오픈 디렉토리 프로젝트를 이용한 문맥 광고*

이정현, 하종우, 박상현, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail : {jhbslpd, okcomputer, condols, yalphy}@korea.ac.kr

Using Open Directory Project to Contextual Advertising

Jung-Hyun Lee, JongWoo Ha, Sang-Hyun Park, SangKeun Lee
Division of Computer and Communication Engineering, Korea University

요 약

문맥 광고에서 웹 페이지의 내용과 의미적으로 연관된 광고를 매칭하기 위해, 최근 웹 페이지와 광고를 동일한 분류 트리에 분류하여 의미적으로 매칭하는 방법이 제안되었다. 그러나 이 방법에서 사용된 분류 트리 및 분류기를 작성하기 위해선 많은 시간과 노력이 필요하다. 따라서 이를 용이하게 하기 위하여, 본 논문에서는 오픈 디렉토리 프로젝트의 공개 데이터를 활용하여 웹 페이지와 광고의 주제 분류를 위한 분류 트리 및 분류기를 작성하는 기법을 제안한다. 또한 실험 결과를 통하여 제안한 기법이 문맥 광고에서 웹 페이지와 광고의 의미적 매칭의 높은 정확성을 보장하는 것을 입증한다.

1. 서론

문맥 광고란, 웹 사용자가 웹 페이지를 요청할 때, 웹 페이지의 내용과 연관된 광고를 함께 표시하는 온라인 광고의 한 형태이다. 웹 페이지의 내용과 광고의 연관성이 높을수록 광고의 효과는 커지며, 이를 위해 최근 연구[2]에서는 웹 페이지와 광고를 동일한 주제 분류 트리에 분류하여 의미적으로 매칭하는 방법이 제안되었다. 여기서 웹 페이지와 광고의 주제는 매우 다양하기 때문에, 분류 트리는 웹 페이지와 광고의 모든 주제를 포함할 수 있는 매우 포괄적인 주제 분류 트리가 되어야 한다. 또한, 웹 페이지와 광고의 분류 정확도는 웹 페이지와 광고의 매칭 정확도에 직접적으로 영향을 미치기 때문에, 높은 분류 정확도를 가지는 웹 페이지와 광고의 분류기를 생성하는 것이 필수적이다.

이를 위해, 기존 연구[2]에서는 Yahoo! US 에서 다수의 사람들에 의해 직접 만들어진 분류 트리를 사용하였고, 각 주제마다 사람들이 연관된 광고 키워드를 수집하여 만들어진 학습 데이터를 분류기 생성에 사용하였다. 여기서 분류 트리 및 분류기 작성에 소요되는 사람의 시간과 노력이 매우 많고 효율적이지 않다. 본 논문에서는 이러한 사람의 노력을 줄이기 위한 방법으로, 오픈 디렉토리 프로젝트(ODP)[1]의 공개 데이터를 활용하여 웹 페이지와 광고의 주제 분류를 위한 분류 트리 및 분류기를 작성하는 기법을 제안한다.

2. ODP 에 기반한 분류 트리 및 분류기 작성 기법

ODP 란, 웹 사이트들을 주제별로 분류하기 위해 만들어진, 사람의 손에 의해 편집되는 가장 포괄적인 웹 디렉토리로서, 세계규모의 거대한 자원 편집 커뮤니티에 의해 구축, 관리되고 있다. ODP 에서 공개하는 데이터에는 웹 사이트를 분류하기 위한 디렉토리 구조와 각 디렉토리마다 사람에 의해 분류된 웹 사이트들의 제목과 설명, URL 이 있다. 디렉토리 구조는 트리 형태로 이루어져 있으며, 전체 디렉토리의 수는 약 59 만개, 최대 트리 깊이는 15 레벨, 분류된 전체 웹 사이트들의 개수는 약 460 만개이다. 최근 연구에서 ODP 의 데이터는 웹 페이지의 분류, 검색 쿼리의 분류 등에 활용되어왔다[4].

본 논문에서는 웹 페이지와 광고의 모든 주제를 포함할 수 있는 분류 트리를 작성하기 위하여, 먼저 ODP 의 전체 디렉토리 구조 중 일부만을 휴리스틱 룰에 의해 추출한다. 이때 사용된 휴리스틱 룰은 주제가 아닌 디렉 토리 제거, 적은 웹 사이트가 분류된 디렉토리 제거, 각 디렉토리 탐색 경로에서 하위 레벨 제거이다. 이를 통해 추출된 디렉토리의 수는 5,177 개, 최대 트리 깊이는 9 레벨이다. 추출된 각 디렉토리는 분류 트리의 하나의 노드로 간주되며 하나의 주제를 나타낸다. 각 노드는 ODP 내에서의 디렉토리 포함관계에 따라 'is-a 관계'로 연결되어, 계층적인 트리 구조를 갖게 한다.

작성된 분류 트리에 웹 페이지와 광고를 분류하기 위해 본 논문에서는 수정된 Rocchio 분류기를 사용한다[3]. 기존의 Rocchio 분류기는 벡터 스페이스 모델를 이용하여 각 주제별 학습 문서들의 중심 벡터들을 계산하고, 분류할 문서와 중심 벡터들간의 코사인 유사도 값을 계산하여, 가장 큰 중심 벡터를 가지는 주

* 이 연구에 참여한 연구자(의 일부)는 '2 단계 BK21 사업'의 지원비를 받았음.

제를 문서의 주제로 선택한다. 그러나, 이 분류기는 Flat 분류기로서, 분류 트리의 계층적인 특성을 반영하지 못하고, 단순히 분류 트리의 각 주제를 집합으로 간주한다. 따라서, 본 논문에서는 Rocchio 분류기에 계층적인 분류 트리의 특징을 반영할 수 있는 방법을 제안한다.

이를 위해 먼저, 각 주제들의 중심벡터를 구하기 위하여, 기존의 ODP 에 분류된 각 웹 사이트들의 제목과 설명을 하나의 학습 문서로 간주하고, 모든 ODP 의 디렉토리에 대한 중심 벡터를 계산하였다:

$$\overrightarrow{c_k} = \frac{1}{|D_k|}\sum_{u \in D_k}\frac{\vec{u}}{\|\vec{u}\|} \qquad (1)$$

(1)에서, $\overrightarrow{c_k}$는 k 번째 디렉토리의 중심 벡터이고, D_k는 그 디렉토리에 분류된 학습 문서들의 집합이며, $\vec{u}$는 한 학습 문서의 단어 벡터이다.

다음으로 분류 트리의 계층적 특징인 'is-a 관계'를 적용하기 위하여, 자식 디렉토리들의 중심 벡터들의 특징을 부모 디렉토리의 중심 벡터에 포함시켜, 부모 디렉토리의 병합 중심 벡터를 계산하였다:

$$\overrightarrow{m_k} = \frac{1}{1+|child(k)|}\left(\frac{\overrightarrow{c_k}}{\|\overrightarrow{c_k}\|} + \sum_{l \in child(k)}\frac{\overrightarrow{m_l}}{\|\overrightarrow{m_l}\|}\right) \qquad (2)$$

(2)에서, $\overrightarrow{m_k}$는 k 번째 디렉토리의 병합 중심 벡터이고, *child(k)* 는 그 디렉토리의 자식 디렉토리 집합이다. (2)는 재귀적 수식이며, 부모 디렉토리의 병합 중심 벡터를 구하기 위해서는, 자식 디렉토리의 병합 중심 벡터를 구해야 한다.

마지막으로, 분류할 웹 페이지와 광고에 대해 생성된 분류 트리의 각 노드들의 병합 중심 벡터들과의 코사인 유사도 값을 구하여, 가장 높은 값을 가지는 노드의 주제로 웹 페이지와 광고의 주제를 결정한다:

$$C_{\max} = \arg\max_{c_k \in C}\frac{\overrightarrow{m_k}}{\|\overrightarrow{m_k}\|}\cdot\frac{\vec{d}}{\|\vec{d}\|} \qquad (3)$$

(3) 에서 C 는 분류 트리의 전체 노드들의 집합이고, $\vec{d}$는 분류할 웹 페이지와 광고를 벡터 스페이스 모델로 표현한 단어 벡터이며, $\overrightarrow{m_k}$는 k 번째 노드의 병합 중심 벡터이다.

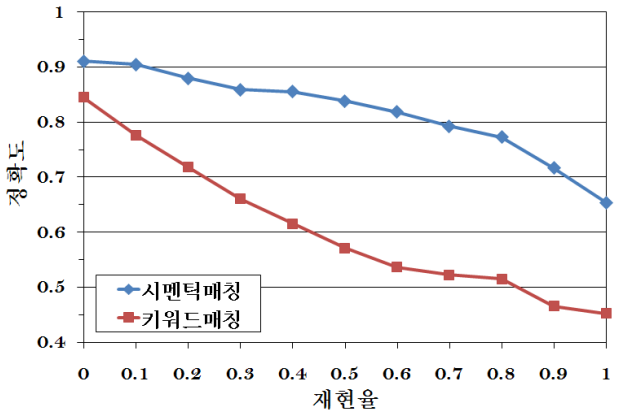
3. 성능 평가

본 논문에서는 먼저 작성된 분류기의 웹 페이지와 광고 분류 성능을 측정하였다. 표 1 은 실험에 사용된 평가 셋 및 웹 페이지와 광고의 분류 정확도를 나타낸다. 실험 결과 광고의 분류 정확도가 웹 페이지의 분류 정확도보다 약 5%정도 낮았다. 이는 광고의 텍스트가 웹 페이지에 비해 매우 적어서 분류에 필요한 충분한 정보를 가지고 있지 않았기 때문이다.

<표 1> 실험 평가 셋 및 분류 정확도

웹 페이지 수	131	페이지당 평균 광고 수	43.8
광고 수	5300	페이지 분류 정확도	83.2%
웹 페이지-광고 쌍	5745	광고 분류 정확도	78.9%

다음으로 작성된 분류 트리를 활용한 웹 페이지와 광고의 의미적 매칭 성능을 측정하였다. 이를 위해, 기존 연구[2]에서 제안된 시멘틱 매칭 방법에 작성된 분류 트리와 분류기를 사용하였고, 키워드 매칭 방법과 비교 평가 하였다. 여기서, 시멘틱 매칭 방법이란, 분류기를 통해 웹 페이지와 광고의 주제를 결정하고, 두 주제가 분류 트리 내에서 떨어진 정도를 웹 페이지와 광고의 유사도로서 측정하는 방법이다. 키워드 매칭 방법이란, 벡터 스페이스 모델에서 웹 페이지와 광고 텍스트 사이의 코사인 유사도 값을 웹 페이지와 광고의 유사도로 측정하는 방법이다.



(그림 1) 정확도 & 재현율

그림 1 은 키워드 매칭과 시멘틱 매칭의 정확도와 재현율의 성능을 나타낸다. 재현율이 증가할수록 시멘틱 매칭의 정확도가 키워드 매칭의 정확도에 비해 월등히 향상된 것을 볼 수 있다. 이를 통해, 본 논문에서 제안한 기법으로 만들어진 분류 트리 및 분류기가 문맥 광고에서 웹 페이지와 광고의 의미적 매칭의 높은 정확도를 보장하는 것을 입증한다.

4. 결론 및 향후 연구

본 논문에서는 문맥 광고에서 웹 페이지와 광고의 의미적 매칭에 사용되는 분류 트리 및 분류기를 작성하기 위해 ODP 의 공개 데이터를 활용하는 방법을 제안하였다. 이를 통해, 분류 트리 작성 및 분류기를 위한 학습 데이터 작성에 사람의 시간과 노력을 줄일 수 있었으며, 실험 결과 웹 페이지와 광고의 높은 분류 정확도와 웹 페이지와 광고의 높은 의미적 매칭 정확도를 얻을 수 있었다. 향후, 분류기 작성에 지지벡터 머신(SVM), 고유 베이지안(Naïve Bayesian)등의 분류기를 활용하는 연구와 광고의 분류 정확도를 높이기 위한 방법에 대한 연구를 진행할 것이다.

참고문헌

[1] The open directory project, http://www.dmoz.org/.

[2] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel, A semantic approach to contextual advertising, in *SIGIR*, 2007, pp.559–566.

[3] J. Rocchio, Relevance feedback in information etrieval, in *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, 1971, pp. 313–323.

[4] X. Qi and B. D. Davison. Classifiers without borders: incorporating fielded text from neighboring web pages. In *SIGIR*, pages 643– 650, 2008.