Twitter기반 K-POP아티스트 상승 키워드 분석 및 군집화

권태용*, 이왕광*, 김성환*, 염성웅+, 김경백+
메이크스타*, 전남대학교 전자컴퓨터공학부+

Twitter based Analysis and Clustering of Trending Keywords for K-POP Artist

Taeyong Gwon*, WangKwang Lee*, SeongHwan Kim*, Sungwung Yeom+, Kyungbaek Kim+
Makestar*
Dept. Electronics and Computer Engineering, Chonnam National University+
E-mail: taaii6569@makestar.co, kwang9092@gmail.com, tjdghks0531@gmail.com, kyungbaekkim@jnu.ac.kr, yeomsw0421@gmail.com

요 약

최근 외국인을 대상으로 한 설문조사 결과에 따르면 한국 연상 이미지 1위가 'K-POP' 일 정도로 K-POP의 인기는 상승하는 추세이다[1]. 이에 따라 아티스트들은 활동 내용들을 SNS를 통해 팬들에게 공유하고 글로벌 팬층의 호응을 유도하고 있다. SNS 중 하나인 Twitter는 K-POP에 대한 이슈, 상황 공유가 활발히 이루어지는 매체이다. 본 논문에서는 Twitter기반 데이터를 기반으로 K-POP 아티스트들의 이슈 되고 있는 상황을 분석하였다.

1. 서 론

최근 K-POP의 세계적인 인기가 이슈가 되고 있다. 외국인을 대상으로 한 설문조사 결과[1]에 따르면 한국 연상 이미지 1위가 'K-POP' 일 정도로 외국인에 대한 한국의 이미지가 K-POP일 정도로 K-POP의 인기는 상승하는 추세이다. 이와 함께 소셜 미디어를 통해 K-POP 커버댄스, 플래시 몹 등의 활동이 이어져 현재의 K-POP열풍이 되었다.

이러한 글로벌한 K-POP 열풍에 따라 한국 엔터테인먼트 회사들 또한 타겟 팬층을 국내 팬층에서 글로벌 팬층으로 확대하였다. 소속 아티스트들의 활동 내용을 글로벌 매체인 SNS를 통해 팬들에게 공유하고 글로벌 팬층의 호응을 유도함으로써 해외 시장진출의 발판을 마련하고 있는 추세이다[2]. SNS 중 하나인 Twitter는 K-POP에 대한 이슈, 상황 공유가 활발히 이루어지는 매체이다. 간단한 글 등록 방법과 리트윗을 통해 자신이 좋아하는 아티스트들의 글을 공유하거나 작성함으로써 자신의 팬심을 표출하고 있다. 본 논문에서는 이러한 팬들이 올린 트윗 내용과 아티스트들이 공유한 자신의 공연내용 등의 트윗을 분석하여 아티스트별 팬들의 관심사와 현재 아티스트의 이슈를 해시태그 기반으로 분석해보고 관련 해시태그들을 군집화 하는 분석을 진행하였다.

2. 관련 배경 기술

2.1. AWS Athena

아마존의 대화식 쿼리 서비스 Athena[3]는 CSV, JSON, Avro 또는 컬럼 방식 데이터 형식으로 S3에 저장된 데이터를 호출 할 수 있다.

2.2 상관분석

상관분석[4] 또는 '상관관계' 또는 '상관'은 확률론과 통계학에서 두 변수간에 어떤 선형적 또는 비선형적 관계를 갖고 있는지를 분석하는 방법이다. 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수로p를 사용하며 표본 상관 계수로 r을 사용한다.

상관관계의 정도를 파악하는 상관 계수는 두 변수 간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다. 두 변수 간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해 볼 수 있다.

2.3 Graph Algorithm - PageRank

페이지랭크(PageRank)[5]는 월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 중요도에 따라 가중치를 부여하는 방법이다. 이 알고리즘은 서로 간에 인용과 참조로 연결된 임의의 묶음에 적용할 수 있다.

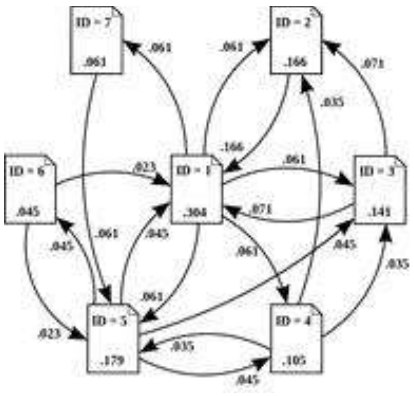

그림 1 페이지 랭크 알고리즘 개념도

3. 초기 설계 및 평가

본 논문의 분석 대상 K-POP 아티스트는 약 1300 이다. 위 아티스트들의 많은 양의 대규모 트윗 데이터를 저장 및 데이터 쿼리의 효율성에 맞춰 설계하였다. AWS S3에 압축하여 저장하고 압축파일에 직접 쿼리를 보낼 수 있도록 S3 버킷의 구조는 '/dt=yyyymmdd/ idx=integer/ tweet_json.gzip'와 같은 형식으로 지정 하였다. 위 경로에서 dt는 크롤링한 날짜이고 idx는 아티스트를 지칭하는 임의의 번호이다. 위와 같이 저장 폴더 주소를 설정한 뒤 Athena을 통해 dt, idx를 변수로 쿼리를 보내 데이터를 추출할 수 있다.

본 논문에서 검증한 데이터 분석은 각 아티스트 별 3일간의 데이터를 분석하였다. 분석의 예를 들기 위해 이하 논문 내용에서 '마마무' 걸그룹 관련 트윗으로 분석 결과를 검증한다. 수집된 트윗의 내용은 그림(2)과 같다.

	text	hashtags	lang
0	191003 KDF CONCERT\n\nUFU AD CAP_ORANGE - ₩ 39...	[MAMAMOO, WHEEIN, 마마무, 휘인]	in
1	191003 KDF CONCERT\n\nUFU AD CAP_ORANGE - ₩ 39...	[MAMAMOO, WHEEIN, 마마무, 휘인]	in
2	I still get teary when I hear this song, so, f...	[mootober, MAMAMOO]	en
3	I still get teary when I hear this song, so, f...	[mootober, MAMAMOO]	en
4	Soooo funny and cuteee 😍😍😍 #solar #moonbyul #m...	[solar, moonbyul, mamamoo]	en

그림 2 마마무 트윗 예시

'hashtags' 컬럼은 유저가 하나의 트윗에 태그한 해쉬태그 리스트값 이다. 위 트윗을 'created_at' (트윗 생성일)을 기준으로 그룹화 한 뒤 일별 트윗을 그래프로 시각화 한다. 분석결과 그림(3)와 같이 19년 10월 15일 13시에 가장 많은 트윗이 수집되었다.

created_at	hashtags	count
2019-10-14 00:00:00	[[MAMAMOO, WHEEIN, 휘인], [MAMAMOO, WHEEIN, 휘인],...	48
2019-10-14 01:00:00	[[MAMAMOO, SOLAR, MONBYUL, moonsun], [MAMAMOO,...	38
2019-10-14 02:00:00	[[AutumnWithHwasa, HWASAXWOOGIE, HWASA], [Autu...	17
2019-10-14 03:00:00	[[BOF, 원아패, 대리찍사, 댈찍, 하성운, AB6IX, 아스트로, 스트레이키즈...	54
2019-10-14 04:00:00	[[더쇼], [더쇼], [김재환, 하성운, 뉴이스트, AB6IX, 아스트로, 슈퍼주...	32
2019-10-14 05:00:00	[[더쇼], [더쇼], [MOONBYUL, Yein, Soojin, YooA, Ch...	48
2019-10-14 06:00:00	[[kpoppoly, 케이팝쓸리, BTS, CLC, Twice, EXO, Seven...	33
2019-10-14 07:00:00	[[14ottobre, MondayMotivation, AutumnWithHwasa...	59

그림 3 시간별 트윗 횟수

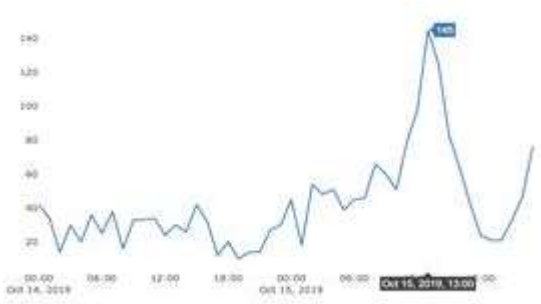

그림 4 시간별 트윗 횟수 그래프 시각화

그림(4)의 그래프에서 15일 경 가장 큰 상승폭이 발견 되었고 14일과 15일 사이에는 작은 상승폭들이 빈번히 발생하였다. 위 상승폭의 트윗 내용을 파악하기 위하여 각 해시태그들의 일별 태그 횟수를 그림(5)의 테이블과 같이 나열한 뒤 이를 그래프로 시각화하여 트윗 그래프와 그림(6)의 해시태그 그래프를 통해 상승폭을 비교한다.

WHEEIN	휘인	mootober	solar	moonbyul	더쇼	InTheFall	HWASA	HWASAXWOOGIE	AutumnW
2.0	4.0	2.0	2.0	2.0	2.0	14.0	16.0	14.0	
0.0	0.0	0.0	4.0	2.0	2.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	4.0	0.0	2.0	2.0	
4.0	5.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	
2.0	4.0	0.0	0.0	0.0	4.0	0.0	2.0	2.0	

그림 5 시간별 해시태그 횟수

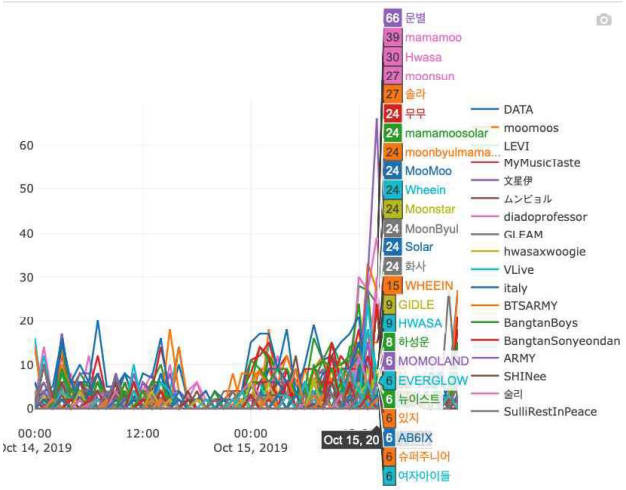

그림 6 일별 해시태그 횟수 시각화

위 그림(4)과 그림(6)에서 그래프는 15일 12시에서의 상승 그래프가 유사함을 확인할 수 있다. 이 구간에서 '문별' 태그가 가장 많이 언급되었다. 위 해시태그들의 일별 변화량이 가장 높은 태그를 아티스트 이슈키워드라 판단하고 이를 확인하기 위해 일별 변화량이 높은 해시태그를 N개 선택한 뒤 시각화 하였다.

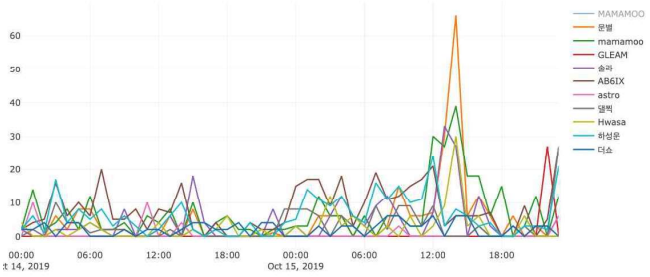

그림 7 변화량이 큰 해시태그의 일별 언급횟수

그림(7)은 변화량이 큰 해시태그들을 시각화한 것이고 해시태그들 중 그래프의 형태가 유사한 그룹들을 발견할 수 있다. 이는 주제가 유사한 트윗의 경우 동일한 해시태그를 사용하기 때문이다. 이러한 해시태그들의 유사성을 파악하고 군집화하기 위해 해시태그 간에 상관관계 분석을 실시한다.

그림(5)의 시간별 해시태그 횟수 테이블을 기반으로 그림(8)을 에서의 상관관계분석 결과 테이블을 기반으로 태그간의 상관관계를 확인한다. 동일한 태그들 사이의 상관 유사도는 1의 유사도 값을 확인할 수 있고 각 태그들 간의 유사도 값이 0.3 이상일 경우 태그간 유사도가 있다. 판단한 뒤 태그 횟수가 더 많은 태그로 군집화하였다.

	MAMAMOO	WHEEIN	휘인	mootober	solar	moonbyul	mamamoo	
MAMAMOO	1.000000	0.748427	0.718241	-0.032698	0.444631	0.685950	0.698710	(
WHEEIN	0.748427	1.000000	0.866211	-0.090355	0.305125	0.550173	0.694906	(
휘인	0.718241	0.866211	1.000000	-0.018782	0.354419	0.689983	0.618147	(
mootober	-0.032698	-0.090355	-0.018782	1.000000	-0.113721	-0.104517	-0.165663	-(
solar	0.444631	0.305125	0.354419	-0.113721	1.000000	0.405747	0.540837	(
moonbyul	0.685950	0.550173	0.689983	-0.104517	0.405747	1.000000	0.613701	-(
mamamoo	0.698710	0.694906	0.618147	-0.165663	0.540837	0.613701	1.000000	(
더쇼	0.092484	0.287717	0.192345	-0.071566	0.038393	-0.058191	0.205470	'
InTheFall	0.144035	-0.002955	-0.027678	0.362806	-0.005260	-0.020267	-0.045599	-(
HWASA	0.645505	0.358096	0.330592	0.020315	0.162434	0.564006	0.362269	(
HWASAXWOOGIE	0.090282	-0.031980	0.028873	0.253449	-0.148155	0.053579	-0.062900	(

그림 8 해시태그 상관관계 유사도 테이블

군집화 키워드를 그림(9)와 같이 시각화 한 결과 7개의 태그로 군집화 되었고 14일부터 16일 사이의 '마마무' 이름의 아티스트들의 이슈 키워드를 확인할 수 있다.

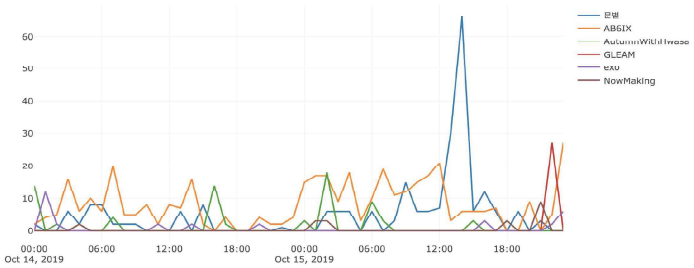

그림 9 군집화 키워드 시각화

군집된 키워드들의 언급 총합을 가중치로 하여 군집별 워드 클라우드를 그림(10)과 같이 생성하여 키워드 내용을 확인해 보았다.

그림 10 각 군집화 태그들의 워드클라우드

그림(10)은 '화사', 방탄소년단', 'KNK(크나큰)', '박봄' 관련 트윗의 키워드를 워드클라우드로 시각화한 결과이다. 각 태그들은 '화사 - 가을속에서', '방탄소년단 - PeopleChoiceAwards 투표', '크나큰 - 죽네투어 방송출연', '박봄 - 퀸덤 출연' 등을 기반으로 이슈화 된 것을 확인할 수 있다.

4. 결 론

본 논문에서는 세계적인 SNS 플랫폼인 Twitter을 통해 공유되는 K-POP 아티스트들에 대한 일상 공유 활동 정보와 함께 제공되는 해시태그 기능을 기반으로, 팬들의 관심사 및 아티스트의 이슈 상황을 분석할 수 있는 인기 상승 키워드 분석 기법을 제안하고, 이를 실제 기사, 공연, 음반 활동과 비교하여 기법의 실행 가능성을 검증하였다.

향후 제안된 기법을 기반으로 아티스트뿐만이 아니라 여러 주제들을 선정하여 기간별 관심사들을 데이터화하고 신뢰성을 높이려 연구하여 볼 것이다.

Acknowledgements

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터 지원사업의 연구결과로 수행되었음 (IITP-2019-2016-0-00314). 이 논문은 정보(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF- 2017RIA2B4012559).

참 고 문 헌

[1]http://file.mk.co.kr/imss/write/20190418134426__00.pdf
[2]http://commres.net/wiki/_media/k-pop_한류의_성공요인분석과_한류_지속화_방안연구.pdf
[3]https://aws.amazon.com/ko/athena/
[4]https://ko.wikipedia.org/wiki/%EC%83%81%EA%B4%80_%EB%B6%84%EC%84%9D
[5]https://ko.wikipedia.org/wiki/페이지랭크