문서 클러스터링을 이용한 문맥 광고 시스템

이 동 광[†] · 강 인 호[††] · 안 동 언[†††]

요 약

본 연구에서는 문서 클러스터링을 이용하여 동음 이의어와 핵심단어 선정 실패로 인해 발생하는 자동 광고 시스템의 오류를 해결하는 광고 키워드 추출방식을 제안한다. 먼저 대규모 뉴스기사를 대상으로 유사한 내용을 가지며 동일한 광고 키워드와 연관이 있는 기사들을 자동으로 분류하여 광고 키워드에 대한 문맥 정보를 구축한다. 또한 광고 대상물에 대한 광고주의 요약 정보나 광고 대상 웹페이지를 분석하여 광고 키워드에 대한 문맥 정보를 추출하는 방식을 보인다. 이렇게 구축된 문서 분류와 광고 키워드용 문맥 정보를 이용하여 광고 대상 문서가 속한 문서 분류를 추정하여 단어들의 의미적인 애매성을 해결하고, 추정한 문서 분류와 관련 있으면서 문맥적으로 중요성을 가지는 핵심 단어들을 선정하여 광고 키워드를 추출한다. 상용 광고 시스템과의 비교 분석 결과 신문 기사나 일반 블로그를 대상으로 최소 21%의 성능 향상을 얻었다.

키워드 : 광고 키워드 추출, 문서 클러스터링, 문맥 기반 광고

Contextual Advertisement System based on Document Clustering

DongKwang Lee[†], In-Ho Kang[††], Dongun An[†††]

ABSTRACT

In this paper, an advertisement-keyword finding method using document clustering is proposed to solve problems by ambiguous words and incorrect identification of main keywords. News articles that have similar contents and the same advertisement-keywords are clustered to construct the contextual information of advertisement-keywords. In addition to news articles, the web page and summary of a product are also used to construct the contextual information. The given document is classified as one of the news article clusters, and then cluster-relevant advertisement-keywords are used to identify keywords in the document. We could achieve 21% precision improvement by our proposed method.

Key Words : Finding Advertisement-Keyword, Document Clustering, Contextual Advertisement

1. 서 론

사용자가 보고 있는 웹 문서의 내용을 기반으로 사용자의 관심 영역을 유추하고 거기에 적합한 광고를 선정하는 연구의 필요성이 대두되고 있다. 이러한 연구는 문서에서 사용된 중요 단어들을 찾아내어[1,2,3] 거기에 적합한 광고를 찾는 문제로 설명된다[4]. 그러나 문서의 문맥적 특성을 고려하지 않은 기존의 단순 키워드 매칭 방식은 정확한 단어를 찾는데 문제점이 발생한다. 예를 들어 '걸레 만두'라는 '만두'에 대한 설명에서 '걸레'가 중요 단어로 선택된 경우 문서의 주된 내용인 '만두'에 대한 설명과 상관없는 '걸레'(청소용품)를 판매하는 쇼핑몰을 광고하는 문제점이 발생한다.

문서의 문맥적 특성을 파악하기 위해서 문서의 분류를 생각할 수 있다[5,6,7]. 그러나 기존의 문서 분류(Document Classification)에서 사용하는 스포츠, 정치, 경제와 같은 문서의 분류는 광고 키워드를 정하는데 충분하지 못하다. 동일한 분류의 문서도 서로 다른 내용을 포함할 수 있기 때문에 문서의 내용에 따른 광고 키워드 추출 방법이 필요하다. 동일한 다이어트 영역의 게시물이라도 내용에 따라서 운동을 통한 다이어트, 과일을 통한 다이어트, 또는 약물을 통한 다이어트를 광고 키워드로 추출할 필요가 있다. 즉 많은 사람들이 관심 있으면서 광고에 유용한 형태의 문서 분류가 필요하다. 예를 들어 "부동산 정책, 대출규제 - 세제완화 어떻게 되나?"라는 제목을 가진 기사의 경우 부동산, 대출, 세금, 정부 정책 등 매우 많은 주제들로 이루어진다. 이때 이 기사에 적합한 광고를 찾기 위해서는 다양한 주제 중에서 어떤 주제로 분류를 선택할지에 관한 기준이 필요하다. 이러한 기준이 되는 것을 광고용 문서 분류셋이라고 본 연구에서는 정의한다. 본 연구에서는 대규모 뉴스기사를 대상으로 클러스터링[8,9]을 수행하여 광고용 문서 분류셋을 생성하고, 이를 활용해서 문서의 문맥 정보를 분석하는 광고

† 준 회 원 : 전북대학교 컴퓨터 공학과 박사과정
†† 정 회 원 : CMU/LTI 연구소 박사후과정
††† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2007년 5월 16일, 심사완료 : 2007년 10월 28일

DOI: 10.3745/KIPSTB.2008.15-B.1.73

키워드 추출 방법을 제안한다. 광고 키워드 추출은 웹 검색 기용 광고 시스템[1]을 기반으로 사용자가 보고 있는 문서에 대해서도 검색기에 사용자가 질의어를 입력한 것과 같이 가장 적합한 광고 키워드를 선정하는 것을 목적으로 한다. 광고 키워드에 매핑되는 광고물은 실시간으로 변경될 수 있다. 그러나 광고 키워드 뿐만 아니라 광고 키워드에 딸린 실제 광고물의 내용을 알 수 있다면 보다 좋은 광고 키워드를 선정할 수 있다. 본 연구에서는 광고 키워드에 관한 정보와 광고 대상물에 대한 정보를 각각 활용하여 두 정보의 유용함을 실험을 통하여 살펴본다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 광고 키워드 추출 방식에 대해서 간략하게 정리한다. 3장에서는 대상 문서 집합에서 문서 분류셋 추출을 위한 문서의 주제별 클러스터링 방식을 설명하며, 4장에서는 주어진 문서에 적합한 광고 키워드 추출 방식을 보인다. 그리고 제안하는 모델의 성능과 실제 서비스에서의 영향에 대한 내용을 5장에서 설명한다.

2. 관련 연구

2.1 검색 키워드 기반 광고 추출

인터넷을 사용하는데 있어 검색기의 비중이 커짐에 따라, 많은 기업들은 검색기의 상위 결과에 자신의 홈페이지가 나타나게 하거나 검색 결과 주위 화면에 광고 문구를 삽입하는 형태로 제품을 홍보하고 있다. 즉 사용자가 검색에 사용한 질의를 기반으로 사용자의 관심 영역을 추정하고 이에 적합한 광고물을 보여주는 것이다. 그러나 이러한 검색 기반의 광고 방식으로는 사용자의 인터넷 활용에 있어서 많은 시간 동안 홍보하기에는 제약점이 있다.

<표 1>은 KoreanClick[2]의 자료를 활용하여 2006년 4분기 PV(Page View)의 분포를 서브 도메인 별로 나타낸 표이다. <표 1>을 통해 알 수 있듯이 검색으로 인한 Page View는 평균 6%정도 밖에 안된다. 나머지는 Community, News등의 게시물을 통한 Page View이다. 따라서 보다 많은 광고 노출을 제공하기 위해서는 검색기에 대한 광고 방식에서 일반 컨텐츠에 대한 광고 방식으로의 확장이 필요하다.

〈표 1〉 2006년 Q4 포털별 PAGE VIEW - KOREANCLICK

Daum			Naver			Nate		
Sub Domain	PV (백만)	%	Sub Domain	PV (백만)	%	Sub Domain	PV (백만)	%
café	27,418	45.9	café	11,382	15.2	cyworld	62,082	85.4
mail	8,264	13.8	news	9,981	13.3	SMS	1,895	2.6
www	5,974	10.0	search	10,019	13.4	news	1,252	1.7
news	2,573	4.3	www	9,410	12.5	bbs	1,060	1.5
search	2,024	3.4	blog	6,253	8.3	mail	810	1.1
dnshop	1,140	1.9	kin	3,928	5.2	www	660	0.9
agora	1,074	1.8	shopping	2,065	2.8	search	206	0.3
ETC	11,230	18.8	ETC	21,969	29.3	ETC	4,723	6.5

1) 본 연구에서 기반하는 overture 광고시스템에서는 광고를 위한 키워드 셋을 마련한 뒤, 광고를 원하는 업체에게 판매한다. 사용자의 검색 질의어와 유사한 광고 키워드가 존재할 경우 해당 광고를 보낸다.

2) http://www.koreanclick.com

2.2 채널 정보 기반 광고 추출

대형 포털 사이트의 블로그, 게시판, 뉴스 등에 대해서 매체가 속한 영역(채널)의 특성에 따라 광고 키워드를 추출하고, 광고 키워드에 해당하는 광고를 노출하는 방식이다. 예를 들면 다이어트 관련 게시판에 광고 키워드 '다이어트'를, 경제 관련 뉴스에는 '증시', '부동산' 등을 고정적으로 광고하면서, 사용자의 광고물 선택율, 시간적, 계절적 변화 등을 고려하여 그때 그때 광고 키워드를 변경하는 방법이다. 하지만 동일한 채널의 문서도 서로 다른 내용을 포함할 수 있기 때문에 문서 내용에 따른 광고 키워드 추출 방식이 필요하다. 예를 들어 다양한 다이어트 방법에 대한 질문과 답변이 있는 게시판에서 무조건 다이어트로 광고를 한다면, 사용자는 그 광고를 그저 항상 떠있는 메뉴와 같은 것으로 생각을 하게 될 것이며, 광고 효과도 그 만큼 떨어진다.

2.3 검색 기술 기반 광고 추출

검색 기술을 응용하여 문서 안에서 검색 광고 키워드를 추출하는 방식이다. 문서 안의 다양한 키워드를 제목에서의 단어 사용여부, 본문에서의 위치, 문장에서의 위치, 검색 광고에서의 단가, 검색 로그의 출현 빈도 등을 고려하여, 문서에서 검색 광고 키워드를 추출한다. 즉, 문서내의 여러 단어 중에서 어떤 단어를 검색어로 사용했을 때 해당 문서가 가장 상위에 랭크 될 수 있는지를 역추적해서 검색 키워드를 추출하여 광고를 하는 방법이다[4]. 문서에 나타나는 단어 중에서 사용자의 검색 질의로 많이 나타나는 단어를 선호하거나[10] 문서와 가장 유사한 광고 문구를 가진 광고물을 선택한다. 이러한 방식의 대표적인 제품으로 구글의 AdSense[3]를 들 수 있다. AdSense는 문서 안에서 광고 키워드를 찾아내는 방법으로, 단어의 빈도수, 사용위치, 광고 키워드의 단가 등 여러 가지 요소를 이용하여, 가장 유사한 광고물을 제공한다.

본 연구에서는 문서를 구성하는 단어들의 매칭을 이용한 문서와 광고 키워드간의 직접적인 추출 방식과는 달리 문서 전체 내용을 이용한 관련 분류를 선택하고 거기에 따른 차별된 단어 매칭을 이용하는 이단계 방식을 보인다.

3. 문서 수집 및 문서 클러스터링을 이용한 광고용 문서 분류셋 생성

본 장에서는 임의의 문서를 분류할 수 있는 광고용 문서 분류셋 작성에 대해서 설명한다. 문서 분류셋을 작성하기 위해서 대표 문서의 수집, 수집된 문서의 클러스터링, 생성된 클러스터의 검증 그리고 최종적으로 선택된 클러스터(문서 분류셋)에 광고 키워드 할당의 4가지 과정을 수행한다.

3.1 문서 수집 및 분석

수집되는 문서들이 특정 분야에 치중되지 않도록 양질의

3) http://www.google.com/adsense/

문서를 수집하는 방법이 필요하다. 예를 들어 여름 휴가철 기간의 문서일 경우 "여행, 휴가, 숙박업소 등"의 내용에 치중될 수 있으며, 연말의 경우 "각종 시상식이나, 연말정산, 송년모임 등"의 내용에 치중될 수 있다. 따라서 클러스터를 생성할때 사용하는 문서는 기간적으로나 출처에서 한정되지 않도록 대규모의 문서를 수집한다. 이를 위해 각 포털 사이트에서 서비스 되고 있는 뉴스기사를 이용한다. 이는 정제된 다양한 분야의 문서를 좀더 쉽게, 또한 대량으로 수집하기가 매우 쉽기 때문이다. 또한 대상 문서를 뉴스로 한정함으로써 카테고리 정보도 이용할 수 있는데, 뉴스는 "정치, 경제, 사회, 문화, IT, 스포츠, 연예" 정도의 대분류가 기본적으로 분류되며, 다시 대분류 경제는 "생활경제, 증권, 금융, 부동산, 산업, 국제" 정도의 중분류로 나뉘어 진다. 이러한 분류정보는 대부분의 뉴스에서 이루어지는 분류이며, 본 연구에서는 네이버[4]의 대분류와 중분류 정보를 수집 때부터 정보를 이용하였다. 이러한 분류정보는 다음 단계인 문서 클러스터링 단계에서도 이용하였다. 문서는 수집 후 메뉴나 다른 기사로의 링크 정보 등은 모두 제거 하였으며, 사이트에서의 분류 정보, 제목, 본문만을 이용하였다.

3.2 문서 클러스터링

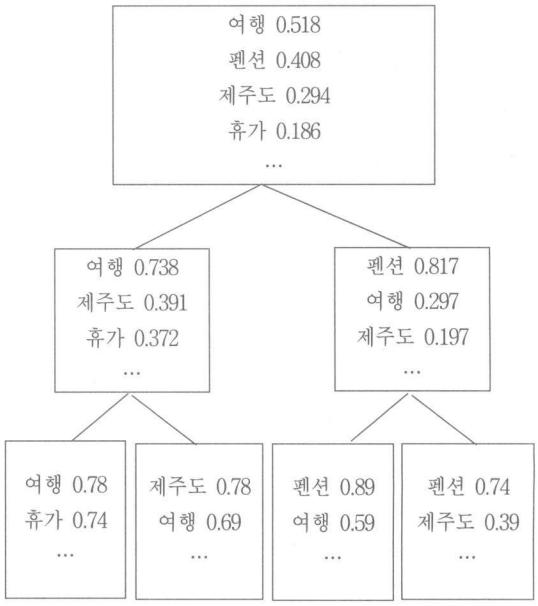
수집된 문서가 정제된 양질의 데이터이긴 하지만 클러스터링을 하기 위해서는 약간의 가공이 필요하다. 문서에 나타난 띄어 쓰기 오류나 오자를 수정하면서 중요 단어를 추출하기 위해서 형태소 분석기를 이용한다. 분석기의 결과 중 숫자, 요일, 및 기호 등은 제거한다. 숫자 정보가 정보 검색에서는 매우 중요한 정보로 이용되지만, 숫자에 의해서 기사의 분류가 정해지는 경우는 매우 적다. 13초, 30초 또는 10만원, 10억원 등은 정보검색의 입장에서 본다면 매우 큰 차이를 나타내는 정보이지만, 주제라는 입장에서 본다면 시간이나 경제적 가치를 나타내는 하나의 단위에 불과하다. 즉 이러한 단위가 바뀐다고 해서 주제가 바뀌는 것은 아니다.

추출된 단어들의 가중치는 *tf*(term frequency)와 *df*(document frequency)를 이용하는, 정보검색에서 가장 많이 사용되는 단어 가중치 기법인 TFIDF를 이용하였다[11]. 오자나 유용하지 않은 표현을 제거하기 위해서 *df*가 낮은 단어를 제거한다[5]. 단어의 가중치는 문서에서의 위치에 따라서 수식 1과 같은 차별된 가중치를 가진다. 대상 문서가 뉴스기사이기 때문에 제목에서 사용된 단어는 본문을 함축해서 표현할 수 있는 매우 중요한 단어들로 이루어진다. 수식 1에서의 α는 제목에서 나타난 단어들에 대한 가중치 배수이다. 예를 들어 α가 .8일 때 A라는 단어가 제목에서 1번, 본문에서 6번 나타났다면 A라는 단어의 TF값은 (0.2*6 + 0.8*1)가 되어 TF=2가 된다.

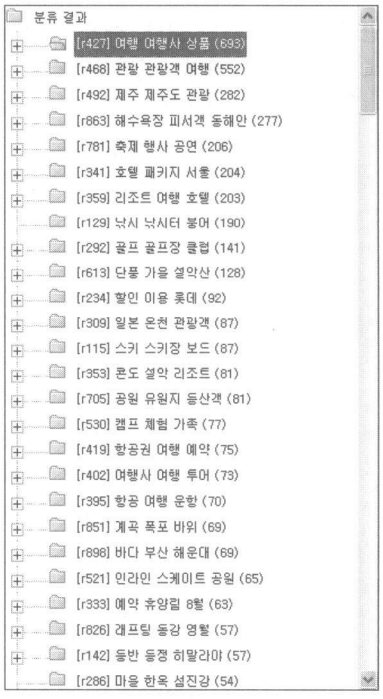
$$w(x) = \frac{\alpha \cdot tf_t(x) + (1-\alpha) \cdot tf_c(x)}{df} \qquad \text{(수식 1)}$$

$tf_t(x)$= term frequency in a title
$tf_c(x)$= term frequency in a content

각 문서는 문서에서 선별된 단어와 그 단어의 가중치로 이루어진 벡터로 표현된다. 구축된 벡터들은 cosine similarity를 이용한 계층적인 클러스터링 방식을 사용하여 문서를 분류한다[9]. 즉, 가장 유사한 벡터로 선택된 두 벡터는 단어의 가중치를 평균을 내어서 새로운 부모 벡터 1개를 생성하며, 부모 벡터에 속한 두 벡터는 더 이상 다른 벡터들과 비교를 하지 않는다 (그림 1).



(그림 1) 이진 트리 형태의 Hierarchical 클러스터



(그림 2) 생성된 문서 클러스터 예

4) http://www.naver.com
5) 본 연구에서는 df가 10 미만인 단어를 제거한다

이 때 새로운 벡터는 각 단어가 가지는 가중치를 기준으로 최대 100개의 대표 단어를 포함할 수 있으며 가중치가 낮은 단어는 버린다. 두 벡터의 결합 여부는 cosine similarity를 통해서 얻어낸 유사도와 두 벡터의 원래 뉴스 문서가 속한 분야 정보를 이용한다[9]. 서로 다른 분야의 문서들의 결합을 위해서는 높은 유사도를 요구하며, 문서들이 속한 분야 정보에 따라서 차별적인 유사도 threshold를 사용한다. 이러한 방법을 지속적으로 수행하여 이진 트리 형태의 클러스터를 만든다. (그림 2)은 생성된 클러스터 중 '관광'과 관련된 클러스터들과 그 대표어의 일부를 보인다.

3.3 클러스터(문서 분류셋)의 검증

클러스터 후보들에 대해서 클러스터링이 이루어지게 된 대표 단어들의 특성에 따라 광고에 적합한 클러스터인지를 검증한다. 검증에 실패한 경우 이진 트리 형태의 클러스터 구조에서 하위의 두 개의 클러스터로 나뉘어지며 다시 검증을 수행한다. 적합한 클러스터를 판별하기 위해서 다음과 같은 3가지 조건을 이용한다.

- 대표 단어의 클러스터 표현력
- 클러스터의 응집력
- 생성된 클러스터의 문서 변별력

첫번째 조건은 클러스터를 구성하는 대표 단어들이 클러스터에 속한 문서들을 잘 설명하는 대표 단어여야 한다. 즉, 대표 단어로 선택된 단어들이 클러스터에 속하는 문서들을 잘 표현할 수 있어야 한다. 이러한 조건이 필요한 경우는 특정 분야에서 자주 쓰이지는 않지만, 특정 문서 몇몇에서 중요하게 쓰임으로써 그 가중치 값이 매우 높을 때이다. 이 경우, 클러스터 후보의 다른 문서에서는 거의 사용되지 않지만 대표 단어로 선택될 수 있다. 예를 들어 여행에 관한 클러스터가 만들어졌다고 하면, '안면도, 꽃지' 등의 단어가 클러스터의 대표 단어로 정해지는 경우가 있는데, 이러한 단어는 특정 지역을 소개하는 문서에서 중요하게, 자주 등장함으로써 대표 단어로 선택이 된 경우이다. 만약 '안면도, 꽃지'가 해당 클러스터의 많은 문서에서 사용되는 표현이라면 태안반도의 여행지에 관한 내용으로 이루어진 클러스터일 가능성이 높지만, 극히 적은 문서에서만 존재한다면, 일반적인 여행에 관한 클러스터에 '안면도, 꽃지'의 단어가 남음으로써 휴양지와 여행정보간에 혼선을 일으키는 역할을 하게 된다. 그렇기 때문에 클러스터의 대표 단어들은 해당 클러스터에 속한 대부분의 문서에서 사용되는 표현력이 높은 단어로 구성되어야 한다. 본 연구에서는 각각의 대표 단어들이 클러스터에 속한 문서의 최소 20% 이상에서 나타나야 한다는 조건을 사용한다.

두번째 조건은 클러스터가 하나의 주제로 구성되어야 한다. 즉 다양한 주제를 가진 문서들로 구성된 클러스터의 경우 효과적인 광고 키워드를 할당하기 힘들다. 이러한 경우는 주로 문서의 내용 주제보다는 흔히 사용되는 표현이나 디자인이나 스타일을 위한 표현과 같은 외적인 표현에 의해서 클러스터링이 이루어진 경우이다. 클러스터에 포함된 주제의 개수를 추정하기 위해서 전체 대표 단어들이 가지는 가중치에서 매우 높은 가중치를 가지는 대표 단어들의 비를 이용한다. 매우 높은 가중치를 가지는 대표 단어들을 핵심 단어라고 정의하고 전체 대표 단어 개수에 비해 핵심 단어들의 개수가 많은 경우 다양한 주제로 구성된 클러스터로 간주한다. 본 연구에서는 가중치 순서로 정렬한 대표 단어 리스트에서 전체 가중치 합의 30% 이상을 가지게 되는 최소 개수의 상위 대표 단어들을 핵심 단어로 정의하며, 핵심 단어의 수는 전체 대표 단어 개수의 최대 5%를 넘지 않아야 한다는 조건을 사용한다.

세번째 조건은 클러스터가 문서들을 제대로 변별하고 있는지를 확인한다. 즉 문서와 가장 유사한 클러스터를 찾았을 경우, 자기가 속한 클러스터와 일치하는지를 살펴본다. 이때 가장 유사한 클러스터와 원래 속해있던 클러스터가 일치한 문서의 비를 이용해서 클러스터의 변별력을 평가한다. 일치하는 비가 떨어지는 클러스터의 대표 단어들을 보면 중요단어들이 뉴스기사를 쓴 기자의 이름이나, 신문사 또는 신문사의 저작권에 대한 기술로 클러스터링이 이루어진 경우 등이 많다. 만약 기사의 끝부분에 나타나는 기자의 이름이나 신문사명, 저작권에 관한 문구를 잘 제거할 경우 이러한 문제 현저히 줄어든다. 본 연구에서는 75%-85% 이상의 일치율을 보이는 클러스터를 선택한다. 일치율은 뉴스에 사용되는 표현에 특성에 따라서 상대적으로 한정적인 표현을 사용하는 스포츠 분야는 높은 값을 사용하고 다양한 표현이 가능한 생활문화는 약간 낮은 값을 사용한다.

3.4 광고 키워드 할당

구축된 문서 분류셋에 적합한 광고 키워드를 할당하기 위해, 실제 광고주가 입력한 광고 대상 사이트의 제목 및 설명문을 이용한다. 즉 각각의 광고물과 가장 유사한 문서 분류를 찾아 그 광고물에 해당하는 광고 키워드를 해당 문서 분류에 할당한다. 이때 광고 설명문에 나타나는 단어는 사전에 작성된 유의어를 이용해서 확장된다. '꽃배달'과 같은 광고 키워드는 생일, 이벤트, 남녀 교제에 관한 문서 분류에 매핑을 해야하지만, 광고물이 가지는 제목이나 요약 설명문에는 "꽃바구니, 꽃다발, 화환" 등의 단어만 기술 되어 있다. 따라서 이러한 설명문에는 '생일'이나 '백일' 같은 유의어를 추가하여 유사한 문서 분류를 찾는다. 유의어는 광고 키워드별로 관련 높으면서 많이 사용되는 표현을 작성한다. 이렇게 광고 키워드 유의어 표현과 광고물에 사용된 표현을 이용해서 각각의 문서 분류에 적합한 광고 키워드를 할당한다. 광고물과 유사 문서 분류를 찾기 위해서 광고물은 단어와 그 단어의 가중치로 구성된 벡터로 변환하고 문서 분류의 대표 단어들로 구성된 벡터와 cosine similarity를 이용해서 유사도를 계산한다. 수식 2는 광고물의 각 단어가 가지는 가중치를 계산하는 방법을 보인다. 여기서 $tf(x)$는 단어 x가 나타난 빈도를 나타내며 N은 광고물에 사용된 총 단어

수를 뜻한다.

$$q(x) = \frac{tf(x)}{N} \qquad \text{(수식 2)}$$

이때 동일한 광고 키워드가 동일 문서 분류에 두 번 이상 할당이 될 경우 유사도의 평균값을 그 광고 키워드와 문서 분류의 유사도로 사용한다. 유사도가 낮은 경우에는 키워드 할당을 하지 않는다. 이와 같은 키워드 할당을 통하여 각각의 문서 분류는 관련 광고 키워드와 그 키워드를 할당할 때 계산되었던 유사도를 가진다.

광고물과 유사한 문서 분류를 찾는 과정에 더불어, 각 광고 키워드에 연결된 광고물의 내용을 이용해서 광고 키워드의 대표 단어 벡터를 작성한다. 즉 동일한 광고 키워드를 가지는 광고물을 합쳐서 하나의 문서로 간주하고 그 문서를 이용해 가중치 벡터를 만들어 해당 광고 키워드의 대표 단어 벡터로 사용한다 (수식 1).

4. 광고 키워드 추출

광고용 문서 분류셋을 이용하여 사용자가 보고 있는 문서에 광고 키워드를 추출하는 방법을 보인다.

4.1 문서의 정규화

3.4절에서 사용되었던 문서 분류에 광고 키워드를 할당하기 위해 광고물을 처리하는 형태와 동일한 방법을 사용한다. 사용자가 보고 있는 문서는 형태소 분석을 수행하여 중요 단어를 추출한다. 추출된 각 단어의 *tf*와 전체 단어의 개수(N)를 이용하여 수식 2와 동일하게 가중치를 계산한다. 문서를 표현하는 단어와 그 단어의 가중치로 구성된 벡터는 cosine similarity를 이용해서 가장 유사한 문서 분류를 찾는다.

4.2 광고 키워드 추출

선택된 문서 분류에 매핑된 광고 키워드도 문서의 내용에 따라서 순위가 결정된다. 사용자가 보고 있는 문서가 여행과 관련된 분류에 매핑이 되었다고 하더라도 문서의 내용에 따라서 국내여행, 해외여행, 유럽여행 등의 적합한 광고 키워드는 달라진다. 따라서 문서의 분류가 결정된 후에도 광고 키워드와 문서간의 유사도를 비교하여 보다 좋은 광고 키워드를 추출해야 한다. 이를 위하여 가장 유사한 문서 분류를 찾기 위해 만들었던 문서의 단어 가중치 벡터와 광고 키워드의 가중치 벡터간의 유사도를 비교하여 보다 좋은 광고 키워드를 추출한다.

5. 실험

본 논문에서 제안하는 광고 추출 방법의 유용성을 AdSense와의 비교 실험을 통해서 보인다.

5.1 테스트셋 구축

실험대상 문서는 현재 구글의 AdSense가 서비스 되고 있는 조선일보의 뉴스를 대상으로 하였다. 실험 문서는 주식, 부동산, 산업, 스포츠의 4개 분야에 대하여 40건의 문서를 각각 수집하였다. 추출된 광고의 유용성은 5명의 평가자가 0, 1, 2, 3, 4 의 5단계로 점수를 주는 방식으로 평가하였다. 점수 0, 1, 2, 3, 4 의 기준은 다음과 같다.

0 : 본문 내용과 전혀 관련이 없는 광고물의 경우
1 : 추출한 3개의 광고 중에서 1개의 광고가 연관성이 있는 경우
2 : 추출한 3개의 광고 중에서 2개의 광고가 연관성이 있는 경우
3 : 추출한 3개의 광고 중에서 3개의 광고가 연관성이 있는 경우
4 : 모든 광고가 문서와 매우 높은·연관성이 있는 경우

5.2 실험 결과

구글의 AdSense외에 문서 검색 기술을 응용한 두 가지 방법을 추가로 비교한다. 즉 사용자가 보고 있는 문서와 광고 키워드를 대표하는 문서들을 비교하여 가장 유사한 광고 키워드를 사용하는 방법이다. 광고 키워드를 대표하는 문서는 광고물에 딸린 요약문을 이용하는 경우와 해당 웹사이트의 main page를 이용하는 두 가지 경우로 작성된다. 가중치는 수식 1을 이용한다.

<표 2>는 뉴스 문서를 대상으로 한 실험의 결과를 보인다. 평가는 4개 분야 각 40개의 문서를 5.1절에서 설명한 0~4까지의 평가 정도를 각 문서에 대하여 5명의 평가자가 평가한 값을 합한 후 평균을 구한 값이다. 전체적으로 문서 분류 이용, 요약문 이용, 사이트수집, AdSense의 순으로 좋은 결과가 나타났다.

문서 분류를 이용한 방법은 각 문서 분류의 대표 단어들이 100여개 정도로 꽤 많은 단어들로 구성됨으로써 광고를 매핑할 문서의 내용이 극히 짧지 않는 경우를 제외하고는 좋은 성능을 보였다. 광고 요약문은 광고하고자 하는 목적이 함축적으로 담겨 있어, 광고하고자 하는 관련 분야의 단어의 수가 작긴 하지만 변별력이 높은 유용한 단어인 반면에, 광고하고자 하는 페이지를 수집하여 비교한 경우는 상품을 판매하고자 하는 내용이 많아서 광고물과 관련 없는

〈표 2〉 뉴스 문서에 대한 광고 추출 평가 결과

분야	Ad-Sense	요약문 이용	사이트 수집	문서 분류	만점
주식	10	88.8	56	123.2	160
부동산	20	72.8	51.6	128	160
산업	11.8	81.2	42.8	122.4	160
스포츠	1.2	30	3.6	111.4	160

〈표 3〉 지식인 문서에 대한 광고 추출 평가 결과

분야	클릭 초이스	요약문 이용	사이트 수집	문서 분류	만점
지식인	92.2	98.8	28.6	112	168

단어들도 많아서 좋은 결과를 얻지 못하였다. 주로 나타나는 비관련단어는 결제, 상품, 세일, 가격 등등으로 온라인으로 광고를 한다면 흔히 나타날 수 있는 단어들이다.

실험 결과에 따라 구글 AdSense의 한글 문서에 대한 결과가 매우 좋지 않아 네이버가 지식인 서비스를 통해 서비스 중인 클릭 초이스[6]를 추가로 실험하였다. 네이버 지식인은 대규모의 사용자들이 게시판을 이용해서 다양한 질문을 올리고 그에 대한 답변을 다른 사용자들로 부터 얻는 서비스이다. 문서의 내용이 뉴스와 달리 정형화되지 않은 형태이며 오타와 비문이 상대적으로 많은 특징이 있다. 84개의 문서를 수집하여 평가를 수행하였다. <표 3>은 실험 결과를 보인다.

네이버 클릭 초이스와의 비교에서는 문서분류 이용, 요약문 이용, 네이버 클릭 초이스, 사이트 수집의 방법 순으로 좋은 결과가 나타났다. 두 실험 결과를 통하여 본 논문에서 제안하는 문서 분류를 이용하는 광고 추출 방법이 뉴스 및 일반 문서에도 유용함을 알 수 있다.

5.3 결과 분석

전체적인 결과를 종합해 보면 광고용 문서 분류를 이용한 방법이 단연 좋은 성능을 보였다. 문서 분류를 이용함으로써 뉴스기사를 클러스터 단위의 관련 분야로 나눌 수 있고, 뉴스의 관련분야별(클러스터 단위)로 적합한 광고 키워드를 매핑함으로써, 광고 키워드를 추출할 문서와 가장 유사한 클러스터를 찾는 것만으로도 좋은 광고 키워드를 추출할 수 있었다.

다음으로 좋은 성능을 보인 것은 요약문을 이용한 방법이다. 사이트 제목과 요약문을 광고주가 선정하고, 광고 대행사가 심사 관리함으로써 사이트 요약문 내용의 정확도가 높았으며, 사이트를 광고하고자 하는 단어들이 적절히 선정됨으로 인해 사이트의 핵심적인 내용이 함축적으로 표현되어 비록 짧은 문장이지만 좋은 결과를 얻을 수 있었다. 하지만 스포츠의 결과를 보면 요약문을 이용한 방법의 단점이 극명하게 나타난다. 야구에서는 주로 홈런, 안타, 타점, 삼진등의 용어가 자주 등장하며, 축구는 스트라이커, 센터링, 수문장, 골, 월드컵 등의 용어가 자주 등장한다. 이러한 스포츠 용어들이 스포츠 관련 광고주의 사이트나 요약문에는 거의 나타나지 않는다. 즉, 광고를 할 때의 용어와 문서를 작성할 때의 용어가 다르다는 것이다. 이러한 분석은 광고주 사이트를 수집한 결과에서도 잘 나타나 있다. 광고주의 사이트를 수집하여 기사에 광고를 매칭시켜본 결과 평균점수가 0.8로 거의 매칭이 되지 않았다. 기사는 내용을 중심으로 서술하지만, 광고를 상품을 위주로 서술을 한다. 이러한 차이 때문에 광고와 기사를 직접적으로 매칭을 시키는 것은 좋지 않은 결과를 냈다.

실험을 하면서 놀라운 점은 Google AdSense의 결과인데, 영어 문서와는 달리 한글 문서에 대해서는 좋은 결과를 보여주지 못했다. 주식분야에서는 금융관련 광고, 부동산분야에서는 부동산, 공인 중계사 등의 광고, 산업분야에서는 자격증 관련 학원 등의 광고, 스포츠분야에서는 판촉물 판매 광고가 노출되었으나, 그 정확도가 매우 낮아서 문서 내용별 광고 매핑이라기 보다는 사이트의 카테고리를 이용한 광고처럼 보였다. 부동산 영역에서는 주식보다는 좀더 좋은 결과가 나왔는데, 부동산 섹션의 기사에 노출되는 광고에는 부동산 담보대출에 관한 광고가 종종 노출되어 사람마다의 주관적 판단에 따라 넓은 의미에서 부동산 관련 광고라고 생각한 사람이 있어 좀더 좋은 결과를 나타낼 수 있었다.

네이버의 클릭초이스는 구글의 AdSense보다 좋은 결과를 나타내었다. 하지만 검색기반 광고 기법의 단점인 동음이의어의 문제가 종종 나타났으며 이러한 이유로 유사도 비교를 통한 광고 시스템보다는 못한 성능을 나타내었다. 실험자에 따라서 네이버의 클릭초이스에 대한 평가가 큰 차이를 보였는데, 본문 내용중 일부가 노출된 광고와 맞아도 좋은 점수를 준 경우와 전체적인 문서의 맥락을 봤을 때 맞는 광고인가를 판단해서 높은 점수를 주지 않은 경우로 나뉘었다. 예를 들어, 네이버 클릭초이스에서 쥬얼리에 따른 동음이의어 문제로 그룹 '쥬얼리' 관련 내용물에 보석을 판매하는 쇼핑몰 리스트가 나열되었다.

6. 결 론

인터넷의 발달로 웹에서 정보를 얻으려는 사람들에게 웹 검색은 필수 요소이다. 이로 인해 TV광고에서도 웹 검색을 통한 마케팅을 심심치 않게 발견할 수 있으며, 이제 검색광고는 발전의 단계를 넘어 완성단계에 접어들고 있다. 그러나 검색광고는 검색이 이루어져야만 광고를 노출할 수 있기 때문에 전체 웹 페이지뷰의 약 6%에 불과하다. 이러한 검색광고의 새로운 대안으로 문서 기반 광고가 대두하였다. 문서 기반 광고는 웹 검색 광고에 비해 월등히 많은 페이지뷰가 존재하며, 모든 웹에 광고를 할 수 있다는 장점이 있다. 이러한 이유로 다양한 방법이 연구되었으나, 대부분의 방법이 검색 기법을 응용한 단어 가중치를 통해 해당 문서에서 광고 키워드를 추출하는 방식이다. 이러한 방법은 동음이의어 문제나 핵심 단어 선정 실패로 인해 추출된 광고 키워드가 본문의 내용과 일치하지 않는다는 단점이 있다.

본 연구에서는 이러한 단점을 보완하기 위한 방법으로 문서 클러스터링을 이용한 문맥 기반 광고 방법을 제안하였다. 문서 클러스터링을 통해서 문서들의 특징을 미리 파악하고 그에 적절한 광고 키워드를 매핑하여 광고용 문서 분류셋을 생성하였다. 문서 분류셋과 광고를 매핑하고자 하는

6) http://clickchoice.naver.com/00

문서의 유사도를 측정하여 제일 유사한 분류를 찾고 이를 통해 광고 키워드를 추출하는 방법을 택하여 문서의 주제 파악의 용이 및 동음이의어로 인해 발생하는 오류를 줄일 수 있었다.

그러나 광고 키워드와 실제 표현의 차이에 따른 유의어 사전을 자동으로 작성할 수 있는 방법의 연구가 필요하다. 또한 문서에 사용된 표현이 별로 없는 짧은 내용인 경우 문서외의 주위 정보를 이용해서 광고를 찾는 기술 또한 연구가 앞으로 필요하다.

참 고 문 헌

[1] P. D. Turney. "Learning algorithms for Keyphrase Extraction," Information Retrieval, vol 2, no. 4, pages 303-336, 2000.

[2] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. "Domain-Specific Keyphrase Extraction," In Proceedings of IJCAI-99, pages 668-673, 1999.

[3] A. Hulth. "Improved Automatic Keyword Extraction given more Linguistic Knowledge," In Proceedings of EMNLP-03, pages 216-223, 2003.

[4] Wen Tau-Wih, Joshua Goodman, and Vitor R. Carvalho, "Finding Advertising Keywords on Web Pages," In Proceedings of the World Wide Web Conference 2006, Edinburgh, Scotland, 2006.

[5] Rafael A. Calvo, Jae-Moon Lee and Xiaobo Li, "Managing Content with Automatic Document Classification," Journal of Digital Information, vol. 5, 2004.

[6] 오장민, 장병탁, 김영택, "SVM 학습을 이용한 다중 클래스 뉴스그룹 문서 분류", 한국정보과학회 가을 학술발표, pages 60-62, 1999.

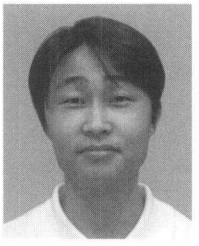
[7] 방선이, 양재동, 양형정, "k-NN 분류 알고리즘과 객체 기반 시소러스를 이용한 자동 문서 분류", 한국정보과학회논문지, vol. 31, no. 9, pages 1204-1217, 2004.

[8] 이경찬, "확률 기법을 이용한 자동 문서 분류 시스템", 석사학위논문 국민대학교, 2004.

[9] Christopher D. Manning, Hinrich Schutze, "Foundations of Statistical Natural Language Processing," MIT Press, Cambridge, MA, 1999.

[10] J. Goodman and V. R. Carvalho, "Implicit Queries for Email,", In Proceedings of the conference on Email and Anti-Spam (CEAS), 2005.

[11] Ricardo Baeza-Yates and Berthier Ribeiro-Neto "Modern Information Retrieval," Addison-Wesley, 2000.



이 동 광

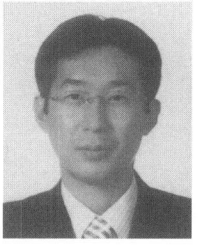
e-mail : dklee@nasmedia.co.kr

2001년 전북대학교 컴퓨터공학과(학사)

2003년 전북대학교 컴퓨터공학과 (공학석사)

2003년~현재 전북대학교 컴퓨터 공학과 박사과정

관심분야 : 정보 검색 및 마이닝, 자연언어처리 등



강 인 호

e-mail : ihkang97@cs.cmu.edu

1997년 경북대학교 컴퓨터공학과 (학사)

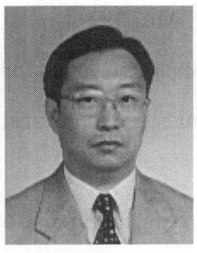
1999년 한국과학기술원 전산학과 (공학석사)

2004년 한국과학기술원 전산학과 (공학박사)

2004년~2005년 삼성종합기술원 전문연구원

2006년~현재 CMU/LTI 연구소 박사후과정

관심분야 : 정보 검색 및 마이닝, 자연언어처리 등



안 동 언

e-mail : duan@chonbuk.ac.kr

1981년 한양대학교 전자공학과 (학사)

1987년 한국과학기술원 전산학과 (공학석사)

1995년 한국과학기술원 전산학과 (공학박사)

2004년~2006년 전북대학교 정보검색시스템연구센터 센터장

1995년~현재 전북대학교 전자정보공학부 교수

관심분야 : 정보검색, 자연언어처리, 한국어정보처리