SNS 비정형데이터 크롤링을 통한 드라마 시청률의 연관어 분석

강선경* · 이현창**· 신성윤***

*원광대학교 컴퓨터소프트웨어공학과

**원광대학교 디지털콘텐츠공학과

***군산대학교

Analysis of related words of drama viewership through SNS unstructured data crawling

Sun-Kyoung Kang* · Hyun-Chang Lee** · Seong-Yoon Shin***

*Department of Computer(Software) Engineering, Wonkwang University, Iksan, 54538, South Korea

**Department of Digital Contents Engineering, Wonkwang University, Iksan, 54538, South Korea

***Scool of Computer Information & Communication Engoneering, Kunsan National University, Kunsan 54150, South Korea

E-mail :{doctor10, hclglory}@wku.ac.kr, s3397220@kunsan.ac.kr

요 약

본 논문에서는 드라마의 시청률에 영향을 미치는 요소가 무엇인지를 파악하기 위해 정형화된 데이터와 비정형화된 데이터를 분석하기 위한 내용이다. 정형화된 데이터 수집은 각 방송사의 드라마 정보, 인물정보, 방송정보, 시청률정보라는 4가지 영역에서 총 19가지항목을 수집하였다. 비정형데이터를 수집하기 위해 각 방송사에서 드라마별로 운영되고 있는 게시판과 방영전블로그와 방영후블로그로부터 크롤링기법을 이용하여 수집하였다. 수집된 데이터로부터 방송사별 드라마 방영시간대, 방영시작시기, 장르, 방영요일에 따른 차이를 비교한 결과 방송사별 서로 유사한 것으로 나타났다.

ABSTRACT

In this paper, we analyze contents of formal and non - standardized data to understand what factors affect the ratings of drama. The formalized data collection collected 19 items from the four areas of drama information, person information, broadcasting information, and audience rating information of each broadcasting company. In order to collect unstructured data, crawling techniques were used to collect bulletin boards, pre - broadcast blogs and post - broadcast blogs for each drama. From the collected data, it was found that the differences according to broadcasting time, the start time, genre, and day of broadcasting were similar among broadcasting companies.

키워드

데이터 분석, 드라마, 시청률, 연관어, 정형, 비정형

Ⅰ. 서 론

최근들어 빅데이터를 활용하여 기업이나 정부 그리고 민간들간의 어떠한 방향제시나 융합지식을 도출할 수 있는 방법을 찾기 위한 예측기술들이 활발히 이루어지고 있다. 또한 바쁜일상에서 사람들이 선호하는 트랜드나 좋아하는 연예인, 노래, 영화, 드라마, 등을 미리 분석하여 예측하려고 하는 노력도 다양한 분야에서 이루어지고 있다. 그 중 인기 있는 드라마에서 사용되는 언어나 패션 스타일, 가전제품, 출연배우 등이 시청자의 정

서나 생활, 그리고 드라마 상영 중 간접 광고로 인한 수익창출 및 시청자들의 소비패턴 등에도 많은 영향력을 끼치는 핵심요소로 자리 잡고 있다. 그로인해 이 드라마의 성공 요인이 사회적, 문화적, 경제적 변화 요인으로까지 자리 잡을 수 있다. 본 논문에서는 드라마의 성공을 판단할 수 있는 시청률에 영향을 미치는 핵심 요소인 연관어를 분석하고자 한다.[1][2]

II. 데이터의 수집

현재 실시되는 시청률 조사는 일부 선정된 가구의 TV에 특정 셋톱박스를 설치한 뒤 시청자들의 TV 시청행동을 데이터화시켜 분석하는 VBM(Viewer Behavior Measurement) 방식으로 진행되고 있다. 이러한 방법은 급증하는 스마트폰과 인터넷, DBM, 태블릿PC 등을 통한 TV 시청인구를 시청률 산출에 반영하지 못하고 있는 상태이다. 따라서 현재 실시되고 있는 시청률 조사 결과의 신뢰도를 높이는 시청률 지표 산출 방법이 필요하다. 이러한 시청률 지표 산출을 위해서는 기존에 시청률이 높았던 드라마에서 그에 영향을 주는 요소가 무엇인지를 먼저 분석해 보아야 된다. 먼저 분석을 위해 그림 1과 같이 드라마 정보, 인물정보, 방송정보, 시청률 정보와 같은 정형데이터와 각 드라마 게시판과 방영전 블로그와 방영후 블로그로부터 비정형데이터를 수집한다.

구분	자료 내용	자료 출처	수집방법
드라마 정보	드라마명	http://movie.daum.net/TV	직접 입력
	장르	http://movie.daum.net/TV	직접 입력
	관람등급	http://movie.daum.net/TV	직접 입력
	총회수	http://movie.daum.net/TV	직접 입력
	연출가	http://movie.daum.net/TV	직접 입력
	극본	http://movie.daum.net/TV	직접 입력
인물 정보	주인공_남	http://movie.daum.net/TV	직접 입력
	주인공_여	http://movie.daum.net/TV	직접 입력
	조연1	http://movie.daum.net/TV	직접 입력
	조연2	http://movie.daum.net/TV	직접 입력
방송정보	방송사	http://movie.daum.net/TV	직접 입력
	방영기간	http://movie.daum.net/TV	직접 입력
	방영개월수	http://movie.daum.net/TV	직접 입력
	방영요일	http://movie.daum.net/TV	직접 입력
시청률 정보	첫회시청률	http://www.agbnielsen.co.kr	직접 입력
	시청률	http://www.agbnielsen.co.kr	직접 입력
	연관어	각 방송사 드라마 게시판	데이터 크롤링
	방영전블로그	NAVER blog 수	데이터 크롤링
	방영후블로그	NAVER blog 수	데이터 크롤링

그림 1. 분석을 위한 정보와 정형데이터 수집

III. 데이터 분석

데이터 분석을 위한 드라마정보, 인물정보, 방송정보, 시청률 정보에 대한 데이터를 분석한 결과 방송사 간의 특성이 거의 비슷하다는 것을 알 수 있다. 그림 2에서 보는 것과 같이 드라마 방영시간대, 방영시작시기, 장르, 방영요일에 따른 차이를 카이제곱을 이용하여 비교 분석한 결과 방송사별 유사분포 값이 큰 차이를 보이지 않는 것을 알 수 있다.

구 분		전체	방송사			카이제곱
			KBS	MBC	SBS	
전체		169	57	53	59	
		100.0%	33.7%	31.4%	34.9%	
방영 시간대	아침(오전)	9.5%	10.5%	9.4%	8.5%	0.843
	저녁(18~22시)	21.9%	22.8%	24.5%	18.6%	
	심야(22시 이후)	68.6%	66.7%	66.0%	72.9%	
방영 시작시기	봄	25.4%	19.3%	26.4%	30.5%	4.282
	여름	23.7%	24.6%	24.5%	22.0%	
	가을	29.6%	28.1%	34.0%	27.1%	
	겨울	21.3%	28.1%	15.1%	20.3%	
장르	멜로	39.1%	40.4%	35.8%	40.7%	4.856
	가정	24.3%	26.3%	26.4%	20.3%	
	사극	11.8%	14.0%	15.1%	6.8%	
	기타	24.9%	19.3%	22.6%	32.2%	
방영 요일	일일	19.5%	22.8%	20.8%	15.3%	7.179
	월화	24.9%	31.6%	17.0%	25.4%	
	수목	29.0%	29.8%	30.2%	27.1%	
	주말	26.6%	15.8%	32.1%	32.2%	

그림 2. 방송사 간 특성비교

드라마의 시청률에 영향을 미치는 연관어를 분석하기 위해 본 논문에서는 각 방송사별 드라마의 게시판 글과 방영전 드라마 블로그의 글 그리고 방영후 드라마 블로그의 글을 분석하여 진행한다. 그림3은 드라마의 어휘의 출현빈도를 분석하고 모든 드라마에서 출현되는 어휘만을 추출하여 출현빈도와 시청률과의 상관관계를 분석해 총 7개의 연관어를 도출한다.

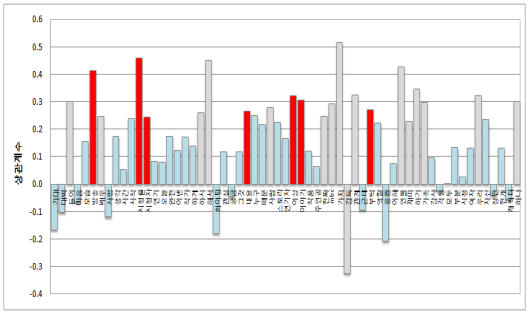

그림 3. 시청률과 출현빈도의 상관관계 분석

IV. 결 론

본 논문에서는 데이터 분석을 통해 드라마시청률과의 관련 요인으로 첫회 시청률, 방영개월수, 방영 시간대, 방영 요일인 것을 알게 되었고, SNS 비정형데이터 크롤링을 통해 연관어 분석을 시행한 결과 7개의 연관어를 찾을 수 있었다. 이는 드라마의 성공여부를 예측할 수 있는 주요 요소로 추후 광고주의 광고선택에 활용될 수 있을 것으로 보인다.

참고문헌

[1] 김유신, 김남규, 정승렬, “빅데이터 감성분석을 통한 지능형 투자의사결정모형,” 한국지능정보시스템학회, 2012

[2] 이미영, 최완, “빅데이터 처리 및 저장관리 기술동향 및 전망,” 한국정보통신학회, 2012