대한 적합 이미지 수로 계산한다. 위키피디아 기반의 검색어 연관성 이용방법에서 정확도와 재현율이 각각 84.4%, 43.4%로 나타나 다른 방법들에 비해 우월함을 보이고 있다.

<표 4> 검색 결과 정확도 및 재현율 비교표

	일반검색(플리커)			워드넷 검색어 연관성			위키 검색어 연관성		
	적합	정확도(%)	재현율(%)	적합	정확도(%)	재현율(%)	적합	정확도(%)	재현율(%)
TOP 4	66	68.8	7.1	73	76.0	7.8	88	91.7	9.4
TOP 8	126	65.6	13.5	124	64.6	13.3	154	80.2	16.5
TOP 12	173	60.1	18.5	178	61.8	19.1	232	80.6	24.9
TOP 16	218	56.8	23.4	223	58.1	23.9	321	83.6	34.4
TOP 20	268	55.8	28.7	279	58.1	29.9	405	**84.4**	**43.4**

태깅되어있는 웹 이미지 검색 시스템에서는 사용자가 원하는 이미지들을 검색된 앞페이지에서 찾는 것을 선호하는데 상위 20페이지 등 앞쪽 페이지에서 더 많이 검색어를 만족하는 이미지를 출력한다면 사용자 만족도를 향상시킬 수 있을 것이다. 따라서 본 연구에서 제안한 위키피디아 기반의 검색어 연관성을 이용하여 검색순위를 조정하는 것이 매우 효과적인 방법임을 알 수 있다.

5. 결론

본 논문에서 제안한 위키피디아 기반의 태깅된 이미지 웹 검색 시스템을 이용한다면, 기존의 연구[4]의 워드넷을 이용하는 것보다 더욱 효과적임을 입증하였다. 또한 단순히 위키피디아 문서를 연결해주는 링크정보만 가지고서 간편하게 코사인 유사도 값을 계산할 수 있었고, 이를 이용하여 정확도 및 재현율이 눈에 띄게 향상된 것을 알게 되었다.

제안된 방법은 기존의 복잡한 알고리즘을 사용하는 기계학습기반의 방법에 비해 매우 빠르고 간편하게 적용할 수 있으므로 실용성이 높은 방법이라 할 수 있다. 향후 연구에서는 위키피디아 DB를 활용한 다양한 검색 서비스에 적용해볼 예정이다.

참고문헌

[1] http://www.wikipedia.org

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 이강표, 김현우, 장충수, 김형주, “FolksoViz: Wikipedia 본문을 이용한 상하위 관계 기반 폭소노미 시각화 기법,” 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제14권 제4호, pp. 341-457, 2008.6.

[9] 김성호, 배상준, 고영중, “ 한국어 위키피디아 기반 질의어 중의성 해소 및 확장 자질 추출 기법,” 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제17권 제3호, pp. 141-204, 2011.3.

[10] http://dumps.wikimedia.org/

[11] http://www.flickr.com/services/api/