방송콘텐츠 영향력 도출을 위한 빅데이터 분석체계에 관한 연구[1)]

*최홍규[2)] **박구만 **최성진 *김성태[3)]
* 고려대학교 ** 서울과학기술대학교

*copernican1978@gmail.com **gmpark@seoultech.ac.kr **ssjchoi@seoultech.ac.kr
*sutkim@korea.ac.kr

Research on the big data collecting system for measuring of broadcast content influence

*Hong-Gyu Choi **Goo-Man Park **Seong-Jhin Choi *Sung-Tae Kim
* Korea University ** Seoul National University of Science and Technology

요약

본 논문은 방송콘텐츠 영향력 도출을 위해 고려되어야 할 요소들에 대해 다뤄보았다. 기존에 방송콘텐츠의 영향력을 나타내는 측정지표로 시청률과 청취율 같은 설문조사 방식의 조사자의 개입을 통한 방식이 활용되었다면, 최근 소셜미디어를 통해 수많은 정보가 교환되는 환경에서는 새로운 측정방식의 제안이 가능할 것으로 보았다. 이에 , 본 연구에서는 소셜미디어상 대용량의 텍스트 데이터인 이른바 '소셜텍스트 빅데이터'를 활용해 방송콘텐츠의 영향력을 분석하는 방식을 제안하였다. 또한 이러한 빅데이터 분석을 위해 일반적으로 발생할 수 있는 문제들과 이 과정에서 유의하여야 사항들에 대해 다뤄보았다.

1. 서론

방송콘텐츠에 대한 영향력을 살펴보는데 있어 현실적인 지표가 필요하다는 논의는 지속적으로 이루어지고 있다. 전파를 통해 TV 및 라디오 프로그램이 송수신되는 환경에서는 시청률과 청취율 등의 지표를 통해 방송콘텐츠가 지니는 파급력을 측정해왔으나, 방송관련 기술이 발달하게 되면서 콘텐츠 영향력을 살펴보기 위해 고려해야 할 요소들이 많아진 것이 사실이다.

오늘날 소셜미디어 공간에서는 인구통계학적 요소, 정치적 성향, 사회적 관심사 이르기까지 다양한 정보들이 공유되고 공론화가 이뤄지기도 한다[1]. 방송콘텐츠와 관련한 정보에 관해서도 마찬가지다. 유무선 전송망을 통해 방송콘텐츠가 송수신되는 환경이 도래하면서 콘텐츠 이용자들의 의견은 가감없이 소셜미디어 웹 공간에서 공유되고 있으며 이러한 정보들은 정형/비정형/반정형의 측정가능한 텍스트 데이터로 축적되게 되었다. 즉, 메시지, 소셜네트워크, 뉴스 등에서 텍스트들이 대용량화 되어 분석가능한 이른바 '빅데이터'로 생산되고 있는 것이다[2]. 이러한 데이터를 가공하여 측정할 경우, 방송콘텐츠에 대한 영향력을 보다 비개입적(unobstrusive)으로 측정해볼 수 있게 되어 방송콘텐츠의 다양한 측면들을 살펴보는 것도 가능하게 되었다.

본 연구는 이렇듯 소셜미디어에 남겨진 시청/청취 관련 정보들을 통해 측정해볼 수 있는 방송콘텐츠 영향력 지표 및 데이터 분석체계에 대해 살펴보기로 한다. 연구를 통해 다양한 방식으로 방송콘텐츠 영향력과 관련된 지표들을 탐색해봄과 동시에 보다 효율적으로 처리가능한 소셜텍스트 빅데이터 분석체계를 제안해 볼 수 있을 것이라 판단된다.

2. 소셜텍스트 데이터 수집 단위

소셜미디어 상에서 대용량의 텍스트 데이터로 존재하는, 이른바 빅데이터는 알고리즘에 기초한 분석이 가능하기 때문에 시간이나 인력의 절약, 표본오차의 부재, 결측값의 정확한 측정, 코더간의 신뢰도 문제 해소, 다채로운 관계분석, 연구자 편견 개입불가 등 다양한 장점을 지니고 있는 것이 사실이다[3].

하지만 분석차원에서 빅데이터의 장점이 제대로 발휘되기 위해서는 데이터를 수집하는 단계에서부터 데이터 단위가 명확히

1) 본 논문은 주저자 박사학위 논문 중 일부내용을 인용하였음을 밝힙니다.
2) 주저자
3) 교신저자

설정되어야 한다. 그리고 무엇보다 웹 크롤러가 데이터를 수집할 수 있도록 공개된 데이터 형식일 경우에 수집 작업이 용이해진다. 따라서, 방송콘텐츠 영향력 측정에 적합한 소셜텍스트 데이터의 범위를 설정하고 이를 통해 분석결과의 도출단계에서의 정확성을 확보하는데 연구의 초점을 맞추고자 한다.

본 연구에서 설정하는 소셜텍스트 데이터의 개념은 뉴스섹션, 블로그, 커뮤니티 공간에서 키워드로 도출되는 게시글 및 댓글등의 텍스트 데이터를 의미한다[4]. 이러한 텍스트 데이터는 수집과 처리가 비교적 용이한 정형데이터와 2차적 가공처리가 필요한 비/반정형 데이터로 나뉠 수 있다. 분석시 용이성과 정확성을 확보하기 위해서는 HTML 파싱(parsing) 작업과정에서부터 방송콘텐츠 측정이 효율적으로 이루어질 수 있는 방향으로 정/비/반정형 데이터에 대한 개념 설정이 필요하다.

1) 게시글 및 댓글의 수치

게시글 및 댓글의 수치는 이들과 관련된 조회수, 업로드수 등의 계량적 수치를 의미한다. 이러한 데이터들은 수집과정에서 구조화된 정형데이터(structured data) 형태로 수집이 가능하다. 무엇보다 특정 콘텐츠의 영향력이 어느 정도인가를 살펴보기 위해서는 이를 언급한 게시글과 댓글의 빈도에 대한 수집이 선행되어야 하겠지만 이들 게시글과 댓글이 포함된 웹공간 자체의 영향력도 변수로 고려해야 한다.

따라서 뉴스의 경우에는 해당 언론기사의 열독률 및 시청률, 블로그의 경우에는 PV(Page View)·공감·블로거 영향력, 커뮤니티의 경우 회원수・구독수·랭킹 등이 방송콘텐츠 관련 게시글 및 댓글의 정확한 수치산정을 위한 가중치로 고려될 수 있다.

2) 게시글 및 댓글 등의 형태소 수치

방송콘텐츠와 관련해 언급된 세부적인 내용을 살펴보기 위해서는 빅데이터의 자연어 처리 과정(NLP : Natural Language Processing)에서 추출되는 게시글과 댓글의 명사, 형용사, 동사 등의 수치 데이터도 필요하다. 이들 데이터를 수집함으로써 방송콘텐츠의 내용에 따른 영향력을 심층적으로 분석해 볼 수 있기 때문이다.

형태소 텍스트를 수집하는 경우에도 데이터 수집이 완료되고 난 이후, 보다 언급된 내용의 파급력에 대한 정확한 측정을 위해서는 뉴스섹션, 블로그, 커뮤니티 각 공간의 영향력이 측정된 가중치를 고려할 필요가 있다.

이와 같이 본 연구에서 수집하고자 하는 소셜텍스트 빅데이터의 수집단위를 나타내면 아래 <표 1>과 같이 정리가 가능하다.

<표 1> 소셜텍스트 데이터 수집 단위

데이터 형태	게시글	수치	가중치 요소의 예	
정형 데이터	게시글 + 댓글	조회수 + 업로드수	뉴스섹션	열독률, 시청률
			블로그	PV, 공감, 블로거 영향력
			커뮤니티	회원, 구독, 랭킹
비/반정형 데이터	게시글 + 댓글	제목 + 내용 (명사, 동사, 형용사)	뉴스섹션	열독률, 시청률
			블로그	PV, 공감, 블로거 영향력
			커뮤니티	회원, 구독, 랭킹

3. 영향력 도출을 위한 언급 범위의 측정

방송콘텐츠 영역에서 시청률과 청취율 등은 해당 콘텐츠에 대한 효과나 성과 등 영향력 측정을 위한 지표로 여겨져 왔다. 따라서 이에 영향을 미치는 변수와 영향력에 대한 다양한 연구가 이루어져 왔다. 그러나 최근에는 이러한 피플미터방식의 지표가 스마트 미디어나 N-스크린 서비스 등 첨단화된 크로스미디어 시청환경에 적합한가에 대한 논의도 나타나고 있는 실정이다[5]. 즉, 기존의 방식은 다변화된 미디어 이용형태를 측정하기 힘들 뿐만 아니라 이용자의 다양한 평가를 반영하기 어렵다는 의견이 나타나고 있는 것이다.

이렇게 기존 방송콘텐츠를 측정하는 방식의 변화를 요구하는 주장은 크게 두가지로 요약이 가능한데 측정을 위한 방법론을 제안하는 주장과 새로운 개념을 도입해야 한다는 주장으로 나눠볼 수 있다[6]. 본 연구에서는 후자에 중점을 두며, 뉴스(news), 블로그(blog), 커뮤니티(community) 등 다양한 소셜미디어 공간에서 생산되고 있는 소셜텍스트 빅데이터를 통해 기존 방식과는 차별화된 방송콘텐츠 영향력 측정방식을 제안해보고자 하는 것이다.

전술한 소셜텍스트 빅데이터들의 수치들을 통해 실질적인 방송콘텐츠 영향력을 계산하기 위해 언급범위(scope of mention)의 개념을 설정하고 수식을 설정하도록 한다. 즉, 소셜텍스트가 어느 정도 소셜미디어 공간으로 파급되었는가를 살펴보는 기준으로 언급 범위를 설정한 것이다.

언급범위는 방송콘텐츠와 관련된 언급이 확산된 공간과 채널의 개념을 토대로 언급빈도(Frequency), 채널수(Channel)등을 통해 아래와 같이 채널 점유율(Channel Share)수식으로 나타낼 수 있다. 그리고 이를 통해 각각의 소셜미디어 공간은 얼마나 많은 채널에서 해당 내용이 언급되었는가를 알 수 있다.

뉴스섹션의 예를 들면, 방송콘텐츠가 아무리 많은 수치로 언급되었다고 하더라도 특정 매체에 편중되어 언급되었다면 실질적으로 전체 소셜미디어 공간에서 언급 범위가 고르다고 할 수 없는 것이다. 이와 같은 수치가 채널 점유율로 설명될 수 있다.

$$CS = \frac{C}{F} \times 100$$

C : 확산 공간 내 소셜텍스트 언급 채널수
F : 확산 공간 내 소셜텍스트 언급 빈도

위와같은 채널 점유율은 뉴스섹션, 블로그, 커뮤니티 등 전체 소셜미디어 공간에서의 언급범위를 나타내는 지표로 활용될 수 있다. 그러나 각각의 공간별 언급범위를 세부적으로 측정해보기 위해서는 각 공간별로 특징적인 가중치 요소들을 고려해서 각각의 측정 지표를 설정해야 할 것이다.

1) 뉴스섹션

앞서 살펴본 바와 같이 특정 방송콘텐츠가 소셜미디어 상에서 어떠한 정도의 언급 범위를 갖게 되는가는 언급 범위를 통해 살펴볼 수 있다. 하지만 뉴스의 경우 언론사가 많고 송출방식이 다양하기 때문에 가중치에 대한 고려를 통해 실질적으로 1개 뉴스가 가지는 언급범위의 수치가 개별로 고려되어야 하는 측면이 있다. 이를 고려한 수식은 아래와 같이 나타낼 수 있다.

$$N = n_1 h_1 r_1 rs_1 c_1 + n_2 h_2 r_2 rs_2 c_2 + n_3 h_3 r_3 rs_3 c_3 + \cdots$$

N : 특정 언론사의 언급 범위 수치
n : 뉴스 기사 건
h : 뉴스 조회수
r : 뉴스 시청률(뉴스 영상콘텐츠인 경우)
rs : 뉴스 열독률(오프라인 신문이 있는 경우)
c : 댓글의 수

위와 같이 하나의 언론사가 지니는 뉴스 언급에 관한 영향력은 뉴스기사 1건당, 뉴스 조회수, 시청률, 열독률, 댓글의 수치가 모두 고려되어 이들의 합계로 이루어진다. 비/반정형 데이터의 경우에는 각 제목과 내용에 포함된 형태소별 수치들을 계산해 위와 같은 가중치를 부여할 수 있다.

2) 블로그

블로그의 경우에는 페이지를 나열하는 방식에 따라 언급범위의 수치가 다르게 계산될 수 있다. 본 논문에서는 일반적으로 하나의 페이지에 하나의 게시글이 보여질 경우를 기준으로 언급 범위 수치를 아래와 같이 설정한다.

$$B = p_1 s_1 br_1 c_1 + p_2 s_2 br_2 c_2 + p_3 s_3 br_3 c_3 + \cdots$$

B : 특정 블로그의 언급 범위 수치
p : 페이지 뷰
s : 타인이 공감한 수
br : 블로거 영향력 수치(파워블로거인 경우)
c : 댓글의 수

3) 커뮤니티

커뮤니티의 경우에는 회원수, 구독자수, 커뮤니티 랭킹이 언급범위의 수치를 구성하는 주요 요소들이다.

$$C = m_1 s_1 r_1 c_1 + m_2 s_2 r_2 c_2 + m_3 s_3 r_3 c_3 + \cdots$$

C : 특정 커뮤니티의 언급 범위 수치
m : 커뮤니티에 가입된 회원 수
s : 커뮤니티 페이지 구독 수치
r : 커뮤니티 랭킹을 점수화한 수치
c : 댓글의 수

4. 빅데이터 수집단계별 방송콘텐츠 분석 체계

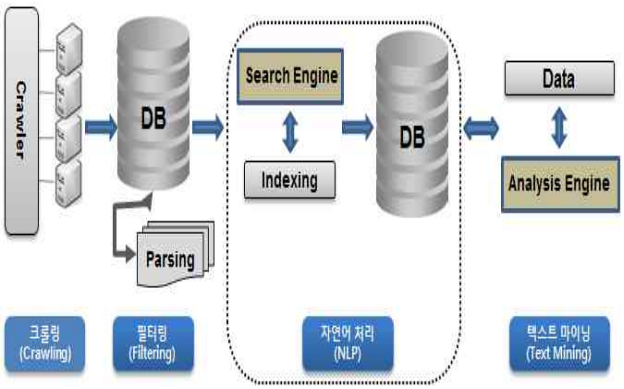
기존에 시청률이나 청취율을 통해 방송콘텐츠의 영향력을 측정하는 방식은 해당 콘텐츠 이용자들을 과학적인 방식으로 샘플링하여 해당 이용자들의 미디어 이용행태를 추적하거나 직접 설문하는 방식이었다. 그러나 이 경우 콘텐츠 이용자들은 조사가 이루어지고 있다는 점을 인지할 가능성이 높고, 질문에 포함된 조사자의 의도를 완벽하게 제거하기도 어려워 완벽한 비개입적 조사라고 보기 힘들었던 것이 사실이다.

소셜텍스트 데이터를 활용한 방송콘텐츠 분석시에도 조사자의 개입을 최소화하기 위해서는 웹크롤링 단계에서부터 영향력 측정을 위한 분석 체계를 명확히 해야 할 것이다.

본 연구에서는 아래와 같이 방송콘텐츠 영향력 분석을 위한 빅데이터 분석플랫폼을 구조화하며, 크롤링-필터링-자연어 처리-텍스트 마이닝 등 세분화된 분석단계별로 어떻게 데이터 분석이 구조화되어야 하는지 고려해야할 사항들을 살펴보고자 한다.

결과적으로는 의미분석의 개념이 포함된 텍스트 마이닝 이전 단계까지의 데이터 수집·분석 과정이 체계화될 경우, 향후 텍스트 마이닝 단계에서의 의미분석도 보다 정확히 이루어지질 수 있을 것이다.

<그림 1> 단계별 소셜텍스트 데이터 분석 체계



1) 크롤링(Crawling)

크롤링 단계에서는 방송콘텐츠가 언급된 텍스트의 게시날짜, 내용, 수량 및 용량, 작성자 정보, URL 등의 데이터가 수집된다. 이 과정에서 중요한 것은 방송콘텐츠와 연관된 내용들 중 영향력 측정에 불필요한 정보들을 수집하지 않도록 하는 것이다. 가령, 방송콘텐츠에는 장르에 따라 다양한 등장인물, 사건, 스토리 등의 정보가 포함되어있는데 해당 내용이 방송콘텐츠 영향력을 측정하는데 전혀 관련성을 가지고 있지 않은 경우도 다수이기 때문에 이러한 정보에 대한 선별 작업이 필요하다.

특히, 최근 우리나라 드라마나 연예인에 대한 관심이 높아지고 있기 때문에 특정 연예인이 드라마에 출연하거나 뉴스에서 기사화가 되는 경우 각 채널을 구분하여 크롤링이 이루어질 수 있도록 체계를 마련해야 한다.

2) 필터링(Filtering)

필터링 과정에서는 일반적으로 키워드가 필터링되거나, 문서가 종류별로 선별되고, 무의미한 정보들이 제거되는 작업이 실행된다.

방송콘텐츠가 소셜미디어상에서 언급되는 경우에 영향력을 측정한다는 기본 취지에 맞게 무의미정보에 대한 선별작업시 고려할 사항이 있다. 시청률과 청취율은 이미 콘텐츠 이용행위가 끝마쳐진 시점을 기준으로 조사된다. 따라서, 빅데이터 분석시에도 제목이나 등장인물에 초점을 맞춰 동일한 텍스트가 무한정 반복되거나, 방송콘텐츠 방송 이전시점에 작성된 텍스트 데이터에 대한 선별작업이 이루어져야 한다.

또한, 필터링 과정에서는 불필요하게 수집된 정보인 가비지(garbage)에 대한 축적도 체계적으로 이루어져야 하겠다. 빅데이터 분석이 전수를 기반으로 하는 조사임에도 불구하고 정확성과 신뢰성을 담보하기 어려운 이유는 바로 이러한 불필요한 정보의 축적 때문이라고 해도 과언이 아니다. 따라서, 해당 가비지가 전체 수집된 정보에서 어느정도의 비율을 차지하고 있는가를 데이터 수집 마지막 단계에서 수치화 해 이를 가비지 스케일(garbage scale)로 공개할 필요가 있다.

3) 자연어처리(NLP : Natural Language Processing)

또한, 자연어 처리 과정에서는 텍스트 데이터에 포함된 형태소나 구문이 추출되면서 실질적인 분석데이터가 색인화되어(Indexing) DB에 축적된다. 이 과정의 DB에 축적된 데이터는 언급 범위를 측정하는 알고리즘에 대입되어 실질적인 분석이 이루어지므로 뉴스섹션, 블로그, 커뮤니티 별 분석데이터가 혼동되지 않도록 수집공간을 기반으로 데이터 분류작업이 우선 이루어져야 한다.

특히 언론사, 블로거, 커뮤니티의 채널명이나 게시글 작성자가 동일하게 나타나는 경우 이를 선별할 수 있게 하기 위해 파급공간명(뉴스섹션-블로그-커뮤니티), 채널명, 해당 제목 및 내용 등을 순차적으로 분류할 수 있도록 체계화 시켜야 할 것이다.

4) 텍스트 마이닝(Text Mining)

본 연구에서 텍스트 마이닝 단계는 텍스트 데이터의 의미분석 단계로 개념을 설정하였으므로 주요 데이터 분석 체계화 과정에서 포함되지 않는다. 그러나, 의미나 네트워크 분석이 이루어지기 이전에 자연어 처리과정에서 잘못 수집된 데이터가 텍스트 마이닝 단계로 전달되는 경우, 다시 해당 데이터를 자연어 처리 단계로 이전해주는 체계를 수립해야 한다.

특히, 방송콘텐츠의 제목에 언급되는 단어들 중 명사, 형용사, 동사를 적절히 구분하되, 해당 형용사나 동사가 텍스트 마이닝 단계의 DB에 축적될 경우 이들 데이터를 자연어 처리 단계의 DB와 연동하여 축적하는 방안을 고려해볼 수 있다.

5. 결론

본 연구는 기존 방송의 영향력을 나타내는 계량적 지표인 시청률과 청취율을 보완해보고자, 방송콘텐츠에 관한 내용이 언급될 경우에 생산되는 데이터의 수집 및 분석체계에 대해 살펴보았다. 시청률과 청취율에 방송콘텐츠의 내용적 판단이 포함되어 있지 않기 때문에 본 연구에서 목표하고자 하는 바는 소셜미디어 상에서 이용자들이 언급하는 정도와 파급 범위정도의 측정을 체계화하는데 초점을 맞췄다.

향후에 보다 심층적인 연구가 이루어지기 위해서는 방송콘텐츠 영향력에 관한 분석결과가 재검증될 수 있는 시스템에 대한 논의가 이루어져야 할 것이다. 또한 방송콘텐츠의 내용적 측면에 대한 분석체계를 논의함으로써 실질적으로 소셜미디어 공간에서 어떠한 방식으로 방송콘텐츠가 논의되고 있는지 심도 깊게 살펴보아야 할 것이다.

참고문헌

[1] 류정호 · 이동훈 (2011). 소셜 미디어로서 마이크로 블로그 심공론장의 정치적 의사소통에 대한 탐색적 연구: 네트워크 동질성 개념을 중심으로. 「한국언론학보」. 제55권 제4호, 309-330.

[2] Dobre, C. and Xhafa, F. (2014). Parallel Programming Paradigms and Frameworks in Big Data Era, International Journal of Parallel Programming, 42(5), 710-738.

[3] 박대민(2013). 뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망 분석. 「한국언론학보」. 제57권 제6호. 234-262.

[4] 최홍규(2014). 정보화 정책 이슈의 확산 과정에 관한 연구 : 소셜텍스트 빅데이터 분석을 중심으로. 고려대학교 일반대학원 언론학 박사학위 논문.

[5] 김관규(2014). 크로스미디어 통합시청률조사의 필요성과 국내외 사례. 「방송문화연구」. 제26권 제1호, 7-32.

[6] 이준웅(2014). 시청률의 해체인가 진화인가? 제도적 유효 이용자와 방송의 미래. 「방송문화연구」. 제26권 제1호, 33-62.