키워드 기반 색인을 이용한 웹 이미지 검색 모델

양재석*, 박정규*, 최영식**, 이긍해**
*한국항공대학교 컴퓨터공학과
**한국항공대학교 전자·정보통신·컴퓨터공학부
e-mail:hijaeseok@mail.hangkong.ac.kr

Web Image Retrieval Model using Keyword based Indexing

Jaeseok Yang*, Jeongkyu Park*, YoungSik Choi**, and Keung Hae Lee**
*Dep. of Computer Engineering, Hankuk Aviation University
**Schools of Electronics, Telecommunication,
and Computer Engineering, Hankuk Aviation University

요 약

대부분의 이미지 검색은 질의 키워드를 이용하여 이루어지기 때문에 웹으로부터 수집한 이미지에는 미리 주제와 연관된 적절한 색인어를 부여하는 것이 필요하다. 웹 문서의 키워드를 이용하는 방법은 이미지와 연관성이 높은 것으로 간주되는 주변 키워드에 대해 각각의 연관도를 계산하여 색인어를 선정하는 방법이다. 본 논문에서는 이미지 주변의 키워드를 이용하여 이미지를 인덱싱한 후 유저 피드백을 통해 정확도를 높이는 웹 이미지 검색 모델을 제안한다.

1. 서론

일반적으로 웹을 통한 정보의 검색은 자신이 원하는 내용에 가장 적합한 키워드로 구성된 질의어를 검색 시스템에 전달하고, 검색 시스템은 이를 이용하여 검색 기능을 수행한 후 그 결과를 사용자에게 제공함으로써 이루어진다. 이를 위해 웹에서 수집한 정보에 적절한 색인어를 부여하여 미리 인덱싱하고 저장하는 것이 필요하다. 웹을 통한 이미지의 검색 또한 유사한 과정으로 이루어진다. 하지만 일반적인 웹 문서의 인덱싱과는 달리 웹 이미지로부터는 직접적으로 색인어를 추출 할 수 없기 때문에, 이미지에 대해 적절한 색인어를 부여하는 것은 이미지 검색 시스템에 있어서 검색의 질을 판가름하는 중요한 과정이다.

웹의 이미지에 대해 적절한 색인어를 부여하는 연구는 두 방향으로 이루어지고 있다. 하나는 이미지의 내용을 분석하여 이미지가 나타내는 주제를 파악하고 그 주제와 연관된 키워드를 부여하는 방법이며, 다른 하나는 이미지 주변의 다양한 키워드로부터 색인어를 선정하는 방법이다. 이미지의 내용을 직접 분석하는 방법은 지금까지 활발히 연구가 이루어지면서 어느 정도 성과를 보이고 있다. 하지만 이미지의 인식 기술은 가능한 대상이 한정적일 뿐만 아니라 유명인이 아닌 '김 아무개'와 같은 특정인, '한라산'이나 '동해'와 같은 구체적인 대상의 인식은 거의 불가능하기 때문에 이미지 인식만으로 색인어를 부여하기에는 한계가 있다. 이미지 주변의 키워드를 활용하는 방법은 이러한 단점을 보완할 수 있으며 이미지 인식 기법에 비해 적은 비용을 소모하므로 이미지 색인어 선정에 있어서 매우 효과적인 기법이다.

본 논문에서는 이미지 주변의 다양한 키워드를 활용하여 색인어를 선정하고 유저 피드백을 통해 이미지와의 연관도를 보정하는 웹 이미지 검색모델을 제안한다.

※ 본 논문은 과학기술부 한국과학재단 지정 경기도 지역협력연구센터(RRC)인 한국항공대학교 인터넷정보검색연구센터의 지원에 의한 것임

2. 관련 연구

초기의 이미지 검색 엔진인 WebSeer[1]는 이미지 내용 기반의 인덱싱 기법과 더불어 이미지 주변의 구문을 분석하여 이미지를 설명 할 수 있는 키워드를 추가하는 방법을 시도하였다. 이미지가 포함된 웹 문서의 특정 위치에 존재하는 구문은 이미지에 대한 정보를 나타낼 확률이 높다고 가정하고, 이러한 특정 위치의 구문으로부터 키워드를 추출하여 인덱싱에 이용하였는데, 주로 이미지의 파일 이름과 URL, 캡션, 이미지 태그의 ALT 필드, 문서의 타이틀이 사용되었다. 이 구문에서 추출된 키워드들은 각각의 위치에 대해 미리 정해진 가중치를 이용하여 이미지와의 연관도를 부여 받았으며, 그 결과에 따라 이미지와 연관도가 높은 키워드를 이미지의 색인어로 선정하였다. 이를 테면 이미지의 설명이 될 수 있는 ALT 필드에는 문서의 타이틀보다 높은 가중치를 부여하여 ALT 필드에서 추출된 키워드가 상위 순위를 차지하는데 유리하게끔 하는 식이었다. 이러한 WebSeer의 방법은 이미지 주변의 키워드를 이용한 인덱싱 기법의 기본이 되었다. 이후 키워드의 출현 빈도와 위치를 고려하여 좀더 정확한 키워드를 선정하기 위해, 가중치를 적절히 조절하는 LSI(Latent Semantic Indexing)[2]를 도입한 연구가 이루어졌다[3]. 이 연구에서는 수집한 모든 웹의 이미지에 대해 각각 특정 구문의 키워드와 그 구문에 해당하는 가중치 및 빈도를 이용하여 벡터를 생성하고, 벡터로 이루어진 행렬의 SVD(Singular Value Decomposition)연산을 통해 이미지에 대한 각 키워드의 연관도를 조절함으로써 초기의 연구에 비해 발전된 모습을 보여주었다.

Shen 등은 [4]에서 새로운 방법에 대하여 설명하고 있다. 이 연구에서는 Lexical chain[5]의 개념을 도입한 Weight ChainNet 이라는 모델을 선보이고 있는데, 기존의 연구에서 선보였던 특정 구문에 대해 통계적인 가중치 부여를 이용한 연관도의 산출과는 조금 다른 방법을 사용했다. 웹 문서에서 획득할 수 있는 다양한 구문을 모두 이용하기보다 가장 연관성이 높을 것으로 판단된 4개의 특정 구문만을 연관 키워드의 선정에 사용하였고, 이는 각각 이미지 파일 이름과 이미지 태그의 ALT 영역, 웹 문서의 타이틀, 이미지 주변의 구문이었다. 이들은 각각 lexical chain으로 재구성되어 순서대로 높은 가중치를 부여 받아 이미지의 색인 키로 사용되었다. 사용자의 질의가 있을 경우 질의의 lexical chain과 4개의 lexical chain간의 유사도를 측정하고 Match Level 측정을 통해 가장 높은 근접도를 갖는 lexical chain 산출 한 후, 해당 이미지를 검색 결과로 반환하였다. 이 연구는 좋은 결과를 보여주기는 했지만 이미지 검색시마다 늘 사용자의 질의문과 lexical chain간의 유사도와 근접도를 측정해야 하는 단점이 존재했다.

이러한 기존 연구들은 색인어의 선정에 있어서 주로 특정 위치의 키워드에 대한 이미지와의 연관 확률만을 바탕으로 이루어지기 때문에, 실제로 선정된 색인어가 이미지의 주제와 직접적으로 연관성을 가지고 있는지는 보장할 수 없는 한계를 가지고 있었다. 다음 장에서는 웹 문서의 키워드를 이용하여 이미지의 색인어를 선정하고 유저 피드백을 통해 연관도를 보정하는 이미지 검색 모델에 대해 설명한다.

3. 키워드의 이미지 연관도 산출

전통적인 문서 검색 기법에서는 문서 내에 포함된 키워드에 대해 각각 단어빈도 가중치(TF)와 역문헌빈도 가중치(IDF)를 계산하여 단어의 중요도를 결정하고, 이를 바탕으로 색인어를 선정하였다. 이미지의 색인어는 이러한 TF-IDF 가중치 기법에 이미지와의 연관도를 반영하여 선정한다.

[3]에 제안된 연관도 산출 방법은 문서 내 특정 키워드에 대한 가중치와 이미지와의 거리에 따른 가중치를 바탕으로 산출한다. 표 1은 키워드가 추출되는 이미지 주변의 특정 위치와 특정 HTML 태그를 나타낸다. 각각의 항목에 대한 가중치(W)는 통계를 바탕으로 산출된 이미지와의 연관도를 나타낸다.

표 1. 특정 구문의 키워드에 대한 가중치

HTML 태그	설 명	가중치(W)
ALT field	이미지 태그의 ALT 필드	6.00
TITLE	웹 문서의 타이틀	5.00
H1	헤딩(heading)	4.00
H2	〃	3.60
H3	〃	3.35
H4	〃	2.40
H5	〃	2.30
H6	〃	2.20
B	굵은(bold) 스타일	3.00
EM	강조(emphasis)	2.70
I	이탤릭(italic) 스타일	2.70
STRONG	강조(string)	2.50
<No Tag>		1.00

한편, 이미지 주변의 구문 또한 이미지와의 거리에 따라 다음의 계산식을 이용하여 연관 가중치를 산출한다[3].

$W = \rho * e^{-2.0 * pos/dist}$

ρ : 5.0

pos : 이미지와의 상대적인 거리

dist : 고려되는 이미지 주변의 키워드 수

이미지 주변의 키워드는 ρ를 5.0으로 고정함으로써 ALT 필드에서 추출한 키워드나 타이틀의 키워드보다 약간 작은 가중치를 부여 받도록 조절한다.

TF-IDF 기법에서의 TF 가중치는 문서 내 등장 빈도를 바탕으로 키워드의 중요도를 강조한다. 문서 내에 자주 등장하는 주요 단어는 등장 빈도에 비례하여 이미지와 더 높은 연관도를 가질 확률이 높다. 이러한 TF함수는 소수의 키워드로 구성된 문서의 경우에도 효과적으로 적용하기 위해 대개 비례식보다 로그 형태의 분포 특성을 갖는다[6]. IDF 가중치는 문서 집합 내 등장 빈도를 바탕으로 키워드의 중요도를 경감시킨다. 임의의 문서에서 자주 등장한 키워드가 전체 문서 집합에서도 자주 등장한다면 오히려 중요도는 낮다고 볼 수 있으며, 이를 IDF 가중치를 통해 반영한다.

이러한 TF-IDF 기법은 예전부터 정보 검색에서 키워드의 중요도를 결정하기 위한 방법으로 사용되어 왔으며, 다양한 모델이 제안 되어졌다. 그 중 대표적인 2-포아송 모델은 확률 검색 모델과 포아송 분포 함수를 이론적 기반으로 하여 개발된 정보 검색 모델로써 어떠한 문서 집합에서도 타 검색 모델보다 뛰어난 성능을 지닌 것으로 판단되어 진다[7].

표 2. 2-포아송 모델의 TF-IDF 가중치

$$TF = \frac{tf}{k_1 \cdot (1-b) + b \cdot \frac{doc\ length}{average\ doc\ length} + tf}, \quad k_1 = 2,\ b = 0.75$$

$$IDF = \log \frac{N - df + 0.5}{df + 0.5}$$

문서에 포함된 각 키워드에 대해 TF와 IDF 가중치를 산출하고, 산출된 각각의 가중치에 이미지와의 연관도를 반영하여 키워드의 최종 가중치(IW)를 결정한다.

$$IW = W \times (TF \times IDF)$$

키워드의 최종 가중치는 이미지와의 연관도를 나타내며 높은 값을 가질수록 연관도가 높다고 할 수 있다. 이미지는 각 키워드에 대하여 높은 가중치순으로 색인어 리스트를 생성한다.

4. 사용자 피드백에 의한 연관도의 보정

이미지의 색인어를 선정함에 있어서 통계적인 확률만을 바탕으로 색인어를 결정하는 기법은 선정된 색인어가 실제로 이미지의 주제와 직접적인 연관성을 가지고 있는지 보장할 수 없는 한계를 가지고 있다. 이를 보완하기 위해 이미지 인식 기법을 병행하고 있기는 하지만, 아직까지 단순 인식의 한계를 벗어나지 못하고 있는 상태이다. 가장 정확한 색인은 사람이 직접하는 것이지만 이것은 매우 힘들고, 비용이 많이 소모되는 작업이기 때문에 비효율적이다. 때문에 우선 통계적인 확률을 바탕으로 색인어를 선정하고, 유저 피드백을 통해 이를 보정하는 방법은 색인어의 정확도를 높이는 좋은 대안이 될 수 있다.

그림 1은 우리의 이미지 검색 모델을 보여주고 있으며, 다음과 같이 4단계로 요약 할 수 있다.

(1) 이미지에 대하여 키워드를 이용한 색인어 리스트 생성

(2) 색인어 리스트를 이용한 첫 번째 검색 결과 산출

(3) 검색 결과 중 사용자가 선택한 이미지의 색인어 선택

(4) 선택된 색인어의 가중치 증가 및 차후 검색에 반영

즉, 검색된 결과에 대하여 사용자가 선택한 이미지는 검색에 사용된 색인어와 가장 연관이 높은 것으로 간주 할 수 있으며, 이를 가중치 보정에 이용하는 것이다. 이와 같은 과정을 거치게 되면, 검색이 반복 될수록 이미지에 대한 색인어의 가중치가 보정되어 연관도가 높은 이미지일수록 검색 결과의 상단에 위치 할 수 있으며, 사용자는 자신이 필요한 검색을 수행하고 원하는 이미지를 선택할 뿐, 가중치 보정을 위한 추가된 행동을 하지 않아도 된다.

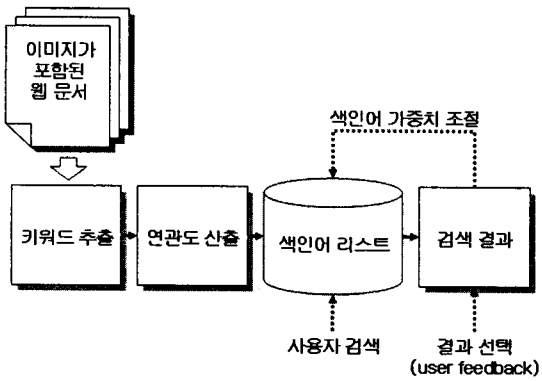

그림 1. 키워드를 이용한 이미지 검색

5. 결론

웹의 다양한 이미지로부터 사용자가 요구하는 이미지를 정확하게 검색해내기 위해서는 수집된 이미지를 분석하여 이미지의 주제를 파악하고 이를 적절한 색인어로 인덱싱 하는 것이 가장 바람직한 방법이다. 그러나 현재의 이미지 인식 기술은 아직까지 만족할 만한 수준에 도달하지 못하고 있기 때문에, 이미지 주변의 키워드를 이용하여 간접적으로 이미지를 대표하는 색인어를 선정하는 방법이 대안이 될 수 있다.

우리의 모델은 키워드를 이용한 이미지 색인어 선정 기법에 사용자의 피드백에 의한 연관 가중치 보정 기법을 도입하였다. 즉, 색인어 리스트를 사용한 검색 결과에서 사용자가 만족한 이미지에 대해 색인어의 가중치를 증가시킴으로써 색인어의 정확도를 증가 시킬 수 있다.

그러나 우리의 모델은 최초의 색인어 리스트에 어느 정도 정확한 색인어가 존재해야만 보정을 통해 색인어의 정확도를 높일 수 있다는 한계를 가지고 있으며 이는 차후 해결해야 할 문제점이다.

향후 우리의 연구는 실험을 통해 우리의 모델을 적용한 시스템과 기존 시스템을 비교 분석하고, 이를 바탕으로 우리 모델의 성능을 검증하고 좀더 발전 시켜나가는 방향으로 이루어질 것이다.

참고문헌

[1] C. Frankel, M. Swain, and V. Athitsos, "WebSeer: Am Image Search Engine for the World Wide Web", *University of Chicago Technical Report*, 96-14, 1996

[2] S. Deerwester, S. Dumais, T. Landouer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis", *JASIS*, 41(6):391-407, 1990

[3] M. La Cascia, S. Sethi, and S. Sclaroff, "Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web", *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, June, 1998

[4] Heng tao Shen, Ooi Beng Chin, and Kian-Lee Tan, "Giving Meanings to WWW Images", *the Proceedings of the 8th ACM International Conference on Multimedia (ACM MM 2000)*, 39-47, 2000

[5] J. Morris and G Hirst, "Lexical Cohesion Computed by Thesaural Relation and an Indicator of the Structure of Text", *Computational Linguistics*, vol. 17 no. 1 pp. 22-48, 1999

[6] Amit Singhal, John Choi, Donald Hindle, David Lewis, and Fernando Pereira, "AT&T at TREC-7", *In Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, 1999

[7] 김지승, 이준호, 이상호, "세 가지 정보 검색 모델의 성능 평가 및 분석", *한국정보과학회논문지*, vol. 28 no. 02 pp. 266-278, 2001