고속연관규칙을 이용한 문맥광고에서의 콘텐츠 추천*

김성민° 이성진 이수원
숭실대학교 컴퓨터학과
{mabak°, ptnrev93}@mining.ssu.ac.kr, swlee@ssu.ac.kr

Content Recommendation Using High-Speed Association Rule Generation for Contextual Advertisement

Sungming Kim°, Seongjin Lee, Soowon Lee
Dept of Computer Science, Soong-Sil University

요 약

인터넷 사용자가 급증함에 따라 온톨로지를 이용한 지능형 웹이나 인터넷 사용자에게 개인 맞춤형 서비스를 제공하기 위한 다양한 연구가 진행되고 있다. 대표적인 예로 문맥광고는 인터넷 사용자들이 뉴스나 커뮤니티 사이트에서 콘텐츠를 조회하고, 해당 콘텐츠와 일치하거나 관련성이 높은 제품 또는 서비스 정보를 제공하는 광고기법이다. 그러나 문맥광고는 사용자에게 다양한 콘텐츠 및 사이트 추천 서비스를 제공하지 못하고 있다. 따라서 다양한 콘텐츠 및 사이트 추천 서비스를 제공하기 위해 본 논문에서는 사용자가 조회한 콘텐츠의 내용을 대표할 수 있는 중요 키워드를 선정하고, 콘텐츠 내에서 추출된 키워드간의 연관성을 분석하여 관련 콘텐츠 및 사이트를 추천하는 방법에 대해 제안한다. 또한 연관키워드리스트 생성방법을 고속연관규칙을 이용하여 처리속도를 줄이고, 사용자가 선호할 만한 다양한 콘텐츠와 관련된 사이트를 제공하는 방법에 대해 제안한다.

1. 서 론

인터넷의 등장으로 온라인에서의 서비스영역이 확대되고 있고, 인터넷 뉴스의 사용자가 급증함에 따라 포털서비스 업체는 문맥광고(Contextual Advertising), 키워드검색 광고 등 다양한 수익모델을 창출하고 있다. 예를 들어, 키워드검색 광고와 문맥광고의 경우 2004년 기준 온라인광고시장의 약 40%를 차지하는 1777억(원)으로서 성장률이 85%에 이른다. 그러나 상품성 키워드에 대한 관리와 키워드검색 광고시장이 성장 한계점에 접근했다는 전문가들의 시각에 따라 문맥광고가 더욱 주목을 받고 있다.

<그림 1> 키워드검색광고와 문맥광고

<그림 1>에서 키워드검색 광고는 사용자가 원하는 정보를 검색할 경우 검색결과로 광고를 제공하고 있고, 문맥광고는 사용자가 콘텐츠를 볼 시점에 해당 콘텐츠에 맞는 광고를 제공한다[1][2][3].

현재 포털서비스 업체는 지식인, 커뮤니티사이트(까페, 블로그), 뉴스 등에서 관련 광고서비스를 제공하고 있다. 그러나 콘텐츠와의 관련성이 다소 떨어지거나 광고주가 선택한 키워드에 한해서만 관련 광고서비스를 제공되고 있으므로 해당 뉴스와 관련된 다양한 콘텐츠 및 사이트를 제공하는 것이 필요하다.

본 연구에서는 [6]에서의 연관키워드리스트 생성방법을 개선하여 DB에 많은 양의 사이트 메타데이터를 보유하지 않아도 뉴스 내에서 추출된 키워드만을 가지고 연관키워드리스트를 생성한다. 생성된 연관키워드리스트를 이용하여 키워드 간의 연관성을 검색하고 연관키워드 묶음을 생성하여 뉴스와 관련 있는 콘텐츠 추천방법을 제안한다.

본 연구의 2절에서는 뉴스와 관련 있는 사이트를 제공하기 위한 연구와 키워드의 연관성에 관한 연구에 대해서 소개한다. 3절에서는 뉴스의 내용과 관련 있는 서비스를 제공해주기 위한 제안방법에 대해 설명하고, 4절에서는 제안방법에 관한 실험 및 결과 대해서 기술하고, 5절에서는 결론 및 향후 연구과제에 대하여 언급한다.

2. 관련연구

본 절에서는 뉴스와 관련 있는 사이트를 제공하기 위한 Impedance Coupling전략에 대해서 소개한다. 또한, 콘텐츠에서의 문맥분석을 위한 키워드간의 연관성에 관한 연구에 대하여 소개한다.

2.1 Impedance Coupling Strategies

Berthier Ribeiro-Neto[10]는 웹페이지와 관련 있는 광고를 제공하기 위해 새로운 10가지 전략을 제안하였다. 웹페이지 p와 광고의 집합 A가 주어졌을 때, 웹페이지 p의 광고 a_i가 $a_i \in A$일 경우, 웹페이지 p에서 추출한 키워드들과 광고에서 추출한 키워드들을 단순매치(Simple Match)시키는 방법이다. 이 방법은 벡터공간모델(Vector Space Model)을 사용한다. 벡터공간모델에서의 질의어와 문서는 n차원 공간에서 가중치벡터(Weight Vector)로 표현된다. 질의 q에서의 단어 t_i와 연관된 가중치 w_{iq}, 문서 d_j에서의 단어 t_i와 연관된 가중치 w_{ij}가 있을 경우, $\vec{q}=(w_{1q}, w_{2q}, \ldots, w_{iq}, \ldots, w_{nq})$와 $\vec{d_j}=(w_{1j}, w_{2q}, \ldots, w_{ij}, \ldots, w_{nj})$는 질의 q와 문서 d_j의 가중치이다. 질의와 문서에서의

* 본 연구는 서울시 산학연 협력산업의 지원으로 이루어졌습니다.

가중치는 TFIDF를 이용하며, 유사도는 코사인유사도 [식1]을 사용한다.

$$sim(q,d_j) = \frac{\vec{q} \bullet \vec{d_j}}{|\vec{q}| \times |\vec{d_j}|} = \frac{\sum_{i=1}^{n} w_{iq} \bullet w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{iq}^2} \sqrt{\sum_{i=1}^{n} w_{ij}^2}} \quad [식1]$$

[10]에서는 웹페이지와 광고간의 매칭 되는 키워드가 없을 때, 웹페이지와 Topic이 같은 다른 웹페이지로부터 새로운 키워드를 확장 추가해주는 방법에 대해 다양한 전략에 대하여 제시하였다.

2.1 연관키워드그룹 추출기법, 신뢰도 및 집중도

연관키워드그룹추출기법[4]은 연관키워드를 추출하기 위한 기법으로서 연관규칙을 이용하여 타겟키워드와 연관 있는 후보키워드를 찾아내는 기법이다. 연관키워드그룹 추출기법은 검색엔진의 Web Log, 디렉토리 등록 사이트 정보, 광고 사이트 정보, 키워드 판매 정보를 이용하며 타겟 키워드별로 사이트 그룹을 형성한 후 각 사이트를 대표하는 중요 키워드를 선별하여 연관규칙과 협력적 추천에 의한 연관 키워드 후보 집합을 조합하는 방법이다.

[4]에서는 타겟키워드에 대한 사이트 그룹의 여러 문서에서 출현하면서 집중적으로 출현하는 키워드가 타겟 키워드와의 연관성(Relevance)이 높은 키워드라고 제시하였고, 신뢰도(Confidence)와 집중도(Concentration) 계산식은 다음과 같이 정의하였다.

$$Confindence(d \rightarrow r) = \frac{df_G(r)}{n_G} \quad [식2]$$

$$Concentration(d \rightarrow r) = \frac{df_G(r)}{df_{all}(r)} \times (1 - \frac{gf(r)}{N_G}) \quad [식3]$$

d : 타겟 키워드 — $df_G(r)$: G중에서 r이 출현한 사이트의 수

r : 연관 키워드 후보 — $df_{all}(r)$: 전체 사이트 중 r이 출현한 사이트의 수

G : d에 대한 사이트 그룹 — $gf(r)$: r이 출현한 사이트 그룹의 수

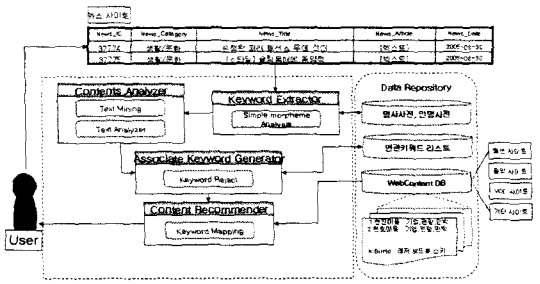
n_G : G에 속한 사이트의 수 — N_G : 전체 사이트 그룹의 수

2.2 ARCS알고리즘

ARCS알고리즘[5]은 FP-growth와 H-mine과 같이 패턴-성장 방법을 사용하여 빈발항목집합을 발견한다. 아이템집합 $X=\{x_1,x_2, \ldots, x_m\}$이 있을 경우 x_1을 포함하는 모든 빈발항목집합, x_1을 포함하지 않으면서 x_2을 포함하는 모든 빈발항목집합, $x_1 \ldots x_{m-1}$을 포함하지 않으면서 x_m을 포함하는 모든 빈발항목집합을 찾는 방식으로 탐사를 수행한다. [5]에서 트랜잭션 모델은 빈발항목의 이름으로만 구성되고, 공통된 트랜잭션 집합에 대해 같은 공간을 사용하기 때문에 매우 경량적이며, 패턴-성장 방법의 경우 조건적 헤더 테이블만을 반복적으로 생성, 탐사를 수행하므로 빠르게 빈발항목집합을 탐사할 수 있다[5].

3. 연관 콘텐츠 및 사이트 추천 시스템

본 논문에서 제안하는 시스템은 뉴스를 입력받아 분석에 필요한 형태소 분석과정과 수정된 TFIDF식을 이용하여 추출된 키워드의 가중치를 구하고, 추출된 키워드들의 연관성을 검색하는 과정을 통해 뉴스 내에서 연관키워드들을 찾아 추천키워드로 사용, 검색엔진에 넘겨주고 얻어진 결과를 뉴스하단에 붙여준다.



<그림 2> 시스템구조도

<그림 2>는 총 3개의 Step으로 진행된다. 첫 번째 과정은 키워드 추출과정으로 형태소분석과정을 거쳐 추출된 키워드는 수정된 TFIDF식을 이용하여 키워드 벡터에 저장한다.

$$MoTFIDF = (1 - \frac{Sdf}{SN}) * (1 - \frac{nSdf}{nSN+1}) * (\frac{tf}{tfmax} * (1 - \frac{df}{N}))$$

[식4] 수정된 TFIDF식

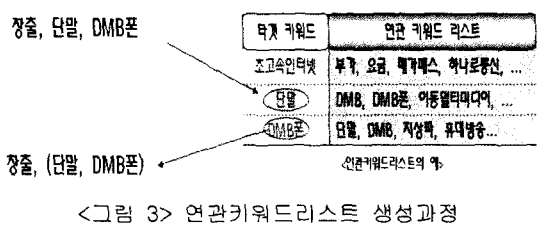
Sdf: 섹션내출현문서수 — SN: 섹션내전체문서수

$nSdf$: 뉴스섹션별출현수 — nSN: 뉴스전체섹션수

tf: 문서내출현빈도수 — $tfmax$: 문서내출현빈도수Max값

df: 전체출현문서수 — N: 전체문서수

[식4]는 일반 TFIDF식에 섹션의 개념을 확장 적용한 것으로 전체뉴스와 섹션에서 적게나오고 사용자가 본 뉴스 내에서는 많이 나타난 키워드를 중요키워드로 간주하였다. 두 번째 과정은 연관키워드 생성과정으로 뉴스로부터 추출된 문장별로 키워드 벡터 내에 저장하고 키워드 벡터내의 각각의 문장에서 추출된 키워드를 트랜잭션으로 간주, 연관규칙을 찾고, 찾아진 연관규칙들의 셋을 <그림 3>과 같이 연관키워드리스트를 생성한다. 연관규칙은 ARCS알고리즘을 이용하였다.



<그림 3> 연관키워드리스트 생성과정

<표 1> 트랜잭션 Set 구성 예

트랜잭션ID	아이템(키워드)
1	청바지, 폭발, 패션, 캠퍼스, 히피
2	예술가, 작업실, 여행자, 배낭
...	.....
n	캘빈클라인, 브랜드, 청바지, 스키니진..

<표 1>은 뉴스에서 추출한 문장들을 대상으로 트랜잭션 Set구성 예이다. 한 문장을 하나의 트랜잭션으로 간주, 각각의 문장에서 나온 명사들을 트랜잭션 아이템으로 구성하여 ARCS알고리즘을 이용하여 연관규칙을 생성한다.

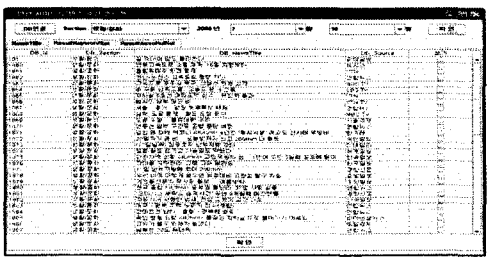
<표 2> 연관규칙 생성 예

A = > B [Support , Confidence, Lift]
죽음 => 주몽
[Support=0.115384616, Confidence=0.5, Lift=1.3]
주몽 => 죽음
[Support=0.115384616, Confidence=0.3, Lift=1.3]
아들 => 주몽
[Support=0.07692308, Confidence=0.5, Lift=1.3]
주몽 => 아들
.....
송일국 => 허준호
[Support=0.07692308, Confidence=1.0, Lift=8.666667]
허준호 => 송일국
[Support=0.07692308, Confidence=0.6666667, Lift=8.666667]
유화 => 주몽
[Support=0.115384616, Confidence=0.75, Lift=1.95]
주몽 => 유화
[Support=0.115384616, Confidence=0.3, Lift=1.95]

<표 2>에서 얻은 연관규칙 결과의 예는 타겟키워드와 단일후보키워드만을 이용한다. 세 번째 과정은 연관키워드 묶음에 있는 키워드를 검색엔진을 통해 검색하여 얻은 결과를 뉴스에 검색결과를 삽입하여 사용자에게 제공한다.

4. 실험 및 결과

본 논문에서 제안하는 시스템은 뉴스로부터 추출된 키워드 내의 연관된 키워드들을 묶어주고, 관련 사이트 및 콘텐츠를 추천해주는 것이다. 데이터는 인터넷 포털사이트에서 제공하는 명사 129,613단어, 인명 113,025명의 사전을 이용하여 뉴스에서 키워드들을 추출하였고, NewsDB에는 네이버뉴스를 2006년 7월 1일부터 31일까지 총 10235개의 뉴스를 사용하였다.



<그림 4> 연관콘텐츠 및 사이트 추천솔루션 UI

본 논문에서 제안한 연관키워드를 이용한 관련 사이트 및 콘텐츠 추천의 예는 다음과 같다.

<표 3> 뉴스 기사 예

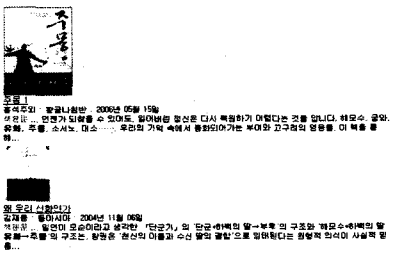
뉴스/날짜	뉴스 본문
TV 리포트 2006/7/4	MBC '주몽'은 3일 방송에서 해모수(허준호)가 짧은 시간 아들 주몽(송일국)과의 애틋한 시간을 보낸 후 부여군사들에 의해 참혹한 죽음을 당하는 장면을 방송했다....(생략)

<표 4> 중요키워드 추출 예

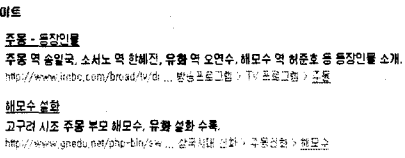
키워드	TF	TFIDF	MoTFIDF
주몽	10	0.9841	1.3952
죽음	6	0.5918	0.5771
눈물	5	0.4889	0.4647
가슴	4	0.3856	0.3620
아들	4	0.3900	0.3706
유화	4	0.3991	0.4963
시청자	3	0.2873	0.2436
사랑	3	0.2735	0.2088
허준호	3	0.2995	0.4466
전광렬	2	0.1997	0.3479

<표 5> 연관키워드 묶음 예

연관키워드리스트	연관키워드묶음 점수
유화, 주몽	0.9458
허준호, 송일국, 주몽	0.7121
주몽, 허준호, 유화, 죽음, 아들	0.5777



<그림 5> 관련 콘텐츠 추천결과(책)



<그림 6> 관련 사이트 추천결과(사이트)

<표 3>은 TV리포트에서 2006년 7월 4일 제공된 뉴스 기사이다. 내용은 드라마 '주몽'에서 해모수(허준호)가 죽었다는 내용이다. 이 뉴스와 관련해서 중요키워드로 <표 4>와 같이 '주몽'이 TF, TFIDF, MoTFIDF 모두 높게 나왔다. 연관규칙을 이용하여 '주몽'과 관련된 키워드로 '유화'가 <표 5>에서 묶였으며 검색키워드로 추천되었다. 연관키워드 묶음 점수는 각 키워드 MoTFIDF의 값의 평균으로 점수가 상위인 '유화', '주몽' 키워드를 가지

고 검색엔진을 통해 추천된 결과를 <그림 5>, <그림 6>에서 보여주고 있다.

<표 6> 뉴스 기사 예2

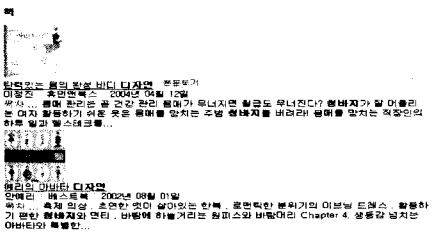
뉴스/날짜	뉴스 본문
경향신문 2006/7/26	청바지의 매력은 흡수력이다. 록바의 폭발하는 젊음 속에서도 고독한 예술가의 작업실에서도 그멋을 드러낸다. .....중략..... 할리우드 스타 모두를 마니아로 만들어버린 청바지...생략...

<표 7> 중요키워드 추출 예2

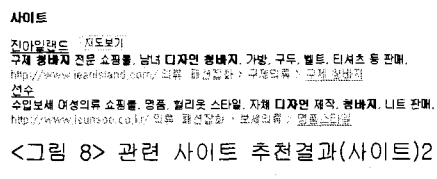
키워드	TF	TFIDF	MoTFIDF
청바지	17	0.9968	0.9903
스키니진	5	0.2939	0.4396
디자인	5	0.2900	0.3486
유행	5	0.2903	0.2807
몸매	4	0.2311	0.2273
브랜드	3	0.1725	0.1674
발목	3	0.1744	0.1722
운동화	3	0.1761	0.1751
허벅지	3	0.1756	0.1736
정장	3	0.1760	0.1745

<표 8> 연관키워드 묶음 예2

연관키워드리스트	연관키워드묶음 점수
디자인, 청바지	0.6694
유행, 청바지	0.6355
디자이너, 청바지	0.6029



<그림 7> 관련 콘텐츠 추천결과(책)2



<그림 8> 관련 사이트 추천결과(사이트)2

<표 6>은 경향신문에서 2006년 7월 26일 제공된 뉴스 기사로 내용은 청바지로 멋을 낸다는 내용이다. <표 7>에서는 중요키워드로 '청바지'가 TF, TFIDF, MoTFIDF 모두 높게 나왔으며 연관규칙을 이용하여 '청바지'와 관련된 키워드로 '디자인', '유행', '디자이너'가 나타났다. 그 이외에도 관련키워드로 '패션', '정장' 등 청바지와 어울리는 키워드가 관련키워드로 뽑혔다. 검색결과는 <그림 7>, <그림 8>에서처럼 뉴스기사의 내용과 관련이 있는 콘텐츠와 사이트를 제공하고 있는 것을 볼 수 있다.

5. 결론 및 향후 연구

본 논문에서는 뉴스로부터 추출된 키워드에 대해서 수정된 TFIDF식을 이용하여 키워드 가중치를 구하고, 키워드벡터 내에 존재하는 키워드들을 ARCS알고리즘을 이용하여 연관키워드리스트를 생성하고, 키워드간의 연관성을 검색하였다. 또한 뉴스와 관련된 연관키워드 묶음을 생성하여 뉴스 관련 콘텐츠 및 사이트를 추천해주는 기법을 제안하였다.

본 연구에서는 연관키워드리스트 생성시 뉴스의 내용이 부족할 경우 연관키워드리스트가 생성되지 않는 것을 확인 할 수 있었다. 또한 뉴스 또는 지식서비스, 블로그 등 타이틀에서 추출되는 키워드의 가중치에 관한 연구와 사용자 모델링을 통해 사용자의 선호키워드를 이용하여 개인화된 서비스를 제공하고, 사용자 취향에 맞게 선호하는 콘텐츠를 다이나믹하게 제공하는 연구가 진행되어야 한다.

참고문헌

[1] 한국광고데이터(KADD) www.kadd.co.kr

[2] (주) 서치안 www.searchian.com

[3]2005검색광고마케팅컨퍼런스(주)나무커뮤니케이션 www.namukorea.com

[4] 이성진, 이수원, "키워드 마케팅을 위한 연관 키워드 추출 기법", 한국정보과학회 2004 추계학술대회, VOL.31 NO.02 p.0124~0126 2004.10

[5] 한영우, 이수원, "고속의 연관 규칙 마이닝을 위한 효율적 공간압축 및 탐사 기법, 데이터마이닝학회 2002년 데이터마이닝 추계학술대회 , 2002

[6] 김성민, 이수원, "문맥광고에서 관련 사이트 추천을 위한 연관키워드 마이닝기법", 한국정보처리학회 2006년 춘계학술대회, VOL. 13NO. 01pp. 0337 ~ 0340, 2006.05

[7] Jiawei Han, Micheline Kamber, "Data Mining Concepts And Techniques", Morgan Kaufmann, 2000.

[8] Tomohiko Sugimachi, Akira Ishino, Masayuki Takeda, and Fumihro Matuo, "A Method of Extracting Related Words Using Standardized Mutual Information", Springerlink, Computer Science, p.478-485, 2003.

[9] Chien-Chung Huang, Shui-Lung Chuang, Lee-Feng Chien, "Mining theWeb for Generating Thematic Metadata from Textual Data" The 20th IEEE International Conference on Data Engineering (ICDE), 2004.

[10] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Gogher, Edleno Silva de Moura, "Impedance Coupling in Content-Targeted Advertising", In *Proc. of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval*, Brazil, p.496-5 , 2005.