뉴스기사의 연관 단어 텍스트 마이닝을 이용한 스타의 분야별 기여도순위 비교기법

강윤정*, 윤재열**, 임지연***, 김응모*
*성균관대학교 정보통신공학부
e-mail : miniringo@skku.edu

Ranking Contribution of Star in Each Domain Using Association Text Mining News Articles on the Web

Yoonjeong Kang*, Jaeyeol Yoon**, JiYeon Lim***, Ung-mo Kim*
* School of Information and Communication Engineering

요 약

스타의 대중에 대한 인기가 브랜드의 이미지 제고와 상업적 영향을 끄는 마케팅 전략을 스타 마케팅이라고 한다. 오늘날의 스타는 방송, 연예활동뿐만 아니라 스포츠, 정치활동, 사회기여활동 등 다양한 분야에서 활약하며 스타의 이미지는 그 활약상에 영향을 받는다. 스타의 이미지는 브랜드 및 기업의 이미지로 직결되므로 그에 대한 사전분석은 마케팅에서 중요한 요소이다. 그래서 일반적으로 스타들이 활약하는 도메인을 분류하여서 그 스타에 대해서 검색을 하였을 때 어떤 분야에서 활약하고 기여를 하는지 그 기여도를 도메인에 따라 랭킹을 매기는 방법을 제안한다.

뉴스기사에서 텍스트 마이닝 기술을 이용하여 스타의 이름과 활동 도메인들에 대해서 관련단어를 빈도에 따라 추출한다. 그리고 관련된 단어들을 이용하여 스타에 대한 뉴스 중 각 도메인과 관련된 기사들을 카운트하며 도메인에 대해서 긍정 혹은 부정적인 보도내용일 경우에는 극성을 부여하여 그 가중치를 달리한다. 빈도 및 극성을 고려한 점수화에 의해 스타가 기여하는 분야에 대한 순위를 매긴다.

1. 서론

최근 기업이나 광고주체는 직접적인 커뮤니케이션을 통하여 소비자를 설득시키기보다는, 일정한 인지도와 지명도를 가지는 특정인물을 통해서 친근한 간접커뮤니케이션을 통해서 광고목적을 달성하려고 한다. 그 대표적인 예가 스타마케팅이다.

스타마케팅은 스타의 대중적 인기를 상품, 서비스, 이벤트, 사회봉사활동 등에 연계한 마케팅 전략이다. 즉 스타가 팬들에 대해 행사하는 상업적 잠재력을 활용하려는 전략으로, 스타마케팅의 영역은 상품 판매나 서비스는 물론 불우이웃돕기 등 사회봉사활동, 정치인의 선거, 기업의 이미지 전략 등 모든 분야로 확대되고 있으며[1], 이러한 스타마케팅은 스타모방심리를 이용한 간접광고 효과가 뛰어나 가장 선호되고 있는 마케팅 수단으로 이미 정착되어 수년 전부터 각광 받는 프로모션방법으로 인식되고 있다. [2] 그러므로 잘 알려진 공인, 특히 연예인 및 스타에 대한 평판과 활약상은 광고주나 기업에게 핵심 마케팅 자료로써 인식되었다.

본 논문에서는 텍스트마이닝 기술을 이용하여 스타들이 활약하는 도메인을 사전에 분류해놓고 스타의 이름을 검색하였을 때 기존에 분류해놓은 도메인에 랭킹을 부여하는 기법을 제안한다. 실제 인터뷰나 설문에 따른 정보가 아닌 기사화된 연예정보를 가지고 분석되는 데이터이기 때문에 사실들에 기반한 스타들의 특징과 사회기여도를 알 수 있다.

2. 관련연구

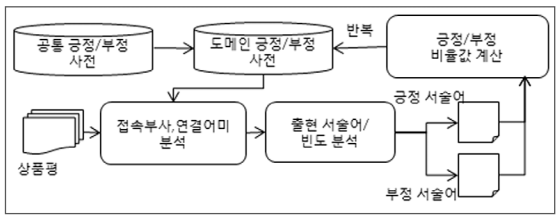
2.1 연관단어 마이닝[3]

문서의 분류를 위해 연관 단어 마이닝이라는 효율적인 특징 추출 방법이 있다. 문서의 특징을 단일 단어가 아닌 해당문서 내에서 연관성이 존재하는 몇 개의 단어로 구성한 연관 단어 벡터로서 표현된다. 연관단어를 구성하는 단어의 수, 신뢰도, 지지도는 문서 분류의 정확도와 재현율에 영향을 미친다. 연관 단어 마이닝을 이용한 특징 추출 방법은 프로파일을 사용하지 않으므로 프로파일 갱신의 필요성이 없으며, 색인어에 대한 확률을 계산하지 않고도 신뢰도와 지지도에 따라 자동으로 명사구를 생성하므로, 단어의 연관성을 이용하여 문서의 특징을 추출하는 기존 방법에 대한 문제점을 해결한다. 연관 단어를 사용하는 기존 문서 분류와 전반적으로 비교하면 연관 단어 마이닝을 사용한 특징 추출의 방법이 성능면에서 높음을 나타낸다. 반면, 연관단어 마이닝은 특징 추출을 하기 전에 사용자 프로파일이나 색인어 사전을 생성해야 하는 단점은 없으나 데이타베이스를 여러 번 검색해야 하는 단점이 있다.

2.2 도메인 긍정 부정어 사전[4]

일반적인 의미를 가진 서술어의 긍정/부정 분류는 평점 긍정/부정 사전만으로도 서술어의 긍정/부정의

분류가 가능하지만 다른 의미방향을 가진 서술어들이 평점에 의해 하나의 의미방향으로 분류되는 문제점을 가지고 있다. 예를 들면, “저렴한 가격이 좋지만 싼 소재를 사용한 것 같다”와 같은 문장은 긍정적인 의미방향인 ‘저렴하다’와 부정적인 의미방향인 ‘싸다’가 모두 부정적인 평점으로 인해 부정적인 의미방향으로 분류되는 문제가 존재한다. 이러한 이유로 한국어 문장에 존재하는 접속부사와 연결어미 정보를 사용하여 한 문장 안에서도 여러 의미방향으로 분류가 가능하도록 한다. 공통 긍정/부정 사전에 구축된 공통 서술어는 도메인별로 도메인 긍정/부정 사전에 초기화 되어 Seed Word 로 활용된다. 초기화된 Seed Word 와 서술어 사이의 접속부사 및 연결어미 정보를 이용하여 도메인별로 새로운 서술어의 의미방향을 분류하게 된다. 새롭게 분류된 서술어들은 도메인 긍정/부정 사전에 추가되고, 추가된 서술어들은 다시 새로운 Seed Word 로 활용된다.



(그림 1) Seed Word 와 접속 정보를 이용한 도메인 긍정/부정 사전 구축 과정

3. 분야별 기여도의 순위 결정시스템[4-5]

본 장에서는 스타로서 활약하는 대표적인 도메인을 10 가지로 사전 분류하였다.

1. 정치	6. 모델
2. 사업	7. 영화
3. 교육 및 학업	8. 스포츠
4. 사회공헌	9. 음악
5. TV 방송	10. 공연 및 뮤지컬

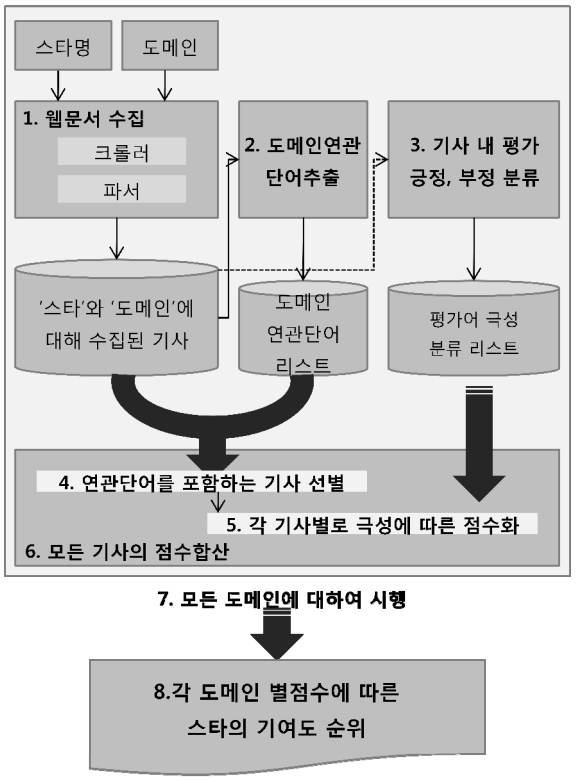
<표 1> 스타들이 활약하는 대표 도메인

먼저 뉴스기사가 스타와 도메인에 대한 뉴스인지를 알기 위해서 도메인과 관련된 단어를 수집해야 한다. 도메인과 관련된 기사를 마이닝하기 위해서 도메인 관련어에 대한 TF(term frequency)뿐만 아니라 기사내용의 감정 극성(sentiment polarity)과 기사에서 평가에 대한 감정 단어 (sentiment word)에 대한 TF 를 바탕으로 카운트하여 도메인 기여도 점수를 부여한다.

3.1 도메인과 관련된 단어의 추출

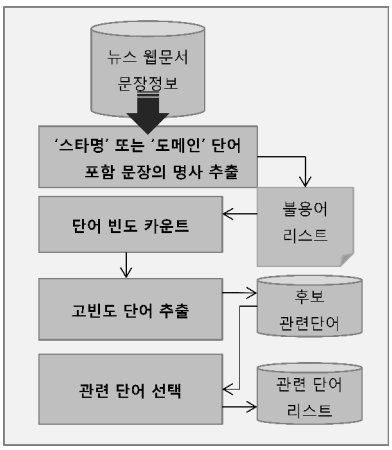
웹크롤러를 통해서 뉴스기사를 수집한 후에 10 가지 도메인과 관련된 연관 단어를 마이닝 하기 위해, ‘스타명’이나 ‘도메인’이 포함된 문장들을 추출한다. 그 문장들에는 연관있는 단어들이 포함될 확률이 높기 때문에 해당 문장들에서 명사들을 추출한다.

그림 2 는 제품 특징을 추출하는 과정이다. 후보 제품 특징을 카운트하는 과정에서 불용어 리스트를 활용하여 필요하지 않는 단어를 필터링하는 단계를 거친다.



(그림 2) 분야별 기여도의 순위 결정시스템

3.2 연관단어 리스트를 포함하는 기사 선별



(그림 3) 도메인 관련단어 리스트 생성 과정

수집된 기사들이 ‘스타명’과 ‘도메인’ 키워드를 포함하고 있지만 쓸모없는 기사를 제거하기 위해서 필터링이 필요하다. 수집된 기사들을 연관단어 리스트와 비교하여 기사 내에 연관단어를 포함하고 있는지 확인한다. 그러기 위해서 기사들의 문서에 나타난 문장에 일련의 번호를 부여하고, 문장

번호와 함께 문장으로부터 추출한 명사를 데이터베이스에 저장한다.

> **김태희**가 지난 2 일 크랭크 인 한 **영화** '그랑프리'의 스틸컷을 통해 싱그러운 단발머리를 공개했다.
> 긴 머리를 층을 내 자른 **김태희**는 어깨에 닿을 듯 말 듯한 길이의 머리카락을 바람에 흩날리며 담담한 표정을 짓고 있다. 사고로 말을 잃은 주희(**김태희**)가 아픔을 추스리기 위해 제주도로 향했고 그곳에서 그를 응원해주는 기수 우석을 만나는 장면이다. 잿빛 청바지에 갈색 셔츠, 흰색 쁘띠 스카프를 목에 두른 **김태희**는 "경마는 내게 새로운 도전인데 경마가 주는 쾌감을 관객들에게 잘 전달하고 싶다"고 소감을 밝혔다. **김태희**는 기수 역할을 위해 일주일에 4 일간 승마 연습을 하고 있다. '아이리스' 양윤호 감독과 **김태희**의 랑데부작인 '그랑프리'는 하반기 개봉된다.

(문서 1) 스타명 '김태희'와 도메인 '영화' 에 의해서 수집된 뉴스

문서 1 에서는 다음과 같이 명사를 추출해 낼 수 있다.

문장번호	추출된 명사
1	크랭크, 영화, 그랑프리, 스틸컷, 단발머리
2	머리, 어깨, 표정
3	사고, 말, 장면
4	잿빛, 셔츠, 스카프, 도전, 관객, 소감
5	역할, 연습
6	그랑프리, 감독, 랑데부작, 개봉

<표 2> 도메인 '영화'와 관련되어 웹문서에서 추출된 연관단어들의 데이터베이스

위와 같이 기사에서 추출된 단어들과 관련단어 리스트를 비교해서 교집합 단어들을 TF-ISF 를 적용하여 빈도를 구한다. TF-ISF 는 마이닝에서 이용하는 정보로, 어떤 단어가 특정 문장 내에서 얼마나 중요한 것인지를 나타내는 수치이다. 문장의 핵심어를 추출하거나, 검색 결과의 순위를 결정하거나, 문장들 사이의 비슷한 정도를 구하는 등의 용도로 사용된다. 우선 사용자가 질의한 단어의 TF 의 값이 큰 순서로 관련 선정한다. TF 는 문서 내에서 특정 단어가 출현하는 회수를 모든 단어가 출현하는 회수로 나눈 값으로 다음의 식(1)로 표현된다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

이 식에서 $n_{i,j}$ 는 단어 t_i 의 출현 회수를, $\sum_k n_{k,j}$ 는 모든 단어의 출현 회수를 의미한다. 하지만 TF 가 동률일 상황이 생길 수 있으므로 순위 선정 기준 중 다음과 같은 신뢰도도 추가한다. 또 신뢰도를 결정하기 위한 식 (2)는 다음과 같이 구해진다.

$$\text{Confidence}(W1 \rightarrow W2) = \Pr(W2|W1) \quad (2)$$

식 (2)는 단어 W1 과 W2 의 모든 단어 항목을 포함하고 있는 문장의 수를 단어 W1 의 단어 항목을 포함하고 있는 문장의 수로 나눈 결과 값을 나타낸다. 연관단어가 기사 내에서 높은 출현빈도와 신뢰도를 보일 때 스타 및 도메인과 관련된 유효한 기사임을 알 수가 있다.

3.3 평가에 대한 극성분류와 점수화

기사라는 것은 객관적인 정보 및 소식에 대해서 스타에 대한 평가가 포함될 수가 있다. 예를 들어 '영화' 의 연관단어 중에는 흥행, 관객, 연기가 있을 수 있다. '흥행하고 있다', '연기가 훌륭하다' 와 같은 표현은 긍정적인 평가이며 '흥행이 저조하다', '연기력이 논란이 되다' 와 같은 표현은 스타의 활동에 대한 부정적인 평가이다. 스타의 도메인에서의 활동에 대해서 <좋다/나쁘다>라는 기본적인 의견부터 연기가 <매력적이다 / 어둡다 / 다양하다> 등의 다양한 의견까지 표현할 수 있다.

본 장은 평가어 긍정/부정 리스트 구축 과정에 대한 내용이다. 문장의 접속부사 및 연결어미 정보를 이용하여 도메인별로 서술어의 의미방향을 분류하게 된다. 분류된 서술어들은 평가어 긍정/부정 리스트에 추가되고, 추가된 서술어들은 새로운 Seed Word 로 활용되며 더 많은 기사를 반복 분석하는 과정을 수행하여 평가어 긍정/부정 리스트가 확장/구축된다.

기사 내용의 긍정/부정 평가어들에 의해서 기사자체의 극성을 결정해야 한다. 단어 극성에 대한 엔트로피는 기사의 평가가 긍정 과 부정 중 어느 한 방향으로 기울었는지 가늠할 수 있는 선정 기준이다. 기사 d 의 엔트로피는 다음과 같이 계산할 수 있다.

$$H_d = \left(-\frac{p}{s}\log\frac{p}{s}\right) + \left(-\frac{n}{s}\log\frac{n}{s}\right) \quad (3)$$

식 (3)에서 기사 내에서 p 는 긍정적인 평가 단어의 출현 회수를, n 은 부정적인 감정 단어의 출현 회수를, s 는 감정 단어 전체의 출현 회수를 의미한다. 엔트로피의 값은 0 에서 1 사이의 값이 되며, 0 에 가까울수록 상품평의 극성이 뚜렷하다는 것을 의미하므로, 사용자의 긍정이나 부정적인 의견이 일관성 있게 표현된 기사임을 알 수 있다.

기사를 추출한 키워드와 평가 단어가 근접해 있을 경우, 도메인의 연관단어들에 대해 스타를 어떻게 평가했는지 더 정확히 알 수 있다. 기사 내에서 연관단어와 평가 단어가 멀리 떨어져 있다면 둘 사이의 직접적인 상관관계는 적을 것이다.

마지막으로 기사 전체의 감정 단어에 대한 TF 로는 상품에 대한 사용자의 감정의 폭을 알아볼 수 있다. 평가 적은 중립적인 기사는 스타에 대한 평가보다는 일반적인 설명이 담겨 있는 경우이기 때문에 가중치를 제외하고 점수를 부여한다.

각 유효한 기사들에 대해서 극성을 적용시켜 합산한다. 도메인과 관련된 기사의 숫자뿐만 아니라 긍정적인 기사가 많을 경우 점수가 높게 나오며, 부정적인 기사가 많을 경우에는 점수가 떨어진다. 그리고 사실을 전달하는 기사에 대해서는 긍정적인 기사와 똑같이 긍정적 기여를 한다고 간주한다.

3.4 최종 결과물

제안한 시스템을 적용하였을 때 수집한 기사들에 대해서 스타의 분야별 기여도를 퍼센트로 나타낸다. 스타가 활동하는 영역 전체를 100%로 하고 점수화되었던 기여도에 따라서 다음과 같이 순위가 매겨지게 된다.

스타명	김태희		스타명	문근영	
기여 순위	도메인	기여도 (%)	기여 순위	도메인	기여도 (%)
1	모델	62	1	영화	40
2	TV 방송	21	2	TV 방송	24
3	영화	12	3	사회공헌	20
4	사회공헌	4	4	모델	13
5	교육 및 학업	1	5	교육 및 학업	3

<표 3> 기사 분석의 예시 결과

4. 결론 및 향후 연구

이제까지 스타들의 평판을 수집하는 것은 특정 층의 대중을 상대로 한 인터뷰와 설문을 통해서 수집되는 것이 일반적이었다. 하지만 이는 시간과 비용이 많이 들고 집단의 특성에 따라서 그 의견이 주관적, 편파적일 수 있다는 단점이 있었다. 또한 스타들의 실제 활동들은 시간단위로 업데이트되는 연예소식들을 통해 알아내기 때문에 조사하는 것도 쉽지 않았다. 본 논문에서는 뉴스기사를 통해 전해지는 객관적인 소식과 사실들을 통해서 스타들의 각 분야에 대한 기여도를 알 수 있게 하는 시스템을 제안하였다. 본 제안은 인터넷 기사를 바탕으로 스타들이 어떠한 활동을 통해 사회에 기여하고 영향력을 주는지에 대해 알 수 있는 지표가 된다. 이를 통해 기업 및 광고주들은 업데이트가 빠르고 객관적인 지표를 활용하여 자신들에게 맞는 스타를 모델로 기용하고 브랜드 이미지 제고 및 효과적인 마케팅을 할 수 있을 것이라 기대한다.

참고문헌

[1] 허행량 "스타 마케팅" 서울:매일경제신문사 (2002)
[2] 신정혜. "TV 의상이 신세대의 의복구매행동에 미치는 영향: 스타마케팅을 중심으로" 숙명여자대학교 대학원 석사학위논문. (2000)
[3] 고수정, 최준혁, 이정현. "연관 단어 마이닝을 사용한 웹문서의 특징 추출", 정보과학회논문지, 데이타베이스, 제 30 권 제 4 호 (2003.8) p352,353
[4] 송종석, 이수원. "상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축", 정보과학회논문지, 소프트웨어 및 응용, 제 38 권 제 3 호 (2011.3) p158~p165
[5] 윤홍준, 김한준, 장재영. "오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법", 정보과학회논문지, 컴퓨팅의 실제 및 레터, 제 16 권 제 2 호 (2010.2) p224