광고 랜딩 페이지를 이용한 문맥 광고 시스템*

이정현, 하종우, 정다운, 이상근
고려대학교 정보통신대학 컴퓨터통신공학과
e-mail : {jhbslpd, okcomputer, daounjung, yalphy}@korea.ac.kr

Contextual Advertising System using ad landing pages

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

요 약

문맥 광고의 의미론적 매칭에서 웹 페이지와 광고의 매칭 정확도는 웹 페이지와 광고의 분류 성능에 종속적이다. 그러나 광고의 짧은 텍스트는 광고 분류 성능을 하락시키는 원인이 되고 있다. 본 논문에서는, 광고 분류 성능을 높이기 위하여, 광고 랜딩 페이지를 활용하여 광고 텍스트를 확장시키는 방법을 제안하고, 실험을 통하여 그 효과를 입증한다. 추가로, 구문론적 매칭과 의미론적 매칭 방법을 적용하여 개발된 문맥 광고 엔진의 프로토타입을 제시한다.

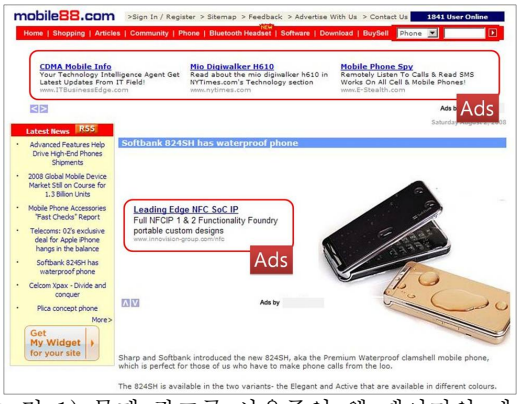
1. 서론

문맥 광고란, 온라인 뉴스 기사 또는 블로그 포스트와 같은 웹 페이지 내에, 페이지의 내용과 연관된 텍스트 광고를 함께 보여주는 웹 광고의 한 형태이다. 그림 1 은 문맥 광고의 한 예제를 보여준다.

문맥 광고 환경에서는 사용자의 광고 클릭에 따라 광고 수익이 발생하며, 광고 클릭율은 웹 페이지의 내용과 광고의 연관성이 높을수록 올라간다는 사실이 최근 연구[3]에서 밝혀졌다. 따라서, 문맥 광고에서 웹 페이지와 광고의 연관성을 정확히 측정하는 것이 가장 중요하며, 여러 연구들이 웹 페이지의 내용과 광고의 연관성을 높이기 위해 노력해왔다. 그 중 최근 연구[2]에서는 구문론적 매칭과 의미론적 매칭을 조합하는 방법이 제안되었다. 이 연구에서는, 구문론적 매칭만을 이용하였을 경우 단어의 다의성으로 인하여 무관한 광고가 매칭되는 문제점을 지적하였다. 이를 해결하기 위하여, 웹 페이지의 주제와 광고의 주제를 이용하는 의미론적 매칭 방법을 제안하였다. 여기서 이들은 공통된 주제 분류 트리에 웹 페이지와 광고의 텍스트를 분류하는 분류기를 학습시켜 사용하였다.

의미론적 매칭 방법에서, 웹 페이지와 광고의 매칭 정확도는 웹 페이지와 광고의 분류 정확도에 큰 영향을 받기 때문에, 높은 분류 정확도를 가지는 웹 페이지와 광고의 분류기를 생성하는 것이 필수적이다. 그러나, 웹 페이지에 비하여 광고는 분류에 활용되는 텍스트가 매우 적고, 사용되는 어휘가 주제를 나타내기에 정확하지 않고 모호한 문제가 있다[6]. 이는 광고 분류 정확도의 하락을 야기하며, 결과적으로 웹 페이지와 광고의 매칭 정확도를 낮추게 된다.

본 연구의 목적은 광고 분류의 정확도를 높여 웹 페이지와 광고의 매칭 정확도를 향상시키기 위한 것이다. 이를 위해, 본 연구에서는 광고의 랜딩 페이지를 광고 분류에 활용하는 기법을 제안한다. 이는, 광고의 부족한 텍스트를 보충하고, 광고의 주제를 나타내는 어휘를 추가하기 위함이다. 제안 기법의 효과를 입증하기 위해, 랜딩 페이지를 광고 분류에 활용했을 때의 광고 매칭 성능과 기존 방법의 광고 매칭 성능을 비교 측정하였다. 추가로, 본 논문에서는 구문론적 매칭과 의미론적 매칭 방법을 적용하여 개발된 문맥 광고 엔진의 프로토타입을 제시한다.



(그림 1) 문맥 광고를 사용중인 웹 페이지의 예제

2. 관련 연구

웹 페이지의 내용과 광고의 내용의 의미적 연관성을 높이기 위하여 지난 몇 년간 다수의 방법들이 제안되었다. 초기에 제안된 방법들은 벡터 스페이스 모델[7]을 기반으로 하여, 웹 페이지와 광고의 구문론적 매칭을 수행한다. 대표적으로, 웹 페이지와 광고의 어

* 이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No.2009-0077925)

휘 불일치 문제를 해결하기 위해 광고 경매 키워드를 활용하여 구문론적 매칭을 수행하는 방법[6]과 웹 페이지로부터 광고 키워드를 추출하여 광고들의 경매 키워드와 매칭하는 방법[7]등이 있다.

최근에는, 구문론적 매칭 방법들의 한계를 극복하기 위한 의미론적 매칭 방법들이 제안되었다. 여기에는, Yahoo! US 에서 만들어진 주제 분류 트리와 분류기를 이용하여 문서 레벨에서 웹 페이지와 광고를 의미를 찾고 이를 의미론적 매칭에 활용하는 방법[2]과, 어휘 레벨에서의 단어들 사이의 의미적 연관성을 계산하고 이를 웹 페이지와 광고의 의미론적 매칭에 활용하는 방법[4]등이 있다. 본 연구는, 문서 레벨에서의 웹 페이지와 광고의 의미론적 매칭 방법에서 광고의 주제를 정확히 찾는 것에 목적이 있다.

3. 광고 랜딩 페이지 활용 기법

본 연구에서, 주제 분류 트리에 광고를 자동으로 분류하기 위하여 센트로이드 기반 분류기[5]를 사용한다. 센트로이드 기반 분류기는 텍스트 분류 분야에서 널리 사용되는 대표적인 분류 알고리즘으로서, 학습 과정에서 각 주제 별로 학습 데이터들을 대표할 수 있는 센트로이드를 계산한다. 분류 과정에서는 주어진 문서와 유사도가 가장 높은 센트로이드를 갖는 주제를 문서에 할당한다. 수식(1)은 센트로이드 기반 분류기의 광고 분류 방법을 나타낸다.

$$C_{max} = arg\max_{c_k \in C} \frac{\vec{\mu}_k}{\|\vec{\mu}_k\|} \cdot \frac{\vec{d}}{\|\vec{d}\|} \quad (1)$$

위 수식에서, C 는 주제 분류 트리의 모든 주제들의 집합이고, $\vec{\mu}_k$ 는 주제 c_k 의 센트로이드이며, $\vec{d}$ 는 광고 텍스트에 대한 단어 벡터이다.

센트로이드 기반 분류기에서 광고 분류 정확도를 향상시키기 위하여, 본 연구에서는 광고 랜딩 페이지를 활용하였다. 광고 랜딩 페이지란, 광고에 링크가 연결된 웹 페이지로서, 사용자가 광고를 클릭하였을 때 볼 수 있는 페이지를 말한다. 광고 랜딩 페이지에는 광고의 주제를 보충 설명해줄 수 있는 텍스트들이 많이 존재하기 때문에 광고 분류 정확도 향상이 도움이 될 수 있다. 그림 2 는 광고 랜딩 페이지의 예제를 보여준다.

(그림 2) 광고 랜딩 페이지 예제

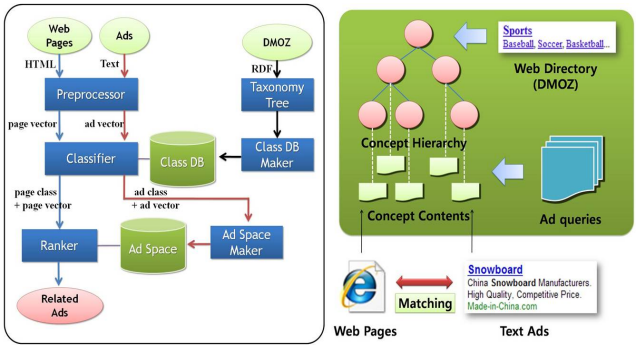
광고 랜딩 페이지를 활용한 센트로이드 기반 분류기의 광고 분류 방법은 수식(2)와 같다.

$$C_{max} = arg\max_{c_k \in C} \left(\beta \times \frac{\vec{\mu}_k}{\|\vec{\mu}_k\|} \cdot \frac{\vec{d}}{\|\vec{d}\|} + (1-\beta) \times \frac{\vec{\mu}_k}{\|\vec{\mu}_k\|} \cdot \frac{\vec{l}}{\|\vec{l}\|} \right) \quad (2)$$

위 수식에서, $\vec{l}$ 은 광고 랜딩 페이지의 텍스트에 대한 단어 벡터이고, β 는 광고 텍스트에 대한 가중치로서 0 부터 1 사이의 값을 가진다. 이 수식은 광고와 각 주제의 센트로이드의 유사도를 구할 때, 광고 텍스트와 센트로이드 사이의 유사도와 광고 랜딩 페이지의 텍스트와 센트로이드 사이의 유사도의 선형조합을 이용하는 방법이다.

4. 광고 엔진의 프로토타입

본 연구에서는 구문론적 매칭과 의미론적 매칭 방법을 적용하여 문맥 광고 엔진의 프로토타입을 개발하였다. 그림 2 는 개발된 엔진의 시스템 구성도를 나타낸다.



(그림 3) 문맥 광고 엔진 프로토타입의 시스템 구성도

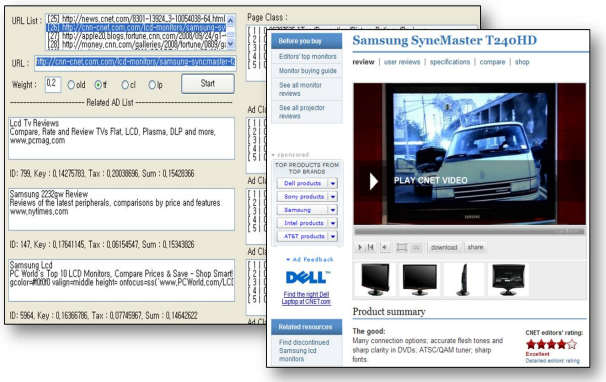
개발된 시스템은 1) DMOZ[1]의 데이터와 광고 키워드를 이용하여 주제 분류 트리와 분류기를 학습시키는 모듈, 2) 광고 데이터를 분석하고 분류하여 Ad space 를 구축하는 모듈, 3) 주어진 웹 페이지에 대해 광고를 랭킹 하는 모듈, 세 가지로 구성된다.

첫 번째 모듈은 DMOZ 의 디렉토리를 활용하여 웹 페이지와 광고를 위한 주제 분류 트리를 생성한다. 또한, 생성된 분류 트리에 웹 페이지와 광고를 자동으로 분류하기 위하여, 주제 분류 트리의 각 노드에 연관된 광고 키워드를 수집하여 학습 데이터를 생성하고, 센트로이드 기반 분류기를 학습시킨다.

두 번째 모듈은 각 광고에 대해 광고 자체와 광고 랜딩 페이지에서 텍스트를 추출 및 분석하고, 학습된 분류기를 통하여 자동 분류하여 주제를 정한다. 주제가 정해진 광고는 Ad space 에 추가되며, 이는 웹 페이지가 주어졌을 때 광고 랭킹을 효율적으로 하기 위한 인덱스 구조를 가진다.

세 번째 모듈은 주어진 웹 페이지의 텍스트를 추출 및 분석하고, 학습된 분류기를 통하여 자동 분류하여 주제를 정한다. 그 후, Ad space 에 저장된 광고들과 구문론적 및 의미론적 매칭을 수행하여 광고의 랭킹을 정한다.

그림 3 은 개발된 문맥 광고 엔진 프로토타입의 스냅샷으로, 우측의 웹 페이지에 대해 시스템이 랭킹한 top 3 광고 결과를 보여준다.



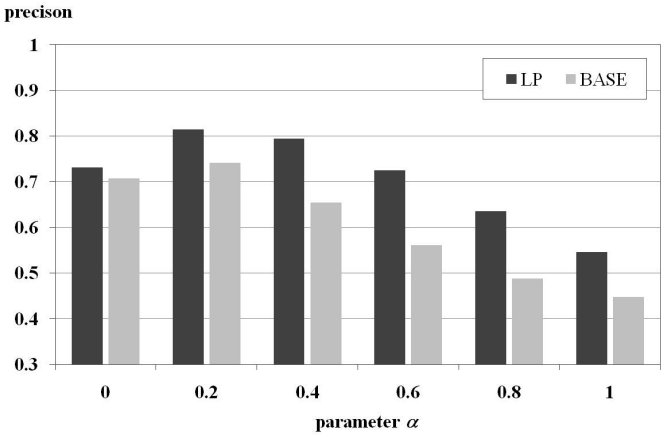
(그림 4) 문맥 광고 엔진 프로토타입의 스냅샷

5. 실험 결과

본 연구에서는, 광고 분류에 광고 랜딩 페이지를 활용한 제안 기법이 웹 페이지와 광고 매칭 성능에 미치는 효과를 알기 위한 실험을 수행하였다. 실험에 사용된 광고 집합은 구글 광고 검색을 통해 수집되었으며, 전체 광고 개수는 6675 개이다. 웹 페이지는 온라인 뉴스 또는 블로그 포스트로 한정하였으며 서로 상이한 주제를 가지는 웹 페이지를 검색 엔진을 통해 직접 수집하였다. 실험에 사용된 전체 웹 페이지의 개수는 30 개이다. 또한, DMOZ 의 디렉토리 중 3 레벨까지 6503 개의 디렉토리를 선택하여 주제 분류 트리를 생성하였다. 각 디렉토리는 분류 트리의 각 노드가 되며, 디렉토리 간의 계층 구조가 그대로 유지되었다.

평가 지표로는 Precision at 5 가 이용되었다. 문맥 광고 환경에서 웹 페이지에 포함되는 광고의 수는 3~5 개 정도로 제한되기 때문에, Precision at 5 의 성능이 매우 중요하다.

그림 5 는 광고 텍스트 자체만을 이용하였을 경우의 광고 매칭 성능(BASE)과 광고 랜딩 페이지의 텍스트를 추가로 이용하였을 경우의 광고 매칭 성능(LP)를 파라미터 α에 따라 비교 측정한 결과이다. α는 구문론적 매칭과 의미론적 매칭의 선형 조합을 위한 가중치로서, 값이 클수록 의미론적 매칭에 더 큰 비중을 두는 것을 의미한다. 이 결과에서 확인할 수 있듯이, 모든 α값에서 광고 랜딩 페이지를 활용하는 것이 더 높은 정확도를 얻을 수 있었다. 가장 높은 정확도를 갖는 α값에서, 광고 랜딩 페이지를 이용하였을 경우 광고 매칭 정확도가 약 10%정도 향상되었다. 이로서, 광고 랜딩 페이지를 사용하는 것이 광고 매칭의 정확도 향상에 도움을 준다는 것을 알 수 있었다. 또한, 의미론적 매칭의 비중이 커질수록 광고 랜딩 페이지를 이용한 경우의 광고 매칭 정확도가 큰 폭으로 향상되었다. 이는, 의미론적 매칭이 광고의 분류 정확도에 많은 영향을 받는다는 사실을 입증해준다.



(그림 5) 실험 결과

6. 결론

문맥 광고에서 의미론적 매칭 방법을 이용하여 웹 페이지와 광고의 연관성을 높이기 위해서, 웹 페이지와 광고의 주제를 정확히 찾는 것이 중요하다. 그러나 광고의 짧은 텍스트의 문제로 인해 광고의 분류 정확도가 낮아지는 문제점이 있었다. 본 논문에서는, 이 문제점을 해결하기 위하여 광고의 랜딩 페이지를 활용하여 광고 텍스트를 확장하였다. 확장된 텍스트를 광고 분류에 활용하는 방법을 제안하였으며, 실험을 통해 제안 방법이 웹 페이지와 광고의 매칭 정확도를 향상시키는데 도움을 준다는 것을 입증하였다.

추가로, 구문론적 매칭과 의미론적 매칭 방법을 적용한 문맥 광고 엔진의 프로토타입을 개발하였으며, 이 엔진의 시스템 구성에 대해 본 논문에서 제시하였다. 향후, 이 엔진을 바탕으로 웹 페이지와 광고의 연관성을 높이기 위한 추가적인 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 *Proc. SIGIR '07*, 2007, pp.559-566.

[3] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541, 2003.

[4] M. Ciaramita, V. Murdock, and V. Plachouras. Semantic associations for contextual advertising. *International Journal of Electronic Commerce Research - Special Issue on Online Advertising and Sponsored Search*, 9(1), 2008.

[5] E.-H. S. Han and G. Karypis. Centroid-based document classification: analysis and experimental results. In *Proc. PKDD '00*, pages 116–123, 2000.

[6] B. A. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted dvertising. In *Proc. SIGIR '05*, pages 496–503, 2005.

[7] G. Salton, A. Wong, and C. Yang. A vector space model for information retrieval. *Journal of the American Society for Information Science*, 18(11):613–620, 1975.

[8] W. tau Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *Proc. WWW '06*, pages 213–222, 2006.