키워드 마케팅을 위한 연관 키워드 추출 기법

이성진° 이수원
숭실대학교 대학원 컴퓨터학과
ptnrev93°@ptnrev93.mining.ssu.ac.kr, swlee@computing.ssu.ac.kr

A Related Keyword Group Extraction Method for Keyword Marketing

SeongJin Lee° Soowon Lee
Dept. of Computing, Graduate, School Soongsil University

요 약

인터넷 광고 시장의 급속한 성장과 함께 보다 효율적인 광고기법을 개발하기 위한 노력들이 이루어지고 있는 가운데 최근 들어 검색엔진의 특성을 이용한 키워드 광고가 주목을 받고 있다. 키워드 광고란 사용자가 입력한 검색어와 유사한 범주에 속하는 사이트의 광고를 검색 결과 페이지 상단에 보여주는 것을 말한다. 그러나, 키워드 광고는 키워드를 판매할 수 있는 위치가 한정적이기 때문에 판매 가능성이 있는 키워드에 대한 관리 및 판매 전략이 요구된다.

본 논문에서는 판매 가능성이 있는 키워드에 대한 관리 전략 수립을 위하여 연관 키워드 그룹을 자동으로 추출하는 기법을 제안한다. 연관 키워드 그룹의 생성은 사용자가 입력한 검색어에 의해 노출되는 사이트들을 묶어 그룹으로 형성하고 사이트 그룹의 중요 키워드를 추출한 다음 키워드간의 연관성을 판단하는 과정으로 이루어진다. 본 논문에서는 연관 키워드 그룹 추출의 각 단계를 구체적으로 설명하고 실험 결과를 분석한다. 마지막으로 연구의 결론과 향후 연구 과제에 대하여 기술한다.

1. 서 론

인터넷 광고 시장의 양적인 규모 확대와 함께 효과적인 광고를 위한 대안들이 모색되고 있는 가운데 최근 가장 주목받고 있는 것이 키워드 광고이다. 키워드 광고란 사용자가 입력한 키워드와 유사한 범주에 속하는 제품의 광고를 검색 결과 페이지 상단에 보여주는 것이다. 키워드 광고는 일반적인 광고에 비해 그 효과가 탁월한 것으로 알려져 있다[1].

그러나 키워드 광고는 키워드 판매 위치가 한정적이라는 점, 특정 키워드는 노출횟수가 높음에도 사이트를 홍보할 수 있는 키워드로 인지하지 못하는 경우가 있다는 점, 노출횟수가 높은 키워드에 대한 구매 비용의 부담으로 구매하지 못하는 경우가 있을 수 있다는 점 등의 문제점이 있다. 만약 특정 키워드와 비슷한, 서로 연관성이 존재하는 키워드들을 찾아내어 관리할 수 있다면 구매 완료된 키워드와 유사한 키워드를 추천하거나, 특정 사이트에 대한 추천 키워드를 선정하여 고객에게 추천하거나, 상대적으로 구매 가격이 저렴한 키워드를 추천하여 판매할 수도 있게 된다. 이는 고객사별 타겟 마케팅을 가능하게 함으로써 검색 엔진을 운영하고 있는 회사의 수입 증대 효과를 가져다 준다. 현재 검색 엔진사에서는 서퍼에 의한 판단으로 연관 키워드를 찾고 있다. 이는 시간과 비용이 많이 소모될 뿐만 아니라 서퍼의 주관에 의해 그 결과가 달라진다는 한계가 있다. 따라서 본 논문에서는 연관 키워드 그룹을 자동으로 추출하는 기법에 대해 제안한다.

2. 관련 연구

본 논문에서 제안하는 연관 키워드 그룹 추출은 각 사이트를 대표하는 중요 키워드를 선별해 내고, 검색 키워드(타겟 키워드 후보)와 각 사이트를 구성하는 키워드 간의 연관성을 판단하여 연관 키워드 그룹을 생성한다. 2장에서는 각 사이트를 대표하는 키워드를 선별해 내는 기법으로 TFIDF에 대해 알아보고, 키워드 간의 연관성을 파악하는 기법으로 연관규칙과 협력적 추천에 대해서 알아본다.

2.1 TFIDF(Term Frequency Inverse Document Frequency)

TFIDF는 각 문서에서 키워드의 중요도를 계산하는 방법으로 TFIDF를 이용한 특정 문서 T에서 키워드 di의 중요도 vi는 di가 T에서 출현한 빈도 tf(di)에 비례하고 그 키워드가 다른 문서에서 출현하는 빈도 df(di)에 반비례한다[2][3]. 그러나 키워드의 빈도는 문서의 길이에 비례하는 경향을 나타내므로 이를 보정하기 위해 최대값을 이용한 정규화와 코사인 정규화를 이용한다.

2.2 연관규칙(Association Rule)

연관규칙이란 대표적인 Data Mining의 기법 중 하나로 항목들의 집합으로 이루어진 트랙잭션을 분석하여 각 항목간의 연관성을 파악하는 기법으로 장바구니 분석, 교차판매전략, 사용자 접근 패턴 분석 등의 분야에서 이용되고 있다[4][5].

분석의 대상이 되는 항목들의 집합을 $I = \{i_1, i_2, i_3, ..., i_m\}$으로 정의할 때 트랜잭션 T는 I의 부분집합($T \subset I$)으로 정의된다. 이러한 트랜잭션들의 집합을 D라 할 때 각 트랜잭션을 구분하는 고유의 트랙잭션 번호를 TID라 정의한다. 트랜잭션 T가 항목집합 X의 모든 항목을 포함할 때 T가 X를 지지한다고 하며, X의 지지도 supp(X)는 X를 지지하는 D에 있는 모든 트랜잭션의 개수를 의미한다. 만약 주어진 최소 지지도 S_{min}에 대하여 $supp(X) \geq S_{min}$이라면 항목들의 집합 X를 빈발항목집합이라 한다. 연관규칙은 빈발항목집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙이다. 연관 규칙은 일반적으로 $R : X \Rightarrow Y$의 형태로 나타내며 X와 Y는 각각 I의 부분집합이며 서로 간의 공통 원소는 존재하지 않는다는 특성을 가진다[4][5].

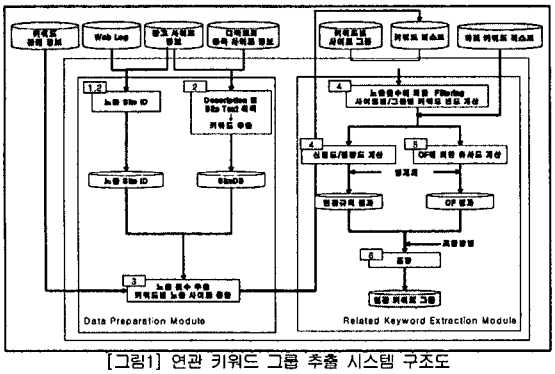
연관규칙을 타당성을 검증하는 척도로는 지지도(Support), 신뢰도(Confidence), 개선도(Lift)등을 이용한다. 일반적으로 최소지지도와 최소신뢰도를 만족하는 경우에만 규칙으로 생성하며 개선도는 생성된 규칙을 평가하는 항목으로 활용되고 있다.

2.3 협력적 추천(Collaborative Recommendation)

협력적 추천은 대표적인 개인화 추천 기법으로 목표 고객과 다른 고객들과의 유사도와 아이템에 대한 선호도를 고려하여 각 상품에 대한 선호도를 예측하고 고객이 구매하지 않은 상품에 대해 선호도 예측값이 큰 상품들을 추천하게 된다[6]. 협력적 추천은 고객간의 유사도를 구하는 과정과 선호도 예측값을 구하는 과정으로 나뉘게 된다. 일반적으로 고객 간의 유사도는 코사인 유사도나 피어슨 상관계수를 이용한다[7].

3. 연관 키워드 그룹 추출 시스템

본 논문에서 제안하는 연관 키워드 그룹 추출 시스템은 원본 Data를 입력받아 이를 분석에 필요한 Format으로 가공하여 주는 Data Preparation Module과 입력 Data를 분석하여 특정 키워드의 연관 키워드들을 추출하고 그 결과를 보여주고 저장하는 기능을 담당하는 Related Keyword Extraction Module로 구성되어 있다. [그림1]은 연관 키워드 그룹 추출 시스템의 구조도이다.



[그림1] 연관 키워드 그룹 추출 시스템 구조도

3.1 1단계 : Web Log로부터 노출 사이트 ID 리스트 추출

검색엔진의 Web Log를 입력으로 하여 사용자가 입력한 검색어와 검색어에 의해 노출된 사이트 ID를 추출한다. Web Log에는 검색 날짜, 시간, 검색 종류 등의 정보가 포함되어 있으므로 분석 작업에 필요한 검색어와 노출 사이트 ID만을 추출한다.

3.2 2단계 : 노출 사이트 ID 리스트 및 SiteDB의 형성

2단계에서는 검색엔진의 디렉토리에 등록된 사이트와 검색엔진의 키워드 샵에서 키워드를 구매한 광고 사이트 정보를 입력으로 받아 노출 사이트 ID 리스트와 SiteDB를 형성한다. SiteDB란 사이트 설명 문구인 Description과 텍스트에서 추출된 키워드 리스트를 말한다.

3.3 3단계 : 키워드별 사이트 그룹과 키워드 리스트 생성

3단계에서는 키워드 리스트와 키워드별 사이트 그룹을 생성한다. 키워드 리스트는 타겟 키워드 또는 연관 키워드 후보가 될 만한 키워드의 리스트로 키워드, 노출횟수, 판매횟수로 구성된다. 키워드별 사이트 그룹은 검색 키워드(타겟 키워드 후보)에 의해 노출된 사이트에서 추출된 키워드들의 리스트이다.

3.4 4단계 : 키워드별 사이트 그룹의 중요 키워드 추출

4단계는 Data Preparation Module에서 생성된 키워드 리스트와 키워드별 사이트 그룹을 입력으로 받아 각 키워드의 사이트별/그룹별 출현빈도를 계산하고 TFIDF에 의한 중요 키워드를 추출하며 이는 내부적으로 다음과 같은 두 가지 과정을 거치게 된다.

3.4.1 타겟 키워드 후보와 연관 키워드 후보의 생성

연관 키워드의 생성은 키워드의 판매 가능성을 보장해야 하므로 키워드의 노출횟수가 최소노출횟수보다 작은 키워드를 키워드 리스트에서 삭제하고 그 키워드에 대한 사이트 그룹을 삭제한다. 이 때 키워드 리스트에 남는 키워드는 연관 키워드 후보가 되며 키워드별 사이트 그룹에 남는 사이트 그룹의 검색 키워드는 타겟 키워드 후보가 된다. 또한 사이트 그룹에 속한 사이트 수가 최소사이트 수보다 작은 경우에 대해서는 타겟 키워드 후보에서 삭제한다.

3.4.2 수정된 TFIDF에 의한 중요 키워드 추출

연관 키워드 그룹은 타겟 키워드와의 연관성이 높은 키워드만을 추출해 내야 한다. 따라서 수정된 TFIDF를 이용하여 각 사이트의 연관 키워드 후보 중 해당 사이트를 대표할 수 있는 중요 키워드를 추출해 낸다. 연관 키워드 그룹에서의 중요 키워드는 한 사이트에서 많이 출현하고 같은 사이트 그룹의 여러 문서에서 출현하는 반면, 해당 키워드가 출현한 사이트 그룹의 수가 작은 키워드이다. 따라서 [식1]과 같이 수정된 TFIDF를 사용한다. [식2]은 수정된 TFIDF에서 df를 계산하는 식이며 [식3]는 그룹 출현 빈도를 계산하여 반영하는 식이다.

$$v_i = \frac{(0.5 + 0.5\frac{tf(d_i)}{tf_{max}})(\frac{n_G}{n_G - df_G(d_i)})(\frac{gf(d_i)}{N_G})}{\sqrt{\sum_{d_j \in T}((0.5 + 0.5\frac{tf(d_j)}{tf_{max}})^2(\frac{n_G}{n_G - df_G(d_j)})^2(\frac{gf(d_j)}{N_G})^2)}}$$

T : 임의의 사이트 G : T가 속한 Site 그룹
d_i : T에서 출현한 키워드 v_i : T에서의 vi의 중요도
tf_{max} : T에서 가장 큰 tf의 값 $tf(d_i)$: T에서 d_i가 출현한 빈도
$df_G(d_i)$: G에서 d_i가 출현한 사이트의 빈도
n_G : G에 속한 사이트의 수 N_G : 전체 사이트 그룹의 수
$gf(d_i)$: d_i가 출현한 사이트 그룹의 수

[식1] 수정된 TFIDF의 식

$$\frac{N_G}{N_G - df_G(d_i)}$$

[식2] 문서 출현 빈도의 반영

$$\frac{gf(d_i)}{N_G}$$

[식3] 그룹 출현 빈도의 반영

[식2]에서 $df_G(d_i)$의 값이 커질수록 [식2]의 값이 커져 같은 사이트 그룹의 여러 문서에서 출현한 키워드의 중요도를 높여 주며 [식3]에서 $gf(d_i)$의 값이 커질수록 [식3]의 값이 작아져 여러 사이트 그룹에서 출현한 키워드의 중요도를 감소시켜 준다.

4단계에서 생성되는 타겟 키워드별 사이트 그룹의 중요 키워드 리스트는 내부적으로 사용되는 테이블로 그 구성은 [표1]와 같다.

[표 1] 타겟 키워드별 사이트그룹의 중요 키워드 리스트의 구성

인덱스	타겟 키워드 후보	노출 횟수	판매 횟수	전체 출현 사이트수	출현 그룹수	연관 키워드 후보 리스트		
						연관 키워드 후보의 인덱스	그룹내 출현 사이트수	...
...	...	...	...	...	...	...	...	...

3.5 5단계 : 연관 키워드 후보 생성

5단계는 연관규칙에 의한 연관 키워드 후보와 협력적 추천에 의한 연관 키워드 후보를 생성하는 단계이다. 연관 키워드 후보를 생성하는 방법은 연관도/유사도가 사용자가 입력한 임계치보다 큰 키워드만을 선택하거나 연관도/유사도가 큰 상위 N개를 선택하는 방법이 있다.

3.5.1 연관규칙에 의한 연관 키워드 후보 생성

본 논문의 연관규칙 탐사에서는 사용자가 연관 키워드 그룹을 생성하고자 하는 타겟 키워드별 사이트 그룹에 속한 각각의 사이트들의 중요 키워드 리스트를 하나의 트랜잭션으로 구성하며 전항목과 후항목을 모두 단일 항목으로 설정하되 전항목은 타겟 키워드, 후항목은 연관 키워드 후보로 한다. 또한 본 논문에서 연관 키워드 후보가 타겟 키워드에 대한 사이트 그룹의 여러 문서에서 골고루 출현하면서, 연관 키워드 후보가 출현한 사이트 그룹의 수가 작은 경우 타겟 키워드와 연관 키워드 후보간의 연관성이 높은 것으로 간주한다. 따라서 연관 키워드 후보가 타겟 키워드에 대한 사이트 그룹의 얼마나 많은 사이트에서 출현하는지를 나타내는 척도로 일반적인 연관규칙 탐사에서의 신뢰도를 이용하며, 연관 키워드 후보가 얼마나 많은 사이트 그룹에서 출현하는지를 나타내는 척도로 집중도라는

새로운 개념을 도입하여 사용한다. 그리고 타겟 키워드와 연관 키워드 후보간의 연관성을 나타내는 척도로 연관도를 사용하는데 이는 신뢰도와 집중도를 곱한 값을 이용한다. 본 논문에서는 [식 4]에 의해 신뢰도를 계산하며 [식 5]에 의해 집중도를 계산한다.

$$Confidence(d \rightarrow r) = \frac{df_G(r)}{n_G}$$

[식4] 신뢰도 계산식

$$Concentration(d \rightarrow r) = \frac{df_G(r)}{df_{all}(r)} \times (1 - \frac{gf(r)}{N_G})$$

[식5] 집중도 계산식

d : 타겟 키워드
r : 연관 키워드 후보
G : d에 대한 사이트 그룹
n_G : G에 속한 사이트의 수
$df_G(r)$: G중에서 r이 출현한 사이트의 수
$df_{all}(r)$: 전체 사이트 중 r이 출현한 사이트의 수
gf(r) : r이 출현한 사이트 그룹의 수
N_G : 전체 사이트 그룹의 수

[식4]에서 $df_G(r)$의 값이 커질수록 R : d→r의 신뢰도 값은 커지고 이는 연관 키워드 후보 r이 사이트 그룹 G의 여러 문서에서 출현함을 의미하며 [식5]에서 $df_{All}(r)$이 작아지고 gf(r)이 작아질수록 R : d→r의 집중도의 값은 커지고 이는 연관 키워드 후보 r이 사이트 그룹 G에서 집중적으로 출현함을 의미한다. 신뢰도와 집중도는 항상 0과 1사이의 값을 가지게 되며 이 둘을 곱한 값은 연관도 역시 0과 1사이의 값을 가지게 된다.

3.5.2 협력적 추천에 의한 연관 키워드 후보 생성

본 본문에서 협력적 추천에 의한 연관 키워드 후보의 생성은 각 타겟 사이트 그룹에서의 키워드의 사이트 출현 빈도를 벡터로 표현하고 이 벡터간의 코사인 값을 계산하여 각 키워드간의 유사도를 계산한다. 즉, 각 키워드가 특정 사이트 그룹의 몇 개의 사이트에서 출현하였는지를 계산하고 그 빈도로 벡터를 표현하는 것이다. 이 때 사이트 출현 빈도는 그룹에 속한 사이트 갯이에 비례하는 경향을 나타내므로 각 그룹에서의 최대 출현 빈도를 이용하여 정규화한다.

3.6 6단계 : 연관 키워드 그룹 생성

6단계에서는 연관규칙에 의해 생성된 연관 키워드 후보와 협력적 추천에 의해 생성된 연관 키워드 후보를 조합(①양 결과에 교집합을 이용하거나 ②합집합을 이용하거나 ③연관규칙에 의한 연관 키워드 후보 집합만을 이용)하여 최종적인 연관 키워드 그룹을 생성한다. 키워드의 판매 가능성을 판단하는 기준으로 키워드의 노출횟수와 판매횟수를 곱한 값을 이용하며 노출횟수와 판매횟수는 각각의 최대값에 의해 정규화한다. 6단계에서 생성되는 연관 키워드 그룹에는 타겟 키워드, 타겟 키워드의 노출횟수와 판매횟수, 연관키워드, 연관키워드의 노출횟수와 판매횟수, 연관키워드의 정규화된 노출횟수와 판매횟수의 값, 정규화된 노출횟수와 판매횟수의 곱(순위값)을 포함한다.

4. 구현 및 실험

본 논문의 실험 Data는 2003년 9월 6일~9월 12일까지의 네이버 통합 검색의 노출 로그, 2003년 10월 14일 현재 네이버 등록 사이트, 2003년 9월 6일 ~10월 5일까지의 키워드 구매 사이트, 2003년 8월 네이버 키워드 샵에서의 키워드 판매 현황을 이용하였으며 실험 방법은 다음의 [표2]와 같다.

[표 2] 실험 방법

최소노출횟수	100회
최소사이트수	10
연관 키워드 후보 생성의 임계치	상위 5개
타겟 키워드	결혼,대출,이사,여행사,창업

다음의 [표3]은 연관 규칙에 의한 연관 키워드 후보 집합과 협력적 추천에 의한 연관 키워드 후보 집합의 교집합을 이용하여 생성한 연관 키워드 그룹을 보여준다.

[표 3] 교집합을 이용한 연관 키워드 그룹

타겟 키워드	노출 횟수	판매 횟수	연관 키워드	노출 횟수	판매 횟수	순위값
결혼	492	5	웨딩	298	10	0.1895
결혼	492	5	예식장	134	8	0.1509
결혼	492	5	웨딩드레스	174	5	0.0949
결혼	492	5	국제결혼	161	5	0.0948
대출	753	15	인터넷대출	870	3	0.0620
대출	753	15	학자금대출	125	3	0.0571
대출	753	15	연체대납	1585	2	0.0480
이사	321	16	포장이사	176	15	0.2824
이사	321	16	용달	777	11	0.2114
이사	321	16	kgb	174	0	0.0012
여행사	1116	5	신혼여행	316	16	0.3021
여행사	1116	5	허니문	136	16	0.3009
창업	944	10	프랜차이즈	133	4	0.0759
창업	944	10	부업	311	1	0.0208
창업	944	10	소호	108	0	0.0007

5. 결론 및 향후 과제

본 논문은 키워드 샵에서 상품 추천을 위한 연관 키워드 그룹 추출 기법에 대한 것으로, 검색엔진의 Web Log, 디렉토리 등록 사이트 정보, 광고 사이트 정보, 키워드 판매 정보를 이용하여 타겟 키워드별로 사이트 그룹을 형성한 후 각 사이트를 대표하는 중요 키워드를 선별하고 연관규칙과 협력적 추천에 의한 연관 키워드 후보 집합을 조합하는 방법으로 연관 키워드 그룹을 추출하는 방법을 제안하였다. 그러나 본 논문에서는 연관 키워드 그룹을 생성하는 단계까지 만을 제안하고 있다. 생성된 연관 키워드를 보다 효율적으로 활용하기 위해서는 모든 구매자에게 동일한 키워드를 제시하기 보다는 구매자의 운영 사이트에서 중요 키워드가 될 수 있는 키워드를 제시해야 한다. 따라서 향후 연관 키워드 그룹에 속한 여러 키워드 중에서 향후 구매자의 운영 사이트에 특성에 따라 최적화된 키워드를 추천하는 방안에 대한 연구가 필요하다.

6. 참고문헌

[1] 김재휘, 김지호, 김용환, 인터넷 검색 사이트의 키워드 광고 효과 연구, 광고학 연구, 일반 제13권 4호, p.91-109. 2002

[2] 백혜정, 사용자 관심도를 이용한 웹 에이전트, 숭실대학교 석사 학위 논문, 1997

[3] Thorsten Joachims, A Probabilistic Analysis of the Rocchino Algorithm with TFIDF for Text Categorization, 1996

[4] 한영우, 고속의 연관 규칙 마이닝을 위한 효율적 공간 압축 및 탐사 기법, 숭실대학교 석사 학위 논문, 2002

[5] Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases. SIGMOD'93, Washington, D.C. (1993) 207-216

[6] Shardanand, U., Maes, P., Social Information Filtering : Algorithms for Automating Word of Mouth. In proceedings of the Computer Human Interaction, 1995

[7] Sarwar, B., Karypis, G., Konstan, J., Riedel, J., Analysis of Recommendation Algorithms for E-Commerc" GroupLens Research Group and Army HPC Research Center Department of Computer Science and Engineering University of Minnesota Minneapolis, MN 55455.