방문 웹 페이지 로그를 이용한 개인화된 문맥광고 기법*

심규선, 이명수, 최재호, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail:{bluesks, lms9711, redcolor25, yalphy}@korea.ac.kr

Personalized Contextual Advertising Scheme using Logs of Web Page Visited

Kyu-Sun Shim, Myong-Soo Lee, Jae-ho Choi, SangKeun Lee
Division of Computer and Communication Engineering, Korea University

요 약

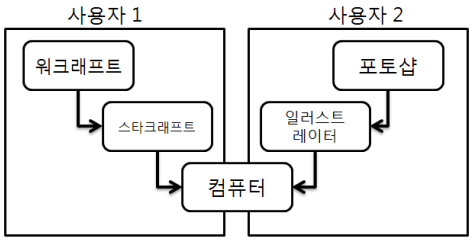
사용자가 방문한 웹 사이트와 연관된 광고를 웹 페이지에 실어주는 문맥광고 관련 연구가 광고 효율성 측면에서 최근 주목을 받고 있다. 이러한 문맥광고 관련 연구의 핵심은 웹 페이지와 웹 광고간의 연관성을 높여주는데 있는데, 연관성 향상 방안으로 최근 사용자 의도 분석을 통한 연관성 향상 기법이 많이 연구되고 있다. 그러나 기존 연구에서는 사용자가 로그인을 해야 하거나, 로그 정보를 일정기간이상 수집해야만 사용자 의도 분석이 가능 하다는 문제점이 존재 한다. 본 논문에서는 로그인이나 많은 양의 로그정보 수집 없이 한 세션 내에서 방문한 웹 페이지 로그만을 이용하여 개인화된 문맥 광고를 제공하는 문맥광고 기법을 제안한다. 실험 결과에서는 제안하는 기법이 기존의 광고 기법에 비해 사용자가 판단하는 웹 페이지와 웹 광고의 연관성 (precison) 이 기존의 기법에 비해 높아짐을 증명한다.

1. 서론

현재, Web의 가장 큰 수익원인 웹 광고는 스폰서 광고와 문맥광고 분류할 수 있다[1][2]. 스폰서 광고는 웹 검색 엔진에서 질의어와 연관된 광고를 배치하는 웹 광고 형태이며, 문맥 광고는 사용자가 웹 페이지를 방문할 때 웹 페이지의 내용과 연관된 광고를 배치하는 광고 형태이다. 광고를 검색 결과와 함께 배치하는 스폰서 광고에 비해, 웹 페이지에 광고를 배치하는 문맥 광고가 광고 효과가 크다는 것이 광고 수입을 통해 증명되고 있으며, 이로 인해 문맥 광고에 대한 관심이 증가하고 있다[1][3].

이런 문맥 광고에서 가장 중요한 이슈 중에 하나는 웹 페이지와 광고 간의 연관성을 높이는 것이다. 웹 페이지와 광고의 연관성은 결국 사용자의 웹 페이지 클릭률에 의해 결정되기 때문에, 사용자가 필요로 하는 광고를 보여주는 것이 중요하다. 하지만 기존의 연구에서는 주로 웹 페이지의 내용과 광고 간의 유사도만을 향상시키는데 초점을 맞추었다[1]. 이는 웹 페이지와 광고 간의 유사도만 측정하기 때문에 특정 웹 페이지에 붙는 광고는 모든 사용자에 대해 동일하다. 다시 말해, 다양한 의도를 가진 사용자의 요구를 웹 광고에 전혀 반영할 수 없다. 예를 들어, 그림 1과 같은 경우 사용자 1, 2는 서로 다른 의도를 가지고 같은 웹 페이지를 방문하였다. 그림 1에서 사각형 내의 단어는 사용자가 보고 있는 웹 페이지의 주제를 나타낸다. 그림 1에서 사용자 1과 2는 모두 '*컴퓨터*'와 관련된 웹 페이지를 보고 있으나, 이전 방문한 페이지를 보면, 사용자 1은 '*게임*' 관련 주제를 사용자 2는 '*그래픽 툴*' 관련 주제를 검색하고 있다. 하지만 사용자 1과 2가 현재 동일한 웹 페이지를 보고 있기 때문에 사용자 의도가 반영되지 않는다면, 사용자 1과 2는 동일한 웹 페이지와 광고를 보게 된다. 이때, 사용자 의도를 반영한 사용자별로 다른 웹 광고를 보여주면 광고 효과를 극대화할 수 있다.

사용자들에게 개인화된 광고를 보여주기 위하여 구글의 Interest-Based Ads, 야후의 APT 등은 사용자들이 직접 관심사를 입력하거나 사용자가 로그인하도록 하는 방법에 기반하여 사용자들의 관심사를 반영하여 사용자가 필요로 하는 광고를 제공하였다. 하지만 위 방법은 사용자들이 직접 관심분야를 입력하지 않거나, 로그인하지 않으면 불가능한 기법이다.



(그림 1) 사용자 웹 행동패턴

최근, 사용자 입력이 없이 클릭트루 (click-through) 로그에 기반한 기법이 제안되었다[3]. 이는 사용자의 클릭 정보를 로그로 남겨 사용자가 방문하고 있는 웹 페이지와 관련된 광고의 유사도를 측정할 때 반영한다. 이 기법의 핵심 아이디어는 클릭을 많이 한 웹 페이지 또는 광고, 질의 단어의 경우 가중치가 증가시키는 것이다. 이 경우 클릭량이 적은 관심사는 반영되지 못할 뿐만 아니라 일정 기간 동안 모은 사용자 정보에 기반하기 때문에 많은 데이터를 저장할 저장소, 계산 오버헤드를 야기시킨다.

본 연구에서는 별도의 사용자 입력 없이, 웹 로그만을 이용하여 사용자의 의도를 기존 문맥 광고 기법에 반영하는 방법을 제안한다. 사용자의 의도를 반영하기 위하여 본 연구에서는 사용자 정보를 수집한다. 제안하는 기법에서는 한 세션 내의 웹 로그만 저장한다. 이는, 저장소 및 계산 오버헤드를 피하기 위해 그림 1과 같이 사용자의 웹 행동 패턴에 지역성 (Locality) 이 있는 점에서 착안한 것이다. 이미 많은 연구에서 사용자가 이전 페이지와 전혀 관계없는 웹 페이지를 방문하는 것은 매우 낮은 확률이라는 것을 발견하였다[4]. 이렇게 웹 행동패턴의 지역성을 이용해 한 세션 내의 웹 로그를 이용한다면 사용자의 의도를 적용하여 개인화된 광고를 제공할 수 있을 뿐 아니라, 많은 양의 데이터를 필요로 하지 않기 때문에 오버헤드 역시

* 이 연구에 참여한 연구자의 일부는 '2단계 BK21사업'의 지원비를 받았음."

줄일 수 있다.

본 논문은 다음과 같이 구성되었다. 2장에서는 제안하는 기법을 설명하며, 3장에서는 성능평가 결과를 소개한다. 마지막으로 4장에서 본 논문의 결론을 맺는다.

2. 사용자 의도 매칭 기법

본 연구에서 제안하는 개인화된 문맥 광고 기법은 사용자가 한 세션 내에서 방문한 웹 페이지의 히스토리 정보를 반영하여 광고를 제공한다. 그림 1의 예에 의하면, 사용자 1은 *'컴퓨터 게임'* 관련 광고를, 사용자 2는 *'컴퓨터 그래픽 툴'* 관련 광고를 볼 수 있다. 이를 위해 본 연구에서는 아래와 같은 기법을 제안하였다.

사용자들에게 개인화된 광고들을 제공하기 위해 먼저 사용자 개인의 웹 로그를 저장하여야 한다. 이를 위해 표 1과 같이 로그 테이블을 만드는데, 로그 테이블은 사용자를 식별하기 위한 식별자와 로그 데이터로 구성되며, 식별자는 사용자의 세션 ID, 로그 데이터는 세션이 유지되는 동안 방문한 웹 페이지가 분류된 주제가 저장된다. 방문한 페이지의 주소 및 내용을 저장한다면, 프라이버시 문제가 발생할 수도 있기 때문에 제안 기법에서는 방문한 웹 페이지가 분류된 주제를 저장한다. 분류는 기존 기법[1]에서 제안된 시멘틱 매칭 방법에서 작성된 분류 트리와 분류기를 사용하였다. 분류기를 통해 분류된 결과는 분류 트리 상의 노드이며, 이 노드는 분류 트리 내의 노드 중 해당 웹 페이지에 가장 적합한 주제를 나타낸다.

<표 1> 사용자 방문 웹 페이지 로그 테이블

식별자	사용자 세션 ID
로그 데이터	방문 페이지 분류 주제(v_1)
	⋮
	방문 페이지 분류 주제(v_k)

표 1과 같이 저장된 데이터는 사용자의 세션이 유지되는 동안만 저장하기 때문에 일정 기간 동안 유지되는 기존 클릭투르 기반의 기법[3]에서의 데이터의 양보다 작다. 작성된 사용자 방문 웹 페이지 로그 테이블을 적용하여 웹 페이지와 광고의 유사도 측정을 한다. 이를 위해 본 연구에서는 기존 연구[1]에서 웹 페이지와 광고간의 유사도를 평가하기 위해 제안된 시멘틱-신테틱 매칭 (Semantic-Syntatic matching) 기법에 본 연구에서 제안된 사용자 로그 데이터를 아래 식 (1)과 같이 적용하여 계산한 후, 상위 n개의 광고를 추출한다.

$$TotalScore(p,a_i,v_j) = \sum_{j=1}^{k} UserScore(a_i,v_j) + Score(p,a_i) \quad (1)$$

식 (1)에서, p는 사용자가 보고 있는 웹 페이지를 나타내며, a_i는 i번째 광고를 의미한다. v_j는 표 1에 저장되어 있는 방문한 페이지의 분류 주제이며, k는 저장된 로그 데이터의 개수이다. $UserScore(a_i,v_j)$은 광고와 방문한 웹 페이지 로그 데이터의 유사도를 측정하여 사용자 의도를 반영하고자 한다. 이를 위해, 먼저, 광고를 분류기를 이용하여 주제를 분류하고, 분류된 주제와 로그 데이터와의 유사도를 측정한다. 이를 위해 본 연구에서는 기존 기법[1]에서 사용된 분류 트리 노드 간의 거리를 측정하는 방법을 이용하여 계산하였다.

$$UserScore(a_i,v_i) = idist(LCA(a_i,v_i),v_i) \quad (2)$$

(2)에서 $LCA(a_i,v_i)$는 광고가 분류된 주제와 로그 데이터에 있는 방문 페이지가 분류된 주제의 분류 트리상의 노드들과 가장 근접한 공통 부모를 뜻한다. $idist(a,b)$는 a 노드의 자식 노드들의 개수 / b 노드의 자식 노드들의 개수를 뜻한다. (2)를 통해 분류 트리 노드 간의 거리를 측정할 수 있다. $Score(p,c_i)$은 기존 기법[1]에서 제안된 기법으로 웹 페이지와 광고와의 연관성을 측정하기 위한 기법이다.

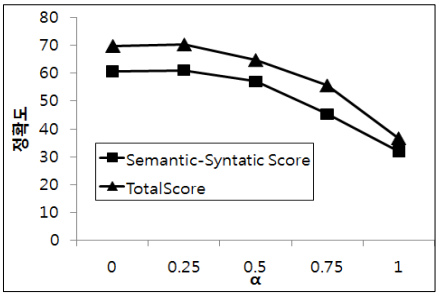
$$Score(p,c_i) = \alpha \cdot TaxScore(p,c_i) + (1-\alpha) \cdot KeywordScore(p,c_i) \quad (3)$$

식 (3)에서 $TaxScore(p,c_i)$는 웹 페이지와 광고의 분류된 주제들의 유사도를 측정하고, $KeywordScore(p,c_i)$는 웹 페이지와 광고 텍스트들 간의 유사도를 측정한다. α 값은 *TaxScore*와 *KeywordScore*간의 상대적 가중치를 결정한다.

본 연구에서는 기존 기법[1]에서 고려하지 않은 *UserScore*를 적용하여 식 (1)을 통하여 그림 1의 예제에서 사용자 1, 2에게 각각 의도에 맞는 광고를 제공할 수 있다.

3. 성능평가

본 연구에서는 사용자 방문 웹 페이지 로그 데이터를 적용하여 웹 페이지와 광고와의 매칭 성능을 측정하였으며, 기존 연구[1]에서 제안된 시멘틱-신테틱 매칭과 비교하였다. 이를 위해 기존 연구[1]에서 제안된 분류 트리와 분류기를 사용하였으며, 웹 페이지와 유사도가 높다고 판단되는 상위 6개의 광고에 대하여 평가를 하였다. 성능 평가에 사용된 웹 페이지의 수는 30개이며, 광고의 수는 15,000개이다.



(그림 2) 웹 페이지와 광고의 정확도

그림 2는 웹 페이지를 방문하였을 때, 유사도가 높은 상위 6개 광고의 정확도 (precison)를 측정하였다. *α* 값은 섹션 2, 식 (3)에서 *TaxScore*와 *KeywordScore*간의 상대적 가중치를 나타내는데, α 값이 커질수록 분류기의 성능으로 인하여 정확도가 떨어진다. 하지만 본 연구에서 제안한 기법은 기존 기법에 비해 모든 α 값에 대하여 향상된 정확도를 보여준다. 또한 섹션 2에서 언급한 *UserScore*의 계산 오버헤드는 1초 미만으로 낮게 측정되었다.

4. 결론 및 향후 연구

본 연구에서는 사용자들이 원하는 광고를 매칭시키기 위하여 사용자 방문 웹 페이지 히스토리를 이용하여 개인화된 문맥 광고 기법을 제안하였으며, 성능 평가를 통해 기존 기법에 비해 정확도가 향상되었음을 보여주었다. 이를 통해 사용자의 로그인을 하지 않고, 적은 양의 데이터만을 이용하여도, 사용자들에게 같은 웹 페이지에서도 사용자들의 의도에 적합한 광고를 제공할 수 있음을 알 수 있다. 본 연구는 사용자 로그 데이터 분류하는 방법이 기존의 분류 기법에 의존되어 있으므로, 본 연구에서 제안한 방법의 정확도를 향상시킬 수 있도록 향후 웹 페이지 분류 기법에 대한 연구를 진행할 예정이다.

참고문헌

[1] A. Z. Broader, M. Fontoura, V. josifovski, and L. Riedel, A semantic approach to contextual advertising, In Proceedings of ACM SIGIR, 559-566, 2007.

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

[3] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang and Z. Chen. How much the Behavioral Targeting can help online advertising? In Proceedings of WWW, 261-270, 2009.

[4] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In Proceedings of ACM SIGKDD, 538-543, 2002.