

# SNS 리뷰 데이터 마이닝을 통한 TV 추천 콘텐츠 검증 시스템 설계

김문지<sup>1</sup> 김형지<sup>1</sup> 이지연<sup>1</sup> 김수민<sup>2</sup> 오유리<sup>1</sup> 김윤희<sup>1</sup>  
숙명여자대학교 컴퓨터과학과<sup>1</sup>, 숙명여자대학교 홍보광고학과<sup>2</sup>  
km0420jj@gmail.com, hyungjee91@gmail.com, bandi28@hanmail.net, ddummi@gmail.com,  
yoori0203@sm.ac.kr, yulan@sm.ac.kr

## A Design of Verification System on TV Recommended Contents Using SNS Review Data Mining

Moonji Kim<sup>1</sup> Hyungji Kim<sup>1</sup> Jiyeon Lee<sup>1</sup> Sumin Kim<sup>2</sup> Yoori Oh<sup>1</sup> Yoonhee Kim<sup>1</sup>

<sup>1</sup> Dept. of Computer Science, Sookmyung Women's University

<sup>2</sup> Dept. of Public Relations & Advertising, Sookmyung Women's University

### 요 약

정보화 시대를 맞이하여 각종 정보들이 홍수처럼 쏟아지면서 소비자들의 보다 나은 의사결정을 위한 다양한 TV 추천 콘텐츠들이 등장했다. 한편, SNS의 발달로 인해 시청자들은 TV 추천 콘텐츠에 대한 의견을 개인의 SNS를 통해 자유롭게 작성하고 타인과 공유하고 있다. 그러나 여러 SNS리뷰에서 TV프로그램에서 추천된 콘텐츠가 실제와는 다르게 과장되었다는 의견이 제기되었고, 특히 맛집 같이 영리를 추구하는 업체를 소개하는 경우 실상과 다르게 포장되어 추천되는 경우가 많았다는 의견이 대다수였다. TV 콘텐츠에 대한 사람들의 의구심이 증가됨에 따라 TV 추천 콘텐츠를 검증해줄 수 있는 시스템의 필요성이 제기되었고, 이에 본 연구에서는 SNS 리뷰데이터를 분석하여 TV 추천 콘텐츠를 검증해 줄 수 있는 시스템을 구축하고자 한다. 본 논문에서는 TV 추천 콘텐츠 중 맛집 추천 TV프로그램을 중심으로 하여 해당 프로그램에서 추천된 맛집에 대한 SNS 리뷰 데이터를 하둠을 통해 분석하고, 오피니언 마이닝을 통해 리뷰에 내포된 감성정보들을 추출한다. 결과적으로 추출된 감성정보를 이용하여 TV 프로그램의 만족도를 도출하고 이를 통해 TV 추천 콘텐츠의 공정성을 검증한다.<sup>1)</sup>

### 1. 서 론

최근 다양한 정보 추천 TV 콘텐츠의 등장과 함께 SNS를 통한 상호교환도 활발해졌다. 사람들은 TV 프로그램에서 추천해준 콘텐츠에 대한 이용후기 및 의견 등을 개인의 블로그, 트위터 등의 SNS를 통해 작성하고 공유한다. 그러나 여러 SNS 리뷰들에서 TV프로그램에서 추천된 콘텐츠가 실제와는 다르게 과장되었다는 의견이 제기되면서 사람들은 TV 추천 콘텐츠에 대한 불신을 갖게 되었다. 특히 맛집 추천 콘텐츠들의 경우, 맛집 프로그램에서 추천된 맛집을 방문한 사람들의 불만 후기가 SNS를 통해 공유되면서 맛집 추천 프로그램이 공익의 정보를 제공하기 보다는 음식점을 홍보하는 광고에 가깝다는 우려와 비판이 제기되었다. 이렇게 TV 추천 콘텐츠에서 제공하는 정보에 대한 사람들의 불신이 증가함에 따라, 프로그램에서 제공되는 정보의 진실성과 공정성을 객관적으로 평가해 줄 수 있는 시스템의 필요성이 대두되었다.

본 연구에서는 TV 추천 콘텐츠가 시청자들에게 공정하고 진실성 있는 정보를 제공해주었는지의 여부를 검증하기 위하여 TV 추천 콘텐츠와 관련된 대용량의 SNS 리뷰데이터를 하둠 맵리듀스를 이용해 분석한다. 그리고 오피니언 마이닝 [1]을 활용하여 작성자들의 감성의견을 추출하고, 이를 통계적 기법에 적용시켜 TV 추천 콘텐츠에 대한 만족도를 도출하는 시스템을 제안한다. 본 논문에서는 특히 TV 추천 콘텐츠 중 맛집 추천 콘텐츠를 대상으로 시스템을 구축하고, 이를 결과 화면으로 설계하여 시청자 및 TV프로그램 관계자가 분석의 결과를 시각적으로 볼 수 있게 하였다.

### 2. 관련 연구

기존의 TV 콘텐츠에 대한 시청자 반응분석 방법에는 [2]와 같은 통계적 척도를 이용한 직접적인 설문조사 방식이 대다수였으며, 설문조사 방식의 시간적인 제약과 표본수집의 공간적인 제한이 있다는 단점이 있었다. 한편, SNS가 발달하면서 기존의 설문조사 방식의 시간공간적인 한계점을 뛰어넘을 수 있게 되었고, 이에 따라 여러 방송사에서는 SNS를 통해 많은 시청자들의 반응을 실시간에 가깝게 분석 할 수 있는 SNS 분석 시스템을 구축하였다. 하지만 KBS의 SNS 분석시스템[3]이나 SBS의 통합 리스닝 플랫폼[4]을 통해 제공되는 데이터들은 온라인 리뷰 데이터의 의미해석 없이, 단지 데이터의 양만을 측정변수로 활용하고 있다는 아쉬운 점이 있다.

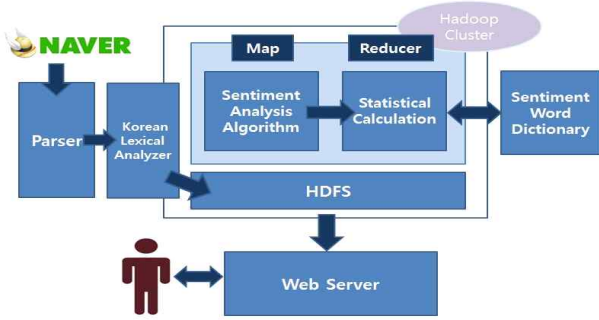
### 3. 설계 및 구현

본 연구에서는 대용량의 SNS데이터를 병렬 저장하기 위한 하둠 기반의 HDFS(Hadoop Distributed File System)를 사용하였으며, 데이터의 효율적인 분석을 위해 맵리듀스 애플리케이션을 사용하였다.

#### 3.1 시스템 설계

TV 추천 콘텐츠 검증을 위해 본 논문에서 설계한 전체 시스템의 구성은 그림 1과 같다. 전체 시스템은 대용량의 SNS 리뷰를 수집하기 위한 데이터 크롤러(Data Crawler), 수집된 데이터를 형태소단위의 표현으로 분류하기 위한 형태소분석기, SNS 리뷰데이터의 오피니언 마이닝을 위한 감성분석기와 사용자에게 분석결과를 보여주는 결과화면으로 나뉜다.

1) 이 연구는 2015년 한국연구재단과 한국여성과학기술인지원센터의 지원을 받아 연구되었습니다



[그림 1] 시스템 구조도

데이터 크롤러에서는 TV 추천 콘텐츠에서 방영된 정보를 기반으로 블로그 리뷰 데이터를 수집하며, 수집 방법은 네이버 개발자 센터[4]에서 제공하는 검색API 라이브러리를 이용하여 PHP언어로 블로그 데이터 수집기를 자체 개발하여 데이터 수집에 활용한다. 형태소의 분석은 오픈 소스로 제공되어있는 KLT 형태소 분석기[5]를 사용하며, 리뷰 데이터들을 형태소 단위의 표현들로 분류한다. 그리고 불필요한 단어와 특수문자 제거의 전 처리 과정을 거쳐 HDFS에 저장한다. 감성 분석기에서는 하둡 맵리듀스를 기반으로 하여 [6]의 연구에서 사용된 PMI기법을 확장한 감성 분석 알고리즘과 선정된 추천 콘텐츠 특징에 맞는 단어 사전을 구축하여 감성 극성을 구분하고, 각 리뷰에 내포되어있는 작성자의 감성을 읽어내어 점수화 한다. 마지막으로 시각화 단계에서는 점수화된 리뷰들을 리커트 5점 척도에 적용하여 아주만족 > 만족 > 보통 > 불만족 > 아주 불만족의 만족도(%)로 표현하여 이를 결과 화면을 통해 보여준다.

### 3.2 시스템 구현

본 논문에서는 TV 추천 콘텐츠 중 맛집 추천 콘텐츠를 실험 대상으로 선정하여, 대표적인 3개의 맛집 추천 프로그램 ‘식신로드’, ‘찾아라 ! 맛있는TV’, ‘테이스티 로드’에서 방영된 맛집에 대한 리뷰 텍스트 데이터 약 6GB를 수집하였다. 또한 하둡 맵리듀스를 기반으로 한 감성분석을 위해 hadoop-2.6.0 버전이 설치된 1대의 Master 서버와 2대의 Slave 서버를 이용하였으며, 운영체제는 Ubuntu 12.04를 사용하였다.

본 시스템의 실험을 위해 맛집 추천 콘텐츠 특징에 맞는 단어사전을 구축하였으며, 과정은 다음과 같다. 먼저 HDFS에 저장된 리뷰 데이터들을 하둡을 통해 워드 카운트 하여 맛과 관련된 서술어와 부사를 중심으로 단어를 추출하고, 이를 긍정, 부정으로 분류하여 데이터베이스에 긍정 사전, 부정 사전으로 각각 저장하여 분석에 활용하였다.



[그림 2] 방영된 맛집의 만족도 리스트 및 해시태그

그림 2은 ‘식신로드’에서 방영한 맛집의 분석 리스트 화면으로 이 중 ‘갈무리 소쿠리’는 만족도가 80%이며, 다음과 같은 과정을 거쳐 점수화 된다. 먼저 ‘갈무리 소쿠리’에 대한 46개의 블로그 데이터들이 형태소 분석과 전처리 과정을 거쳐 HDFS에 저장되면, ‘갈무리 소쿠리’를 key로 ‘갈무리 소쿠리’에 대한 블로그 데이터들을 value로 하여 하둡 맵리듀스의 감성 분석 함수와 구축된 긍/부정의 감성 사전을 거쳐 1차적으로 점수화 된다. 이렇게 점수화된 결과는 리커트 5점 척도에 적용되어 아주만족 17개와 만족 18개, 보통 8개, 불만족 2개, 아주 불만족 1개의 결과로 80%의 만족도를 도출하게 된다. 위와 같은 과정을 통해 식신로드에서 방영된 모든 맛집의 블로그 데이터들이 점수화되며, 점수화된 만족도 리스트들은 다시 리커트 5점 척도에 적용되어 최종적으로 ‘식신로드’의 만족도가 된다. 또한 도출된 만족도와 관련해 사용자들이 SNS에 표현된 실제 방문자 리뷰들의 의견을 요약적으로 볼 수 있도록 “시원하다, 깊은 맛”등의 만족 이유와 “너무 달다, 짜다”의 불만족이유를 해시태그로 보여준다.

### 4. 결론 및 향후연구

본 시스템은 TV 추천 콘텐츠가 시청자들에게 공정하고 진실성 있는 정보를 제공하였는지의 여부를 SNS 리뷰분석을 통해 공개적으로 모니터링 해줌으로써 TV 추천 콘텐츠에 대한 객관적인 판단 기준을 제공한다. 시청자들은 TV 추천 프로그램에서 방영된 정보에 대한 객관적인 판단과 함께 보다 효율적인 의사결정을 할 수 있을 것이며, TV 프로그램 관계자에게는 앞으로의 프로그램 제작에 있어 보다 공정하고 진실성 있는 TV 프로그램 제작을 할 수 있도록 경각심을 줄 것이다. 향후에는 리뷰 작성의 주관성으로 인해 발생하는 일관적이지 않은 글에 대한 분류가 필요하며, 객관적인 분석을 위해 상업적, 광고성으로 작성된 리뷰에 대한 분류도 필요하다. 또한, 본문에서 제안된 TV 추천 콘텐츠뿐만 아니라 모든 TV 콘텐츠를 대상으로 하여 각각의 특징을 고려한 전체적인 TV 콘텐츠 리뷰 분석이 이루어진다면, SNS 리뷰데이터를 통한 TV 콘텐츠 검증의 진정한 목적을 달성 할 것이다.

### 참고문헌

- [1] 박경미, 박호건, 김형근, 고희동 "SNS에서 오피니언마이닝 연구" 정보처리학회지 v.18 no.6, 2011, pp.68-78
- [2] 2014년 ‘방송프로그램 시청자 만족도 평가지수(KI) 조사’,[http://www.kisdi.re.kr/kisdi/common/download?type=D&file=GPK\\_PRESS%7C30218%7C1](http://www.kisdi.re.kr/kisdi/common/download?type=D&file=GPK_PRESS%7C30218%7C1)
- [3] KBS(2014). 시청자 프로그램 품질평가. URL: <http://www.kbs.co.kr/openkbs/customer/sub01.html>
- [4] SBS(2014, 3, 20). 2013 SBS 시청자 만족도(ASI: Audience Satisfaction Index), [http://www.mediacreate.co.kr/information/data\\_view.jsp?seq=3382&brd\\_gubun=K&cpage=1&p\\_SCHSTR=&p\\_SCHBOX=](http://www.mediacreate.co.kr/information/data_view.jsp?seq=3382&brd_gubun=K&cpage=1&p_SCHSTR=&p_SCHBOX=)
- [5] 강승식, “한국어 형태소 분석기와 한국어 분석 모듈”, <http://nlp.kookmin.ac.kr/HAM/kor/download.html>
- [6] Jung-yeon Yang, Jae-seok Myung, Sang-goo Lee, "A Sentiment Classification Method Using Context Information in Product Review Summarization", Journal of Computing Science and Engineering, Vol.36, No.4, pp.18-26, 2009.08 (in Korean)