문맥 광고를 위한 링크 분석 기법*

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

Link Analysis for Contextual Advertising

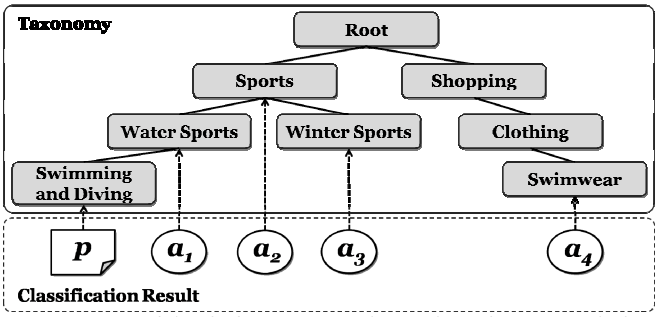
JongWoo Ha, Jung-Hyun Lee, Sang-Hyun Park, SangKeun Lee
Division of Computer and Communications Engineering, Korea University

요 약

문맥 광고에서 계층적인 분류 트리를 활용하여 의미적으로 연관된 광고를 매칭하는 기법이 소개되었다. 하지만 기존 기법은 계층 구조의 특성에 기인하여 임의의 광고의 연관성을 측정할 때에는 적합하지 않다. 이러한 문제를 해결하기 위하여 본 논문에서는 분류 트리를 유사도 그래프로 변환한 후 개인화된 페이지 랭크를 응용한 링크 분석 기법을 적용하여 광고의 의미적 연관성을 측정하는 기법을 제안한다. 실험을 통하여 제안 기법이 문맥 광고에서 광고 매칭의 정확도 성능을 향상시킴을 확인하였다.

1. 서론

문맥 광고는 온라인 광고의 한 형태로서, 블로그 포스트, 온라인 뉴스 페이지와 같은 일반적인 웹 페이지의 내부에 광고를 제공한다. 임의의 광고를 제공하는 배너 광고와 달리, 웹 페이지의 내용을 분석하여 그와 연관된 광고를 제공하는 것이 문맥 광고의 특징이다. 문맥 광고가 구글, 야후와 같은 대표적인 인터넷 서비스 업체의 주요한 수익원으로 부각됨에 따라 광고 매칭의 정확도(웹 페이지와 광고의 연관성)를 향상하기 위한 많은 연구가 진행되고 있다.



(그림 1) 계층적인 분류 트리를 활용한 광고 매칭

최근 광고 매칭의 정확도를 높이기 위한 의미적 매칭(semantic matching) 기법이 소개되었다[3]. 이 기법에서는 먼저, 계층적인 분류 트리에 웹 페이지와 광고를 분류하여 그 주제를 파악한다. 그림 1 은 ODP(Open Directory Project)[1]로부터 추출한 분류 트리에 웹 페이지(p)와 광고(a_i)가 분류된 예시를 나타낸다. 이러한 분류 결과를 바탕으로 하여 웹 페이지의 주제와 의미적으로 연관된 광고를 제공한다.

계층 구조는 'is-a' 관계의 집합이기 때문에, 기존 기법에서는 계층 구조에서 거리가 가까울수록 유사하다고 가정한다. 따라서 그림 1 의 예제에서 a_1, a_2, a_3, a_4 의 순서로 광고의 연관성이 높은 것으로 판단한다. 그 결과, 주어진 웹 페이지와 의미적으로 가장 연관된 광고인 a_4 의 연관성이 가장 낮게 평가되는 문제가 있다. 이러한 결과는 계층 구조에서 'is-a' 관계가 루트 아래의 서브 트리 내에서만 정의되기 때문에, 서로 다른 서브 트리에 위치한 두 노드 간의 관계를 파악할 수 없기 때문에 발생한다. 따라서 임의의 광고에 대한 의미적 연관성을 계산하기 위하여 계층 구조만을 활용하는 것은 적합하지 않다.

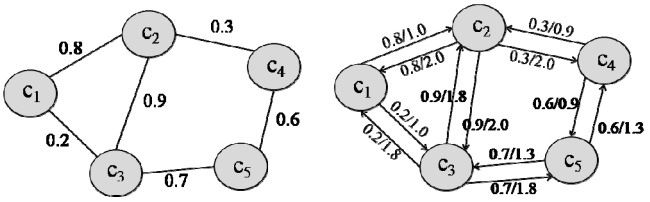
이러한 문제를 해결하기 위해서 본 논문에서는 분류 트리로부터 유사도 그래프를 생성한 후, 개인화된 페이지 랭크(personalized PageRank)[4]에 기반한 링크 분석 기법을 적용하여 광고의 의미적 연관성을 측정하는 기법을 제안한다.

2. 제안 기법

유사도 그래프는 정점들의 집합인 C 와 가중치가 주어진 간선들의 집합인 E 로 정의한다. C 의 각 정점은 분류 트리에서의 노드(클래스)와 1:1 매칭된다. 임의의 두 정점의 유사도가 임계값(threshold) τ 보다 높을 경우 두 정점 간의 유사도를 가중치로 갖는 간선을 추가한다. 정점 간 유사도를 계산하기 위하여 본 제안 기법에서는 분류 트리의 각 노드마다 할당되어 있는 학습 데이터를 활용한다. 벡터 공간 모델[2]에 기반하여 각 학습 데이터를 텀 벡터로 표현한 후 각 정점마다 중심 벡터(centroid vector)[5]를 계산한다. 유사도 그래프에서 임의의 두 정점 간의 유사도는 중심

* 이 연구에 참여한 연구자는 '2 단계 BK21 사업'의 지원비를 받았음.

벡터 간의 코사인 유사도(cosine similarity)[2]로 정의한다. 이러한 유사도 그래프에서는 그림 1 에서의 'Swimming and Diving' 노드와 'Swimwear' 노드 간의 유사도가 높게 측정되는 특징이 있다. 그림 2(a)는 유사도 그래프의 예제를 도식화한 것이다.



(a) 유사도 그래프 (b) 전환 그래프

(그림 2) 유사도 그래프를 활용한 링크 분석

제안 기법에서 광고의 의미적 연관성은 개인화된 페이지랭크를 응용한 랜덤 서퍼가 광고가 분류된 정점을 방문할 확률로 정의한다. 이를 계산하기 위하여 유사도 그래프를 활용한다. 정형적으로, 제안 기법은 연관성 행렬 R 을 계산하며 R 의 (i, j) 성분(r_{ij})은 웹 페이지가 c_j 에 분류되었다고 가정하였을 때 랜덤 서퍼가 정점 c_i 를 방문할 확률로 정의한다. 이러한 방식으로, 웹 페이지와 광고의 분류 결과에 상관없이 광고의 의미적 연관성을 사전에 계산할 수 있다.

랜덤 서퍼가 유사도 그래프를 순회하는 패턴은 두 가지의 경우로 정의되며, 매 순간마다 (1-d)의 확률로 첫 번째 패턴을 따르며 d 의 확률로 두 번째 패턴을 따른다. 댐핑 팩터인 d 는 0 에서 1 사이의 값을 가지며 실험에 따라 최적의 값을 찾는다. 랜덤 서퍼가 첫 번째 패턴을 따를 경우 다음 번 방문하는 정점은 웹 페이지가 분류된 c_j 가 된다. 랜덤 서퍼가 두 번째 패턴을 따를 경우 현재 방문 중인 정점과 연결된 정점 중 하나를 간선의 가중치와 비례하는 확률로 방문하게 된다. 랜덤 서퍼가 두 번째 패턴을 따라서 다음 정점으로 이동하는 확률은 수식 (1)와 같이 계산한다.

$$tw_{ij} = \frac{sim(c_i, c_j)}{\sum_{c_k \in N(c_j)} sim(c_k, c_j)} \quad (1)$$

여기서 tw_{ij}는 j 번째 클래스에서 i 번째 클래스로 이동할 확률, $sim(c_i,c_j)$는 코사인 유사도로 구한 간선의 가중치, $N(c_j)$는 c_j 와 연결된 정점들의 집합이다. 그림 2(b)는 그림 2(a)의 유사도 그래프에서 랜덤 서퍼의 두 번째 이동(전환) 패턴에 따른 확률을 도식화한 것이다. 예를 들어서, 랜덤 서퍼가 현재 c_1 을 방문하고 있으며 d 의 확률로 두 번째 패턴을 따라 유사도 그래프를 순회할 때, 0.8/1.0 의 확률로 c_2 로 이동하며, 0.2/1.0 의 확률로 c_3 으로 이동하게 된다. 랜덤 서퍼의 두 가지 순회 패턴에 대한 정의를 바탕으로 하여 r_{ij} 는 수식 (2)과 같이 계산한다.

$$r_{ij} = d\left[\sum_{c_k \in I(c_i)} tw_{ij} \cdot r_{kj}\right] + (1-d)t_{ij} \quad (2)$$

여기서 $I(c_i)$는 c_i 와 연결된 정점들의 집합이다. t_{ij} 는 첫 번째 패턴을 계산하기 위한 것으로서, 웹 페이지가 c_j에 분류되었을 때 c_i의 신뢰도(trusted weight)이다. 랜덤 서퍼의 확률 분포를 웹 페이지가 분류된 정점에 집중될 수 있도록 하기 위하여 제안 기법에서 t_{ij} 는 c_i 가 c_j 가 서로 같은 정점일 때에만 1 로, 나머지 경우에는 0 으로 정의한다.

3. 실험 평가

기존 기법과의 성능 비교를 위하여 잘 알려진 웹 디렉토리인 ODP[1]로부터 계층적인 분류 트리를 추출하였다. ODP 의 에디터가 각 노드에 할당해놓은 웹 페이지의 제목, 설명, URL 을 학습 데이터로 설정하여 Rocchio 분류기[5]를 활용하였다. 따라서 기존의 의미적 매칭 기법(TaxScore)과 제안 기법(GraphScore)은 동일한 분류 체계와 분류 결과를 바탕으로 계산하였다. 또한 웹 페이지와 광고에 단순히 벡터 공간 모델을 적용하여 유사도를 측정하는 KeywordScore 기법을 구현하여 본 실험에서의 비교군으로 설정하였다[3]. 표 1 은 KeywordScore, TaxScore, GraphScore 의 Top-k 정확도(Precision@k) 성능을 나타낸 것이다. 기존 연구에서도 확인된 바와 같이 의미적 매칭 기법은 단순 벡터 공간 모델을 활용한 기법보다 성능이 뛰어나다. TaxScore 는 웹 페이지가 분류된 노드와 동일한 서브트리에 있는 광고만을 선택하지만 GraphScore 는 동일한 분류 체계 내의 모든 광고 데이터를 고려하여 의미적 연관성을 측정한다. 이로 인하여 표 1 에서와 같이 8~9% 이상 성능이 향상됨을 확인할 수 있다.

<표 1> Top-k 정확도 성능 비교

광고 매칭 기법	k=1	k=3	k=5
KeywordScore	0.733	0.639	0.585
TaxScore	0.794	0.761	0759
GraphScore	**0.870**	**0.863**	**0.849**

4. 결론

본 논문에서는 문맥 광고에서 광고의 의미적 연관성을 효과적으로 계산하기 위한 링크 분석 기법을 제안하였다. 제안 기법은 계층 구조에서 단순 거리 측정에 기반한 기존 기법보다 광고 매칭의 정확도 성능을 향상시키는 것을 확인하였다. 향후 유사도 그래프를 생성하는 고도화된 기법을 연구할 계획이다.

참고문헌

[1] "The open directory project", http://www.dmoz.org/.
[2] R.A. Baeza-Yates and B.A. Ribeiro-Neto, "Modern Information Retrieval", ACM Press/Addison-Wesley, 1999.
[3] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Reedel, "A semantic approach to contextual advertising", in SIGIR, 2007, pp.559-566.
[4] G. Jeh and J. Widom, "Scaling personalized web search", in WWW, 2003, pp.271-279.
[5] j. Rocchio, "Relevance feedback in information retrieval", in The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice-Hall, 1971, pp.313-323.