creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

트위터 사용자의 음악 청취 행태 와 음악 트렌드의 관계성 분석 및 예측

2015 년 2 월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

김 예 경

국문초록

음악이 디지털화되면서 음악의 소비와 유통이 온라인 플랫폼을 통해 이루어지고 있다. 이에 따라 음악 사용자들의 청취 행태 또한 온라인을 중심으로 변화하며 온라인 상에 사용자-음악 관련 다양한 정보가 축적되고있다. 생성된 정보들은 연구자들에게 귀중한 정보원으로서 다양한 방면의 연구가 활발하게 진행되고 있다.

그러나 사용자 관련 정보들은 한정적으로 제공되고 있으며 사용자들이 언제 어떠한 음악을 듣는지에 대한 '사용자 음악 청취 행태'에 대한 정보는 제공되고 있지 않다. MIR(Music Information Retrieval)분야에서 활발하게 연구되고 있는 음악 추천 시스템, 흥행 예측과 같이 음악 사용자와 긴밀하게 연결되어있는 연구 주제에서는 이러한 정보의 부재가 연구의 한계를 초래할 수 있다. 때문에 최근의 연구에서는 마이크로 블로그를 통해 사용자 음악 청취 행태를 규명하려는 시도가 있어왔다. 하지만 이러한 연구는 사용자 그룹 분류나 음악적 특성에 따른 음악 청취 행태의 차이점에 대한 고려는 이루어지지 않았다. 또한 전체 트윗을 사용하지 않고 제한적 정보를 이용했다는 점에서 한계가 있다.

본 연구에서는 대표적인 소셜 미디어중 하나인 트위터의 해쉬태그를 이용하여 사용자 음악 청취 행태를 수집하고 음악 트렌드를 중심으로 분석하려 한다. 첫 번째로 생성한 트윗 수를 기준으로 사용자를 분석한다. 두 번째로 분류된 사용자 그룹에서 추출한 각각의 특성과 빌보드 음악각 사용자 그룹별 음악 청취 행태와 트렌드와의 분포를 통해 어떠한 경향성을 가지는지 알아보고 특징을 정의한다. 세 번째로 음악의 장르와 아티스트 선호도에 따라서 사용자별 음악 청취 행태가 어떻게 달라지는지 관찰한다. 나타난

특징을 반영하여 새로운 특성을 제시한다. 네 번째로 추출된 특성을 기반으로 빌보드 음악 순위를 예측하는 모델을 수립하고 예측 결과와 실제를 비교하여 정량적인 평가를 수행한다.
이러한 과정을 통하여 트위터의 음악 청취 행태를 사용자 그룹별로 트렌드와의 관계성을 규명하고 차이점을 보인다. 나타난 특징들은 사용자들의 음악 청취 행태에 있어서 다양한 요소에 따른 선호도를 추론하는데 사용될 수 있을 것이다.

주요어: 음악 청취 행태, 트위터, 관계성 분석, 소셜 데이터 마이닝, 예측 모델
학번: 2013-22403

목 차

제 1 장 서론 1
제 1 절 연구 배경 1
제 2 절 연구 목적 4

제 2 장 관련 연구 6
제 1절 선행 연구 6
1. 트위터 6
2. #nowplaying 8
3. 트위터와 음악 청취 행태 9
4. 트위터를 통한 예측 10
5. 음악 흥행 예측 12
제 2절 기술적 배경 12
1. API 12
2. 회귀 분석 13
3. 의사 결정 트리 13

제 3 장 연구 문제 15
제 4 장 연구 방법 17
제 1 절 데이터 수집 17
1. 트위터 17
2. 빌보드 차트 19
3. 음악 장르 20
제 2 절 전처리 과정 20

제 4 장 분석 및 실험 22
제 1 절 음악 청취 행태 분석 22
1. 사용자 분석 22
2. 사용자 그룹별 음악 트렌드와 클러스터 24
3. 음악 순위와 특성간의 관계 분석 28

4. 발매 이후 시간과 특성간의 관계 분석 39
5. 음악 차트와 특성간의 상관 관계 41
6. 새로운 특성 제시 및 음악 차트와 상관 관계.....44
제 2 절 음악 순위 예측 46

제 5 장 결론 50
제 1 절 연구 요약 50
제 2 절 연구 의의 51
제 3 절 연구 한계점 51
1. 부족한 데이터 52
2. 실험 설계 및 통계적 검증.....52

제 1 장 서론

제 1 절 연구 배경

마이크로 블로그(micro blog)와 같은 소셜 미디어(social-media)의 성장으로 인해 풍부한 데이터들이 온라인 상에서 사용가능해졌다. 소셜 미디어는 사용자들에게서 생성된 가치있는 사용자 관련 정보들을 제공하기 때문에 중요성은 계속적으로 증가하고 있다. 특히 마이크로 블로그의 대표적 서비스인 트위터와 페이스북의 전세계 사용자는 각각 5 억, 9 억으로 20 억에 이르는 세계 인터넷 사용자중 50%이상에 달한다. 이는 중복을 제외하여도 전 세계의 반 이상의 인터넷 사용자가 마이크로 블로그 서비스를 사용한다는 것을 보인다[30]. 사용자들은 이러한 소셜 미디어를 통해 상호 교류를 하고 자신의 다양한 생각이나 일상을 공유한다. 이렇게 모여진 정보를 통해 전세계 사람들이 어떠한 생각, 행동 등을 공유하는지 실시간으로 마이크로 블로그를 통해 관찰할 수 있다는 점에서 마이크로 블로그는 사용자 중심의 연구를 한데 있어서 귀중한 정보원이라고 할 수 있다. [1,18]

음악은 이전부터 인간의 삶에서 중요한 역할을 하며 하나의 문화로서 자리잡아왔다. 많은 사람들이 음악을 즐기고 소비함에 따라 음악 시장은 거대한 규모를 지니게 되었다[7]. 인터넷의 발전과 함께 음악의 형태가 디지털 매체로 옮겨가면서 디지털 음악은 하나의 중요한 상업요소로

떠오르게 되었다. 닐슨에 따르면 2011 년도에 처음으로 디지털 음악이 오프라인 음악 판매를 넘어섰으며, 그 이후에도 상승세를 유지하고 있다. 또한 2013 년 음악 스트리밍 수는 1 천억을 넘어서는 기염을 통하며 끊임없이 성장하고 있다. 이처럼 발전되어가는 음악산업의 형태에 따라 사람들의 음악 소비 구조는 오프라인에서 온라인으로 이동하고 있다. 사용자들의 청취 행태 역시 온라인을 중심으로 변화하며 온라인 상에 사용자-음악 관련 다양한 정보가 축적되고 있다. 더불어 Last.fm, Spotify 와 같은 거대 온라인 음악 서비스들은 하나의 소셜미디어로서 음악을 유통하기에 소셜 미디어의 성격을 띄고 있어 사용자들은 다른 사용자들과 교류가 가능하다. 또한 소셜 미디어 내에서 음악에 대한 정보와 다양한 생각을 공유하기에 소셜 미디어와 음악은 밀접한 관계를 가지고 있다.[37]

그러나 일반 연구자들이 공개적으로 접근할 수 있는 데이터가 한정적이라는 문제점이 있다. 특히 실시간 사용자 음악 청취 행태에 관한 데이터는 Last.fm 에서 제공되고 있으나, 사용자의 아티스트 별 청취 행태를 제공한다는 점에서 전체적인 사용자와 음악의 정보를 알기 힘들다. 사용자 관련 음악 정보 데이터의 대표적인 예로는 사용자별 음악 평점이 있다. 이러한 데이터는 아마존(Amazon), 야후(Yahoo) 등을 통해 얻을 수 있어 협업 필터링을 이용한 추천 시스템 등에서 데이터 원천으로서 사용한다. 그러나 제공되는 평점만으로는 한계가 있어 사용자의 접근 패턴과 같이 다양한 방법으로 문제를 해결하려는 시도가 있어왔다[8]. 하지만 이 역시 청취 시각과 같은 메타데이터를 고려하지 못했다는 점에서 개선의 여지가 있다. 이는 소셜 미디어 속에서도 마찬가지의 양상인데, 거대 온라인 음악

서비스 중 Last.fm 만이 API(Application Programming Interface)를 제공하고 Last.fm 을 제외하고는 소셜 미디어 데이터를 이용한 음악 정보 검색 연구는 거의 존재하지 않고 있다[37]. 또한 Last.fm 역시 한정적인 데이터만을 제공하며 여기에는 시간별 사용자 음악 청취 행태가 포함되어있지 않다.

앞서 언급한 소셜 미디어의 특성을 미루어보았을 때 소셜미디어는 전세계의 많은 수의 사용자가 자신의 일상 요소를 공유하고 실시간이라는 점에서 사용자 음악 청취 행태를 관찰하기에 적합한 매체이다. 그러나 소셜미디어와 음악의 교차점 연구는 거의 진행되고있지 않다. 이러한 정보를 대상으로 제공하는 소셜미디어는 존재하지 않기 때문에 소셜 미디어에서 사용자 음악 청취 행태를 추출해낼 필요성이 있다.

거대한 사용자 커뮤니티를 통해 확산되는 막대하고 다양한 정보는 집단 인구에 대한 의견을 수렴하고 그들의 행태로부터 의미 있는 통찰을 가능하게 한다. 또한, 특정 개체에 대한 사용자들의 대화 수집은 마케팅이나 홍보에 도움이 될 수 있다[2]. 전세계를 대상으로 하는 트위터의 실시간성은 기존의 데이터베이스에는 존재하지 않는 귀중한 정보이며 이를 이용하여 시공간적으로 다양하게 변화하는 음악 청취 행태에 유연하게 대처할 수 있을 것이다. 또한 트위터는 사용자와 지역적 메타데이터를 제공하기 때문에 추가적으로 다양한 정보를 수집할 수 있어 보다 넓은 분야의 연구에 적용될 것이다.

본 연구에서는 트위터의 음악 청취 행태 데이터를 수집하고 이들의 특성을 추출하여 분석한다. 이를 바탕으로 사용자의 음악 청취 행태를

분석하고 사용자 분석을 통하여 서로 목적이 다른 사용자 그룹을 분류한다. 분류된 사용자를 중심으로 음악 청취 행태와 트렌드 사이의 관계성을 살펴보고 그 차이점에 대해서 분포와 장르, 아티스트 선호도의 요소로 특징을 분석한다. 도출된 특징을 통해 새로운 특성을 제안하며 이들 특성이 음악 순위 예측에서 얼마나 효과를 발휘할지 평가한다. 이를 통해 트위터 사용자의 음악 청취 행태가 일반적인 음악 산업과 어떠한 연관이 있는지 알아보고 새로운 가능성을 제안한다.

제 2 절 연구 목적

본 연구에서는 트위터의 음악 청취 행태를 활용하여 음악 트렌드와의 관계성을 분석하고 이를 이용한 예측 task를 바탕으로 그 효용성을 평가하는데 있다. 대상으로하는 음악 트렌드의 기준은 공신력 있는 음악 차트인 빌보드 순위이다. 음악 청취 행태 관련 트윗을 생성한 트위터 사용자를 목적과 메타데이터에 따라서 사용자 그룹을 분류한다. 분류된 사용자의 음악 청취 행태와 음악 트렌드의 분포를 통해 차이점을 분석하고 비교한다. 또한 장르나 아티스트 선호도와 같은 요소를 바탕으로 사용자 그룹별 특징을 비교하고 어떠한 차이점과 공통점을 갖는지 관찰한다. 이때 발견한 특징을 바탕으로 음악 트렌드를 보다 잘 설명할 수 있는 feature를 제시한다. 마지막으로 추출한 특성의 단계별 모델링을 통해 음악 순위를 예측하는 task에서 각 특성이 얼마나 중요한 요소인지를 보인다.

위의 일련의 과정을 통하여 트위터의 음악 청취 행태를 사용자 그룹별로 트렌드와 함께 분석하고 외부 요인에 따른 차이점을 규명한다. 또한 그 차이점이 왜 생기는지에 대해서 분석하였고 발견한 특징을 통해 새로운 측면을 관찰할 수 있는 feature의 가능성을 발견하였다는 점에서 본 연구의 의의가 있다. 제시된 음악 청취 행태 데이터와 그 특징은 보다 사용자의 음악 선호도에 대한 다양한 연구를 가능하게 할 것이며 음악 흥행 예측과 같은 분야에서 의미 있는 특성으로서 사용 될 것이다.

제 2 장 관련 연구

제 1 절 선행 연구

1. 트위터

마이크로 블로그는 소셜 미디어의 한 종류로, 블로그보다 작은 형태를 가리킨다. 짧은 텍스트 메세지와, 즉각적인 메세지의 전송, 업데이트를 제공받는 것의 세가지의 특징이 있다[34]. 이전부터 사용해왔던 소통 형태(e.g., 이메일, 웹 블로그, 인스턴트 메세지)와는 차별화되어 하나의 일상적인 통신수단으로 자리매김하였다[1]. 트위터는 대표적인 마이크로 블로그 서비스의 하나로, 2006 년에 시작한 이래로 빠르게 증가하며 2012 년 기준으로 5 억 명의 사용자가 등록되었다[2, 22]. 트위터에서는 '트윗'(tweet)이라 불리는 140 글자의 제한이 있는 포스트를 모든 사용자가 작성할 수 있다[34]. 트위터 내의 메세지는 2013 년 기준 하루에 2 억만명의 사용자들에 의해 4 억만건에 생성되기에 이르른다. 이러한 트위터의 시스템적 제한은 사람들로 하여금 포스트의 내용에 간결함을 부여하고 효율적으로 단어를 사용하도록 했다[1]. 사용자는 자신이 업데이트를 받기를 원하는 사용자를 follow 하며, 반대로 다른 사용자가 자신의 업데이트를 받아본다면 그것은 follower 라 한다.

해쉬태그(#)는 특정한 주제를 나타내는 트윗을 분류할 때에 사용된다. Hash(#)라는 심벌과 문자의 조합으로 사용되며 이는 포스트 내의 키워드

단어를 지칭한다. 해쉬태그를 기반으로 사용자들은 같은 해쉬태그를 가지는 트윗을 분류 할 수 있으며 이를 통해 다른 트윗과의 링크가 가능하다[8]. 그림 2 는 #nowplaying 을 검색한 쿼리를 통해서 전달 받은 #nowplaying 을 포함한 결과 트윗 예시이다.

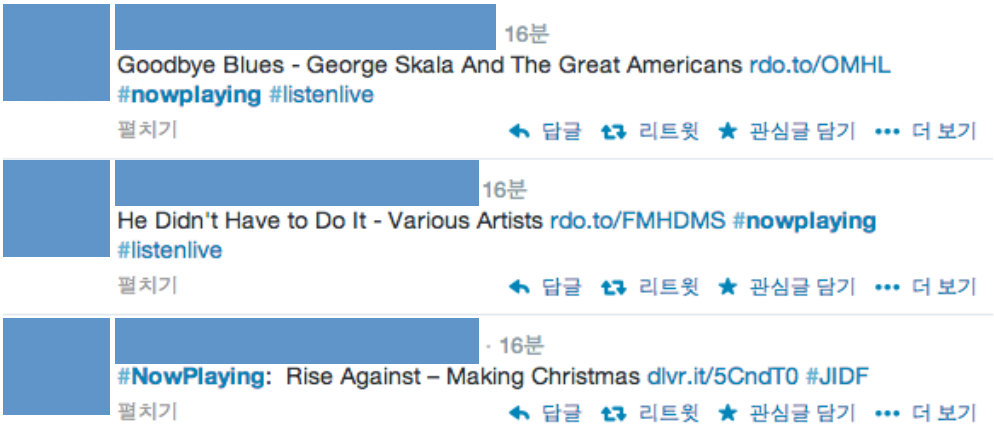

그림 1. 트위터에서 #nowplaying 키워드로 검색 결과

해쉬 태그는 특정 주제로 분류할 수 있다는 점과 140 자로 제한된 트윗에서 주요한 Topic 단어를 나타낼 가능성이 크다는 점에서 해쉬태그는 다양한 연구에 사용된다[8,15,23]. 이 태그는 사용자들에게 의해 만들어지며 동일한 의미에 대해 다양한 키워드로 생성되기도 한다. 만들어진 해쉬태그가 다른 사용자들에게 수용될 수 도 있고 거부될 수도 있으며, 대부분의 경우 생겨난 뒤 바로 사라지는 짧은 수명을 가진다[8].

리트윗은 특정 사용자의 글을 다른 사용자들이 볼 수 있도록 전달하는 기능이다[6]. 리트윗은 정보가 흥미있는 포스트를 전달하는데 유용한

역할을 하며, 트위터 커뮤니티에서 정보 확산의 중심적인 기능을 하고 있다[2]. 대부분 트윗 앞에 'RT'와 같은 양식으로 표시하며 다른 양식도 여럿 존재한다[6].

트위터는 실시간으로 다양한 정보를 빠르게 전달한다는 특징을 통해 전세계에서 어떠한 현상이 일어나고 있는지 효과적으로 관찰 가능하다[2]. 이러한 특성을 이용한 관련 연구로는 병의 전이[26], 지진 예측[30], 관심이 많은 주제[7] 예측이 있다.

2. #nowplaying 연구

#nowplaying(=#np)는 사용자가 현재 어떠한 곡을 재생하고 있는지에 대한 주제를 나타낼 때 사용하는 키워드이다. #nowplaying 은 2009 년 이전부터 존재했으며 현재 하루에 450,000 건의 관련 트윗이 생성되는 주요 해쉬태그중 하나이다. 많은 수의 해쉬태그가 생겨나자마자 사라진다는 점을 고려할 때[8] 대다수의 사용자들이 받아들이고 오랜시간 하나의 약속으로서 지켜지고 있다는 것을 알 수 있다. 전체 해쉬태그 사용양에서 상당히 높은 순위를 차지했으며, 계속적으로 인기를 유지하고 있다[40]. 많은 수의 해쉬태그가 금방 사라진다는 점으로 미루어보았을 때 이는 주목할 만한 현상이다.

[18]에 따르면 음악 관련 트윗은 다음과 같은 패턴을 가지고 있다.

- Song title by artist name [on some platform]
- Artist name: "song title"

- Song title #artist name
- Song title – artist name
- Artist name – song title

그러나 #nowplaying 이 태깅되어있더라도 마이크로 블로그 데이터에는 제한이 존재하지 않기 때문에 반드시 음악 청취 행태와 연관이 있는 트윗이 아닐 수 있다. 예를들어 #nowplaying 은 음악 청취 이외에도 영화나 TV 를 관람하거나, 이외의 전혀 상관없는 다른 곳에도 사용할 수 있다. 또한 음악의 가사와 같이 음악 제목이 포함되지 않은 트윗 역시 존재한다. 이런 경우 분석에 어려움을 겪을 수 있다[16].

3. 트위터와 음악 청취 행태

소셜 음악 서비스 Last.fm 이 대부분의 음악 정보를 제공하기에 이를 벗어나고자, 웹에서 음악 관련 데이터를 추출하려는 시도는 계속적으로 있어왔다. 2012 년부터 최근까지 사용자 청취 행태와 관련이 있는 해쉬태그를 사용한 트위터 사용자의 음악 청취 행태 관련 연구가 진행되기 시작하였다.

수집된 트위터 음악 청취 행태 관련 트윗을 통해 아티스트 유사도, 인기도와 지역적 트렌드를 이용하여 지역적으로 음악 청취 패턴에 대한 그래픽 프래임워크를 제시하였다[16]. 이는 장르적 특성과 아티스트 유사도와 같은 MIR 분야의 연구 주제를 적용했다는 점에과 이를 그래픽 비주얼라이제이션을 했다는 점에서 의미가 있다. 이후 Million Musical Tweet Dataset(MMTD)라는 새로운 데이터 베이스가 제안 되었는데 이는

last.fm 등 에서 사용자의 지역적 정보를 주지 않는 점을 문제점으로 들며, 지역 정보가 잇는 트위터만을 대상을 데이터로서 사용하였다. #nowplaying, #np, #iTunes, #musicmonday and #thisismyjam 과 같은 키워드를 사용하였다. 시공간에 따른 트윗의 수를 장르에 따라 관찰하고 이를 바탕으로 장르에 따른 나라별 음악 청취 패턴을 시각화하였다[17].

[20]에서는 'Streamwatchr'라는 전세계의 실시간 음악 청취 행태 시각화 시스템에 대한 컨셉을 제안하였다. 1) 현재 어떠한 음악과 아티스트가 관심을 받고 있는지 2) 새로이 발견된 음악 3) 지역별,전세계의 음악과 아티스트의 인기도 통계, 음악 정보를 추출하는데 있어서 page extraction, baseline extraction, fuzzy extraction 등 다양한 방법을 사용하였다.

그러나 이러한 연구는 지역적 특성에 집중하고 시간에 따른 음악 청취 패턴은 비교적 관심을 두지 않았다는 점과 샘플수가 작다는 점에서 개선의 여지가 있다[16,17,20]. 또한 시각화를 중심으로 연구가 제안되고 있고 실제 음악 청취 행태와 얼마나 밀접한 연관이 있는지는 고려되지 않았다.

4. 트위터를 통한 예측

강력한 컴퓨팅 파워를 기반으로 인해 거대한 데이터의 기계 학습이 가능해졌다. 이를 바탕으로 소셜 미디어의 집단 데이터를 통해 특정 개체의 미래의 트렌드를 예측하려는 다양한 연구가 존재한다.

이전 연구에서는 소셜 데이터를 통해 박스 오피스 흥행 예측 [2], 독일 대선 예측[34], 지진 예측[36] 등을 하였으며 빌보드 관련 예측은 후에 2.1.5 에서 중점적으로 다룬다. [36]은 지진을, [2]에서는 24 가지 영화제목을

키워드로 트위터를 수집하였고, [34]는 전체 트위터에서 정치적인 트윗을 분류하는 방식으로 데이터를 수집하였다. 트위터를 바탕으로 만든 예측 모델이 실험을 통해 검증되어 트위터가 각 개체의 예측에 유효한 특성이라 제시하였다[2,34,36].

5. 음악 흥행 예측

Hit Song Science는 기계 학습(Machine Learning)에 기반을 두고 있으며 음악의 특성에 따라 음악이 흥행할 것인지 예측 가능한지에 대한 연구이다. 음악 산업에서 관심을 가질만한 주제로 MIR분야에서 많은 관심을 가지고 연구가 되었지만 현재는 그 사용 가능성에 대해서 의문점이 남아있는 상태이다[11,28]. 초기에는 오디오 신호들과 가사들을 특성(feature)으로 한 기계 학습(machine learning)을 통해서 인기도를 설명할 수 있다고 주장하였다 [23].

그러나 오디오 특성으로부터 음악의 인기도를 효과적으로 학습할 수 있다는 [11]의 주장은 거대한 규모의 검증을 통해 아닌 것으로 보여졌으며 음악적 특성을 통한 인기도 예측은 불가하며 아직은 science가 아닌 open field 라고 주장하였다[24]. 또한 “cumulative advantage”나 preferential attachment또한 실제 사용의 가능성에 반박하고 있다[28]. 최근 University of Bristol 이 연구한 Hit song equation은 Hit song의 여러 feature를 분석하고 이를 하나의 “식”으로써 표현하고 있다. 그러나 이 또한 흥미위주이며 초기의 의도와는 달리 상업적인 용도에서 멀어지고 있다. 현재로써 Hit

Song Science는 하나의 풀리지 않는 연구 과제로써 음악 자체의 feature만을 통해 Hit song을 예상하는 것은 힘들다는 연구가 다수 제기되고 있다 [28,29].

이에 따라 음악의 내부적 특성이 아닌 소셜 미디어 속 음악과 사용자 사이의 관계에서 음악의 흥행을 예측하려는 연구가 나타났다. last.fm 같은 소셜 기반 서비스 정보를 통해 음악의 히트 여부를 예측하려 하였으며 이 연구에서는 소셜 데이터만을 사용해 상당히 높은 정확도의 모델을 만들었다는 점에서 의의가 있다. 서비스의 API를 통해 수집한 정보를 통해 다양한 특성을 추출하고[19], 다양한 데이터를 바탕으로 예측모델을 제시하였다. 앨범이 얼마나 P2P네트워크에서 불법적으로 공유되어지는지[4], 앨범 발매를 전후로 Sentiment를 예측하고 비교하는[32] 연구들을 통해 앨범 판매량을 예측하는 것과 같이 독특한 데이터들이 제시되었다. 이러한 연구들은 음악적 특성을 통한 예측을 한계를 인지하고 소셜 미디어 데이터를 이용하여 흥행 예측의 정확도를 높였다는 점에서 의미가 있다.

제 2 절 기술적 배경

1. API (Application Programming Interface)

API 란 응용프로그램이 특정한 서비스에 접근할 수 있도록 따라야 하는 인터페이스이다. 웹 환경에서는 제공하는 웹 API는 웹 서비스와 같이 메세지를 통해 요청을 보내고 응답을 받는 개념으로 사용된다. 메세지의 내용은

JSON(JavaScript Object Notification)이나 XML(Extensible Markup Language)로 이루어져 있으며[21] 사용자가 사용하기 위해서는 파싱(Parsing)작업이 필요하다. 공개적으로 누구나 사용할 수 있도록 허용된 API 는 공개 API(Open API) 라 하며, 비교적 제한이 없이 웹 서비스에서 제공하는 데이터를 다른 서비스 등에서 사용할 수 있다[3].

2. 회귀 분석(Regression analysis)

회귀 분석은 변수들간의 관계를 측정하는 통계 과정이며, 종속 변수를 설명하기 위하여 독립 변수들이 어떠한 관계식을 가지는지 분석한다. 어떠한 변수들이 종속 변수를 변화시키는지 관찰할 수 있기 때문에 예측 분야에서 폭 넓게 사용된다.

종속 변수 Y, 독립 변수, X, 미지의 파라미터 β로 이루어져 있으며 다음과 같이 나타낼 수 있다.

$$Y \approx f(X, \beta)$$

독립 변수들의 관계에 따라서 여러 가지 방식의 회귀분석 방법이 존재하기에 데이터의 특성에 따라서 회귀 분석 방법을 선택하여야 한다[47].

2. 의사 결정 트리(decision tree)

의사 결정 트리는 분류 절차로서 반복적으로 데이터를 트리에 존재하는 각 가지들에서 일련의 검정을 바탕으로 세분하여 분류하는 알고리즘이다. 트리는 그림 1과 같이 루트 노드(Root node), 내부 노드(Internal node), 터미널 노드(Terminal node)로 구성되어 있다.

각각의 노드는 오직 하나의 부모 노드(Parent node)와 두개 이상의 자식을 가진다. 데이터 뭉치(Data Set)는 트리에 의해서 정의된 결정 프레임워크(Decision Framework)를 통해 순차적으로 분류된다.

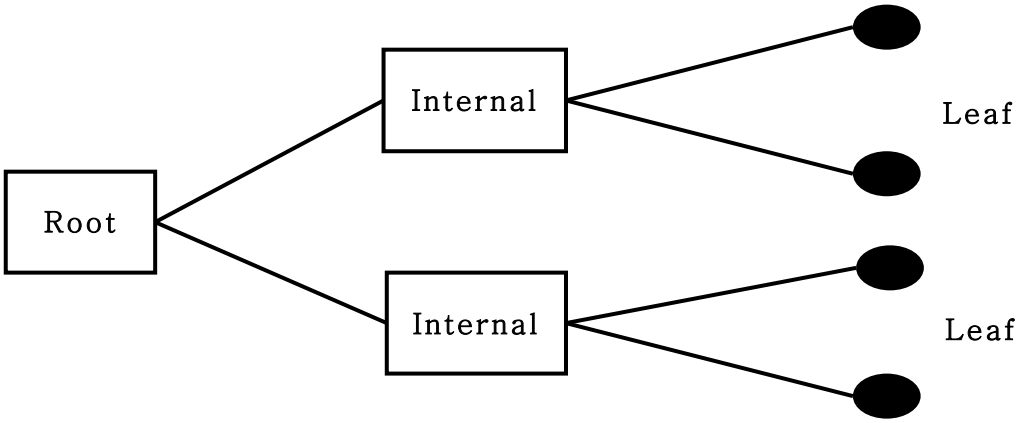

그림 2. 의사 결정 트리의 구성 요소

의사 결정 트리는 엄격하게 비모수적(Nonparametric)이며 인풋 데이터의 분포에 대한 어떠한 가정도 필요하지 않다는 장점을 가지고 있다. 또한 특성과 클래스간의 선형이지 않은 관계를 다루며 유실된 데이터를 허용하는 등의 유연한 특징을 가지고 있다[46].

제 3장 연구 문제

앞서 선행 연구 조사를 통해 사용자의 음악 청취 행태와 트렌드 사이의 관계를 분석하려 시도한 연구가 많지 않으며 특히 사용자 그룹을 분류하거나 외부 요소에 따른 차이점을 규명하는 연구는 거의 이루어지지않은 것을 확인할 수 있었다. 음악 흥행 예측의 경우 많은 수의 연구가 소셜 데이터를 중심으로 이루어졌지만 사용자의 음악 청취 행태를 활용한 연구는 찾기 어려웠다.

따라서 본 연구에서는 이러한 한계점을 극복하기위해 다음과 같은 연구문제를 설정하였다.

(연구 문제 1) 음악 청취 행태 관련 트윗을 생성하는 사용자 간의 차이점을 바탕으로 사용자 그룹을 나눌 수 있는가?

자료 수집 이후, 어떠한 사용자들이 관련 키워드를 쓰고 있는지 파악하기 위해 사용자를 분석하고 도출할 수 있는 정보가 무엇인지 분석한다. 사용자 별로 생성한 트윗의 양을 기준으로 이루어지며 사용자간에 어떠한 차이점이 있는지 관찰한다.

(연구 문제 2) 분류된 사용자의 음악 청취 행태는 음악 트렌드와 어떠한 경향성과 특징을 보이는가?

각 사용자 그룹별 음악 청취 행태와 트렌드와의 분포를 통해 어떠한 경향성을 가지는지 알아보고 특징을 정의한다.

(연구 문제 3) 장르와 같은 음악 특성에 따라서 그룹별 음악 청취 행태

와 음악 트렌드는 어떠한 관계성이 존재하며 이를 통해서 새로운 feature의 제시가 가능한가?

음악의 장르와 아티스트 선호도에 따라서 음악 청취 행태가 어떻게 달라지는지 관찰한다. 나타난 특징을 반영하여 새로운 feature를 제시한다.

(연구 문제 4) 음악 순위 예측에서 추출된 특성들은 얼마나 유용한가?

추출한 특성과 음악 트렌드 사이를 스피어만 상관 계수로 정량화한다. 또한 음악 순위 예측에서 특성들이 얼마만큼 유용하게 사용되는지를 평가한다.

제 4 장 연구 방법

제 1 절 데이터 수집

본 연구를 수행하기 위해서는 트위터 및 음악 관련 데이터의 수집이 필수적이다. 실험을 위해 수집한 기간은 2013년 11월 21일부터 2014년 5월 28일까지이며 이후에 기간을 늘려 추가 연구를 수행한다.

1. 트위터

2장에서 언급했듯이 트위터에서는 공개적으로 다양한 API를 제공한다. 이전의 연구들에서는 Twitter Streaming API에서 'Sample'상태를 이용해 실시간으로 공개 상태로 올라오는 트위터의 1% 샘플을 수집하였다. 이후, #nowplaying 관련 키워드만을 다시 필터링 하였다. 이러한 일련의 과정으로 시간별 데이터의 수는 실제 트윗의 수보다 현저히 작아졌을 수 밖에 없다. 따라서 본 연구에서는 실시간으로 #nowplaying 관련 트윗을 최대한으로 수집하기 위해 Twitter Streaming API의 'Filter'를 이용하였다[35]. 이를 통해 키워드인 #nowplaying 이외의 트윗은 수집하지 않고 오직 키워드가 포함되는 트윗만을 수집할 수 있었다. 모든 트위터 데이터를 가지고 있는 외국 회사 'Topsy'에서 제공하는 그래프를 통해 키워드와 날짜별 전체 트위터 수를 알 수 있다. Topsy 에서 제공하는 데이터와 비교하였을 때 본 연구에서 수집한 트윗은 키워드에 포함되는 트윗 대부분을 수집한 것을 알 수 있었다.

수집 기간 동안 총 79,214,135개를 수집하였으며, 하루에 평균적으로 약 450,000개의 트윗이 관찰되었다. 수집한 데이터의 항목은 표 1과 같다.

범위	항목	내용
User Profile	Id	사용자의 고유 식별 아이디
	Name	사용자의 아이디
	Screen Name	트위터에서 보여지는 사용자의 아이디
	Location	사용자가 설정한 기본 위치
	URL	사용자가 설정한 URL
	Followers	사용자의 Follower 수
	Friends	사용자의 Following 수
	Created At	사용자가 ID를 생성한 시기
	Favorite	사용자가 등록한 Favorite
	Utc Offset	사용자의 시차
	Statuses	사용자의 총 트윗 수
	Language	사용자의 언어
	Time Zone	사용자의 Time Zone
	Description	사용자가 설정한 자신의 설명
	ID	Tweet의 고유 식별 아이디
	Created Time	Tweet이 생성된 시간

Tweet	Text	Tweet의 내용
	Source	Tweet을 포스트한 디바이스
	Geo	Tweet의 위치정보
	Retweeted	Tweet이 Retweet되었는지 여부
	Favorited	Tweet 이 Favorite되었는지 여부
	Language	트윗의 언어

표 1. 트위터에서 수집한 트윗이 포함하고 있는 정보

2. 빌보드 차트

음악 산업에서 공신력있는 Billboard 매거진은 다양한 음악차트를 출판한다 [22]. Hot 100은 현재 가장 인기있는 전 세계 싱글 음악에 대한 순위이며 음악 판매량, 스트리밍, 라디오 에서 수집한 데이터를 바탕으로 집계된다.

Billboard의 개발자 API는 더이상 제공되지 않기때문에 웹 화면에서 직접적으로 수집하는 방식을 사용하였다. Billboard 차트는 매주 목요일 일주일마다 업데이트되며 미래의 빌보드 차트를 예측하기 위해 총 35주의 데이터를 수집하였다. 한 차트당 1위부터 100위까지 순위를 포함하고 있기 때문에 수집한 데이터는 일주일당 100개에 해당한다. 빌보드 에서 수집한 데이터의 항목은 표 2와 같다.

항목	설명
Title	음악의 제목

Artist	음악의 가수
Ranking	음악의 순위
Weeks on Chart	음악이 차트에 머무른 주
Top Position	음악이 차트에서 가장 높게 올라간 순위

표 2. 빌보드 차트에서 수집한 데이터의 항목

전체 빌보드 데이터에서 고유한 음악은 총 356개, 아티스트는 227개에 달한다.

3. 음악의 장르

음악의 장르는 때때로 음악에 따라 구분이 애매모호하기 때문에 Amazon Music의 장르 대분류를 바탕으로 수집하였다. 대분류에 속하지만 데이터의 수집 기간동안 아주 낮은 빈도로 나타난 장르들은 모아서 others로 추가하였다. 이렇게 35주간의 빌보드 차트에 나타난 장르는 Pop, Hiphop, Dance&Electrnoics, Country, R&B, Rock, Soundtracks, Holiday, Others 총 9가지이다.

제 2 절 전처리 과정

Billboard 차트와 트위터 내 음악 청취 행태와 연관성을 분석하기 위하여 트위터에서 빌보드 차트에 등재된 음악을 분류할 필요가 있다. #nowplaying 태그가 된 트윗은 일반적으로 음악 제목, 가수, 가사 등과 같

은 음악에 관련된 정보를 가지고 있다. 이를 바탕으로 수집한 24주간의 빌보드 차트에서 언급된 전체 음악을 이용하여 특정 음악과 관련된 트윗을 분류하였다. 빌보드 차트와 관련된 음악의 수집은 키워드 검색을 이용하여 수행되었다. 첫 번째로, 해당 음악을 청취한 트윗을 찾을 때는 음악의 제목과 가수의 이름을 함께 검색하여 나타나는 트윗을 선택하였다. #nowplaying에서는 2장에서 언급하였던 것과 같이 〈#nowplaying + Title〉과 같이 아티스트를 함께 사용하지 않는 경우도 빈번하나 이는 다른 노래이나 중복된 제목을 가진 음악일 경우에 이를 연구자가 어떤 음악에 해당하는지 판단하기 어려우므로 확실하게 어떤 음악인지를 알 수 있도록 한 것이다. 수집된 데이터는 보다 적어도, 정확성을 높일 수 있다는 점에서 두 키워드를 함께 사용하였다. 이 트윗의 양은 'Song popularity'라 이름붙였고 사람들이 해당 음악에 얼마나 많은 청취 행태 트윗을 작성하였는지 알 수 있는 척도이다.

두 번째로, 해당 가수의 음악을 청취한 트윗을 찾을 때는 가수의 이름을 키워드로 검색한다. 이는 'Artist popularity'로서 특정 가수의 다른 음악이 얼마나 인기 있는가를 알 수 있음과 동시에 인지도의 척도로서 삼을 수 있다. 이렇게 추출된 빌보드 관련 데이터에는 시간 정보가 존재하므로, 후에 이를 사용하여 분석할 때 daily값을 추출하여 적용한다.

세 번째로, Retweet 관련 트윗을 'RT'라는 키워드가 트윗의 맨 앞에 존재한다면 이는 리트윗된 트윗으로 간주하고 따로 수집하였다. 이렇게 수집된 Retweet 데이터 내부에서 다시 Song popularity와 Artist popularity를 수집한다.

제 4 장 분석 및 실험

제 1 절 음악 청취 행태와 사용자 분석

1. 사용자 분석

연구문제 1 의 해결을 위하여 전체 사용자에 대해서 사용자별로 생성한 트윗의 양을 히스토그램으로 나타낸다. 이상치를 가지는 사용자들의 트윗과 메타데이터를 분석하여 어떠한 사용자인지를 정의하고, 기준에 의하여 사용자 그룹을 분류한다.

전체 #nowplaying 관련 키워드를 어떠한 사용자들이 얼마만큼 사용하는지 알아보기 위해 사용자별 트윗 수를 집계하였다. 순 사용자의 수는 약 4,297,968이며 총 트윗 개수가 약 79,214,135여개에 이르기 때문에 평균적으로 약 18개의 트윗을 생성했다는 것을 알 수 있다.

그러나 트윗수 통계를 관찰하면 가장 많이 트윗을 생성한 사용자는 총 100,000개 이상을 생성하는 것을 볼 수있다. Median 값이 1이라는 것으로 볼 때 반 이상의 사용자는 #nowplaying을 한번만 사용하는 것으로 알 수 있다. 이 데이터를 바탕으로 생성한 분포는 다음과 같다.

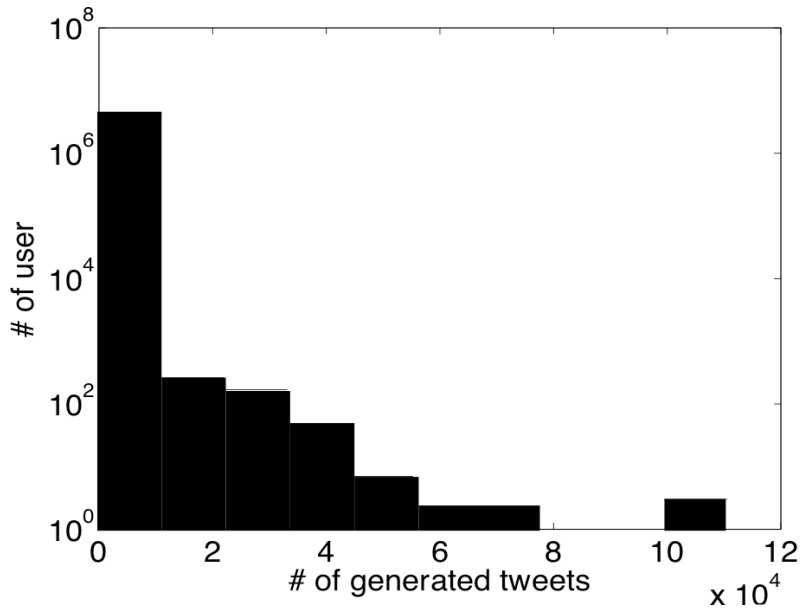

그림 3. #nowplaying 키워드 사용자들의 트윗 수별 분포

그림 5에서 소수의 사용자들이 상당수의 트윗을 생성한다는 것을 알 수 있었다. 가장 트윗을 많이 생성한 10개의 사용자를 보면, 이름에서 부터 알 수 있듯이 라디오 웹에서 관리하는 트위터 계정임을 알 수 있다. 이름에서 추론하지 못하더라도, 사용자 설명에서 라디오임을 알 수 있는 내용을 발견할 수 있었다. 이에 따라서 'Radio', 'FM', 'On Air'등의 키워드로 라디오 사용자를 검색할 수 있었으며 이들은 약 19,021의 개별적인 사용자가 9,031,013의 트윗을 생성한다는 것을 알 수 있었다. 이는 전체 트윗의 20%에 해당한다.

Type	User name	User screen	Tweet 수

		name	
Radio	Securenet Systems	securenet	74511
	Virtual Jukebox	VirtualJukebox	73118
	OutboundMusic	OutboundMusic	70660
	Pulse API	PulseAPI	61346
	EKR Radio	ekrnetwork	53144
	Now Playing on The X	NowOnTheX	52041
	dopefm	dopefm1	48474
Bot	Now Playing	NowPlayingBot_	43715
Radio	iCountry.FM Radio	iCountryNash	39757

표 3. #nowplaying관련 트윗을 가장 많이 한 10명의 사용자

이들은 자신들의 라디오를 홍보하기 위해 트위터를 사용한다. 때문에 라디오에서 재생하고 있는 음악을 트위터 계정에 올리기 때문에 많은 수의 트윗을 올린다고 할 수 있다. 생성된 데이터는 사용자들의 음악 청취 행태라기보다는 인터넷 라디오 운영국의 재생 목록이라고 할 수 있다. 따라서 이러한 목적을 가진 사용자들은 일반적인 사용자들과 집단이 다르다고 할 수 있어 본 연구에서는 라디오 사용자를 Broadcaster로 이름 붙이고 일반 사용자는 Crowds로 분류 하였다.

2. 사용자 그룹별 음악 트렌드와 클러스터

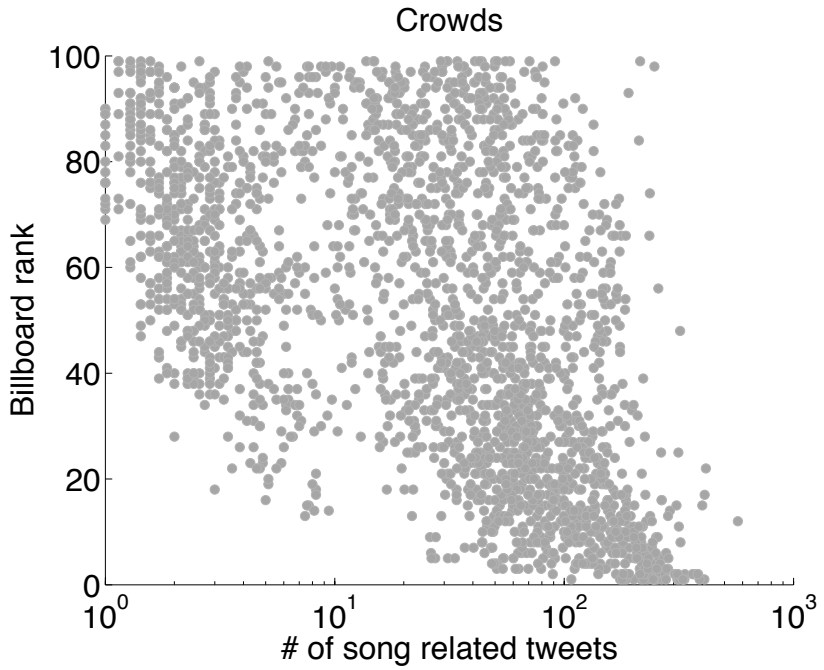

그림 4. Crowds 사용자 그룹의 song popularity에 대한 빌보드 순위

Crowds 사용자 그룹의 song popularity와 빌보드 순위의 분포를 살펴보면, 1-20 위의 순위까지는 분포가 순위에 따라 상대적으로 군집되어있는 것을 알 수 있다. 하지만 이후 20위 이상의 순위에서는 같은 순위라 할지라도 다양한 분포를 가지고 있다는 것이 관찰된다. 이는 사용자가 1-20위권에서는 비슷한 선호도를 가지지만, 그 이상의 순위에 대해서는 확연하게 선호도 차이가 다름을 내포하고 있다.

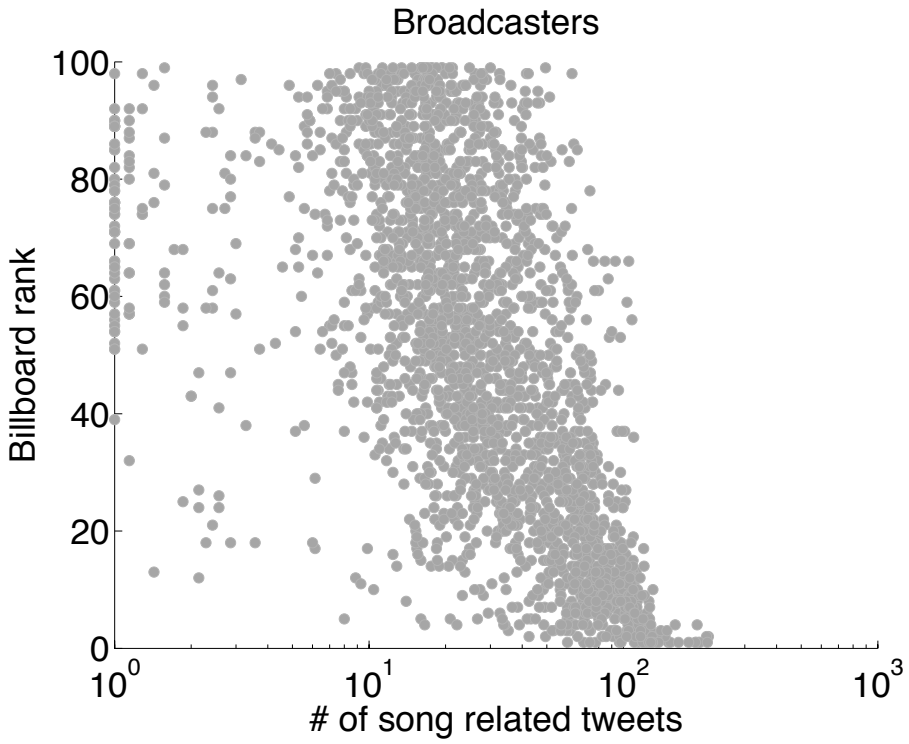

그림 5. Broadcaster 사용자 그룹의 song popularity에 대한 빌보드 순위

Broadcasters 사용자 그룹에서는 좀 더 다른 양상을 보인다. 이 사용자들은 전체적으로 군집화되어잇는 양상을 보이며, Crowds와 비교하여 빌보드 순위와 보다 큰 연관이 있어 보인다. 하지만 주요 군집을 제외하고 10번 이하의 트윗이 생성된 경우들이 관찰되어서 특정 음악들은 아예 broadcaster들에게 선택받지 못함을 알 수 있다.

연구문제 2의 해결을 위하여 각 사용자 그룹별 음악 관련 특성을 추출하고 빌보드 순위와의 분포를 관찰한다. 특이점들이 존재하는 사용자 그룹에 대해서 관계성에 따라 3가지 cluster로 나누고 분석한다.

두 사용자 그룹의 양상은 확연히 다른 차이점을 가지고 있다. 때문에 두 사용자 그룹을 보다 자세히 분석하면 각 사용자들이 어떠한 점에서 차이가 있는지를 알 수 있다. Crowds에서는 빌보드 순위와 song popularity의 분포가 상대적으로 더 분산되어있는데 이것을 바탕으로 세 가지 cluster를 나누었다.

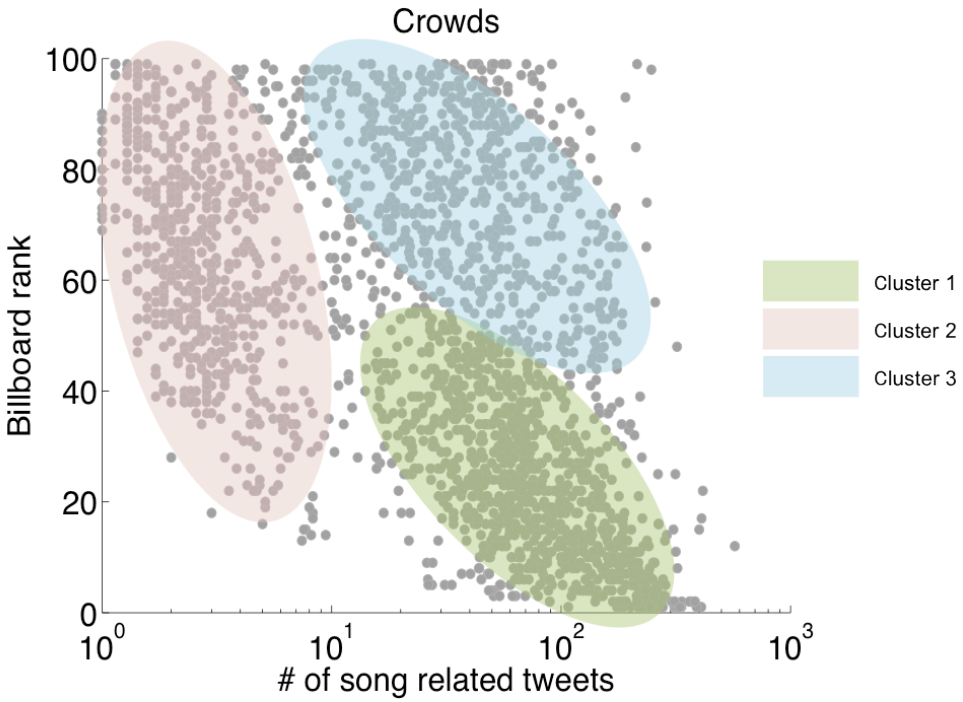

그림 5. Crowds 사용자 그룹의 song popularity와 음악 순위 분포를 3가지 cluster로 분류한 것

첫 번째 cluster는 song popularity와 빌보드 순위가 상대적으로 선형 관계를 이루고 있으며 순위에 따라 군집화된 모습을 보여준다. 이 cluster는 빌

보드 음악중에서도 1-20위 사이에 있는 아주 흥행한 음악에 속한다고 할 수 있다. 이를 통해서 Crowds는 1위에 가까운 순위일 수록 음악 청취의 수가 비례한다고 가정할 수 있다.

두 번째 cluster는 빌보드 순위에 비해 상대적으로 song popularity가 낮은 것들을 의미한다. 이들은 내부적으로는 선형 관계가 있는 것 처럼 보이지만, 같은 순위의 다른 관측값과 비교하여 보았을 때 song popularity가 낮아서 전체적으로 보았을 때는 다르다는 것을 알 수 있다. 이 cluster에 해당하는 관측값은 트위터의 Crowds사용자들에게 별다른 선호도가 없음을 가정할 수 있다.

세 번째 cluster는 빌보드 순위에 비해 상대적으로 song popularity가 높은 것들을 의미한다. 이들은 cluster 2에 비해서 내부 선형 관계가 약하게 보이지만 cluster 1과 비슷하거나 더 높은 song popularity를 가지고 있다. 이 부분에 해당하는 관측값은 빌보드의 트렌드와 상관없이 Crowds 사용자 그룹에서 큰 선호도를 보이는 것으로 해석될 수 있다.

본 연구에서는 Crowds 사용자의 성격이 다른 3가지 Cluster를 장르와 Artist popularity를 기준으로 분석하며 진행된다. 또한 이 분석을 기준으로 빌보드 순위 예측을 위한 새로운 feature를 제시하여 이것이 어떠한 영향을 끼치는가에 대해서도 평가한다.

3. 장르와 아티스트 선호도에 따른 사용자 그룹 분석

연구문제 3의 해결을 위하여 장르, 아티스트 선호도를 바탕으로 각 사용자 그룹의 음악 청취 관련 특성과 트렌드가 어떠한 차이점을 가지는지 살펴보

고 각 Cluster와 비교해본다. 추가적으로 음악 발매 이후 지난 날짜와 음악 청취 관련 특성의 관계를 살펴보고 그룹별로 어떠한 차이점이 있는지 분석한다. 또한 관찰된 특징들을 통해 음악 트렌드를 설명하는 새로운 feature를 제안한다.

사용자 그룹별 순위에 대한 song popularity의 차이점을 보다 자세하게 알아보기위해 장르와 Artist popularity를 통해 분석한다. 더불어 앞서 언급한 Crowds의 cluster에 대한 설명 뿐만 아니라 Broadcaster는 Crowds와 어떠한 차이가 있는지에 대해서 알아본다. 각 장르별 분포의 차이를 통해 사용자 그룹별로 어떠한 차이점과 선호도를 가지고 있는지 분석한 바는 다음과 같다.

Cluster 1은 Song popularity와 빌보드 순위가 동시에 높은 경우인데, 이러한 분포는 Pop 장르에서 찾아볼 수 있다. 그림 6 에서 Cluster 1에 해당하는 관찰값들중 많은 수가 Pop음악임을 알 수 있다. 이는 양쪽 사용자그룹에서 모두 해당되며 대부분의 Pop음악은 순위와 Song popularity가 높다. 그림 7에서 Broadcaster 그룹의 분포는 1-20위에 아주 집중되어있는 것을 볼 수 있다. Crowds 그룹보다는 Broadcaster에서 보다 빌보드 순위와 선형관계를 이루며 분포하고 있는 모습을 볼 수 있다.

전체적인 양상을 살펴보았을 때 pop에 해당하는 관찰값이 808개로 타 장르보다 수가 많고, 전체의 약 1/3을 차지하고 평균 순위가 38.55위로 전체 평균인 50위보다 1위에 가깝다. 이를 통해 빌보드와 트위터 모두에서 pop 은 인기있는 장르임을 알 수 있어 pop이 전체적으로 친숙하고 그 수도 상대적으로 많음을 유추할 수 있다.

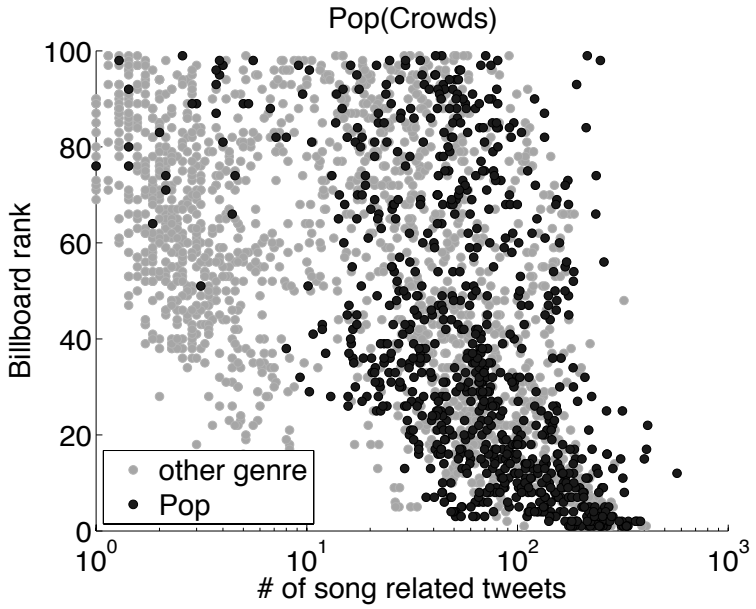

그림 6. Crowds 사용자 그룹의 Pop장르에 대한 분포

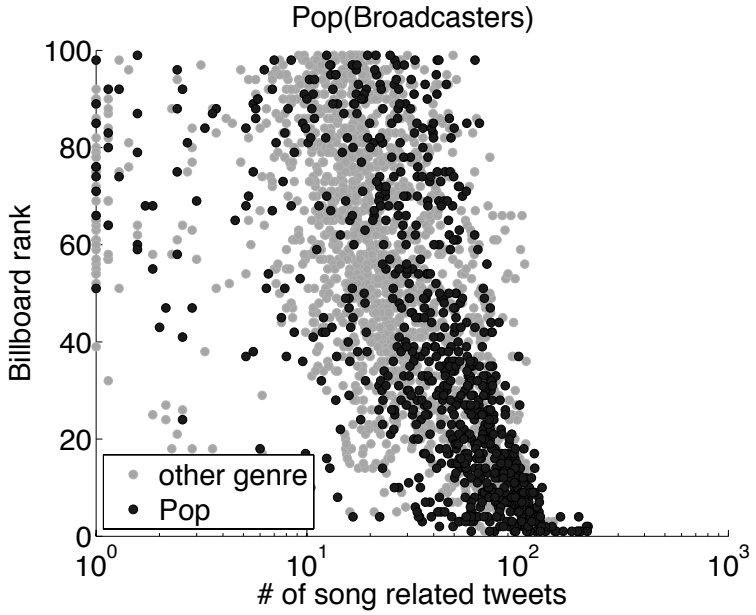

그림 7. Broadcasters 사용자 그룹의 Pop장르에 대한 분포

이와 다르게 Cluster 2는 빌보드 순위와 Song popularity 모두 낮은 경향을 보인다. 그림 8과 그림 9를 통해 country장르가 이외의 장르와 song popularity에서 어떠한 차이점을 보이는지 알 수 있으며 cluster 2의 관찰값들의 특징을 유추할 수 있다. Crowds 그룹에서 Country 장르 음악은 빌보드 순위에 비해 아주 적은 양의 트윗이 생성되었다. Broadcasters 그룹 역시 상대적으로 적은 양이지만, Crowds에 비해서는 보다 군집화되어있는 모습을 보인다. 이를 통해 Crowds 사용자 그룹의 Cluster 2는 Country 장르가 많은 부분을 차지하고 있으며 트위터 사용자들이 Country 장르에 대한 선호도가 낮음을 간접적으로 알 수 있다.

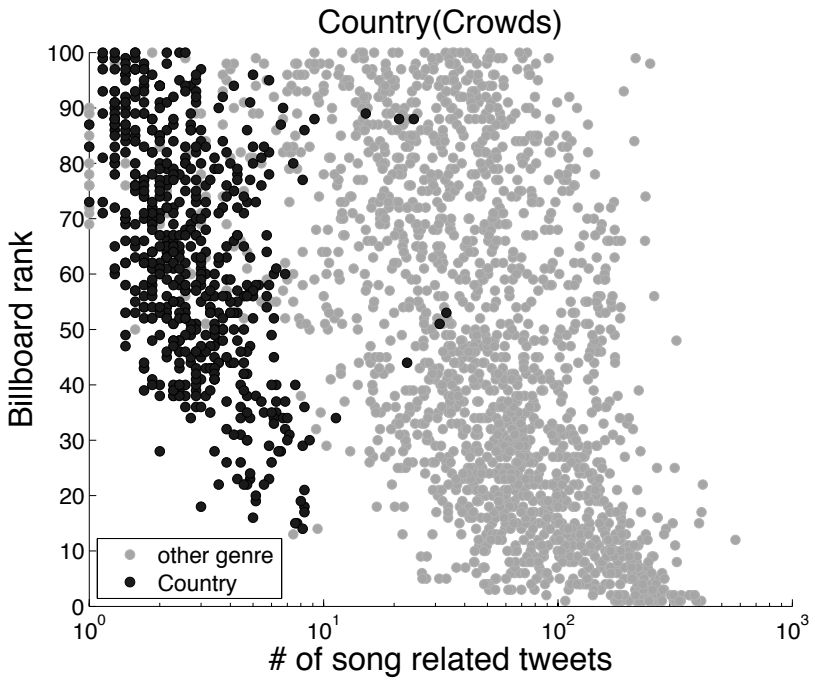

그림 8. Crowds 그룹의 사용자 그룹의 Country장르에 대한 분포

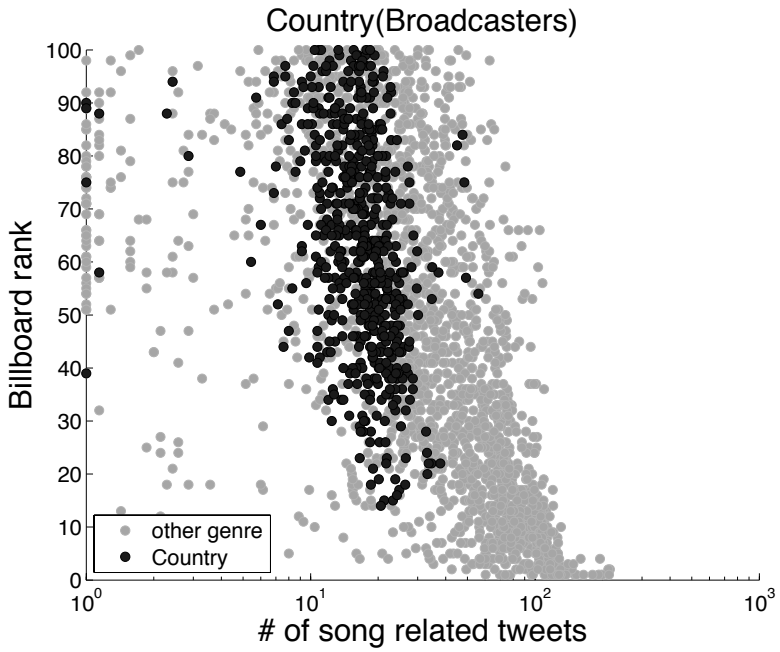

그림 9. Broadcasters 사용자 그룹의 Country장르에 대한 분포

반면에 Broadcasters의 경우 군집에서는 낮은 양을 가지고 있지만 이것이 군집의 범주를 벗어나는 양이 아니어서 상대적으로 선호도가 Crowds보다 높거나 빌보드 순위에 보다 예민하다고 추측할 수 있다. 이러한 Broadcaster도 상대적으로 song popularity가 적은 장르가 존재하는데, 그것은 일반적인 장르가 아닌 영화나 애니메이션에서 제공되는 soundtrack에 해당하는 장르이다.

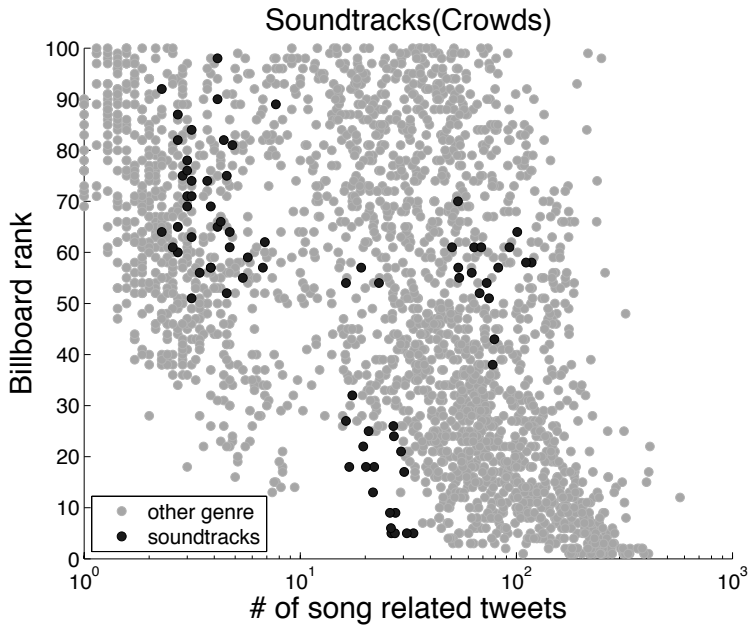

그림 10. Crowds 사용자 그룹의 Soundtracks 장르에 대한 분포

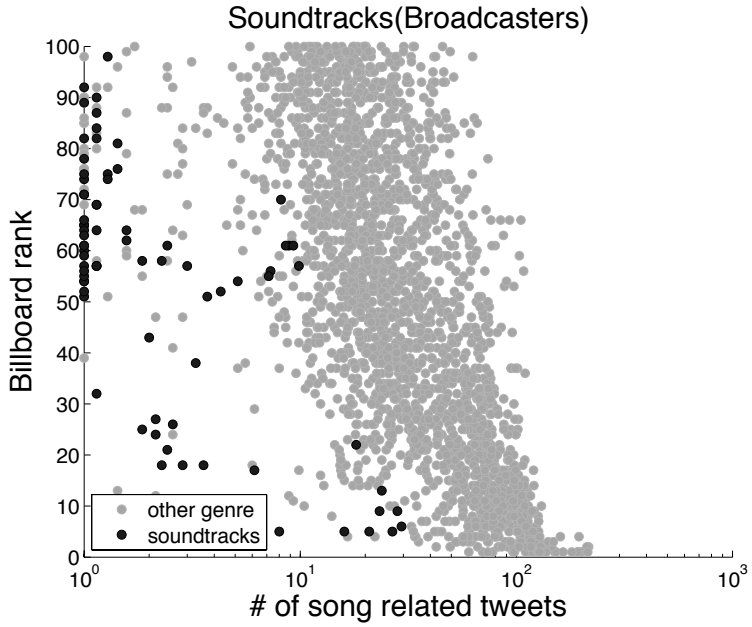

그림 11. Broadcasters 그룹의 Soundtracks 장르에 대한 분포

soundtrack에 해당하는 음악들은 겨울왕국('Frozen')의 'Let it go'와 같은 삽입곡이 대부분이다. 이 음악들은 Crowds 사용자 그룹에서는 별다른 경향성이 보이지 않았지만, Broadcaster 그룹에서는 많은 수가 한번도 청취 트윗이 생기지 않았다는 것을 알 수 있다. 이것은 Broadcaster의 선호도가 Crowds와는 다르며 이 차이점은 목적에서 기인한다고 추측할 수 있다. Broadcaster들은 soundtrack을 선호하는 어린 연령대보다는 보다 성숙한 연령대를 대상으로 운영될 수 있기에 이러한 차이점이 생긴다는 가정을 할 수 있다. 이 부분은 관련 연구나 설문을 통해 규명될 수 있을 것이다.

Cluster 3을 설명하는 특성은 Artist popularity를 통해 찾아볼 수 있다. 데이터를 관찰한 결과 Artist에 대한 음악의 선호도가 높은 경우 전체적으로 song popularity가 높음을 알 수 있었다. 이는 Pop 장르와 유사하여 Pop 장르의 Artist들은 대체적으로 popularity가 높다고 알 수 있지만 Cluster 3의 보다 많은 부분을 설명할 수 있다.

그림 12와 그림 13에서 두 사용자 그룹 모두 Artist popularity가 높을 때 Song popularity 도 같은 순위에 비해 보다 높은 것을 알 수 있다. Broadcaster는 순위에 비해 과도하게 벗어난 song popularity가 비교적 적게 보인다. 하지만 Crowds 사용자는 100에 가까운 순위에서 확연하게 그 차이점을 관찰할 수 있다. 이것은 일반적인 사용자들이 유명한 Artist인 경우에 그/그녀의 노래를 좀 더 선호한다는 것으로 해석할 수 있다.

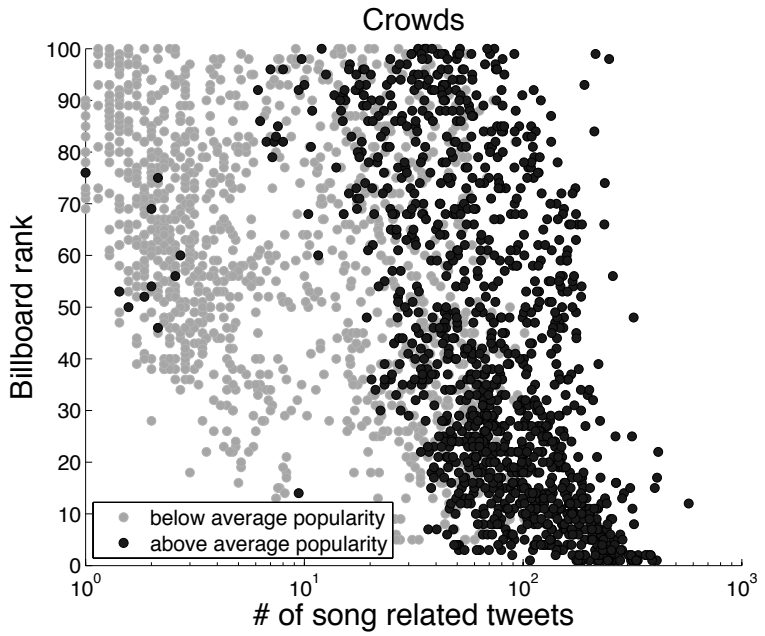

그림 12. Crowds 사용자 그룹의 Artist popularity에 따른 분포

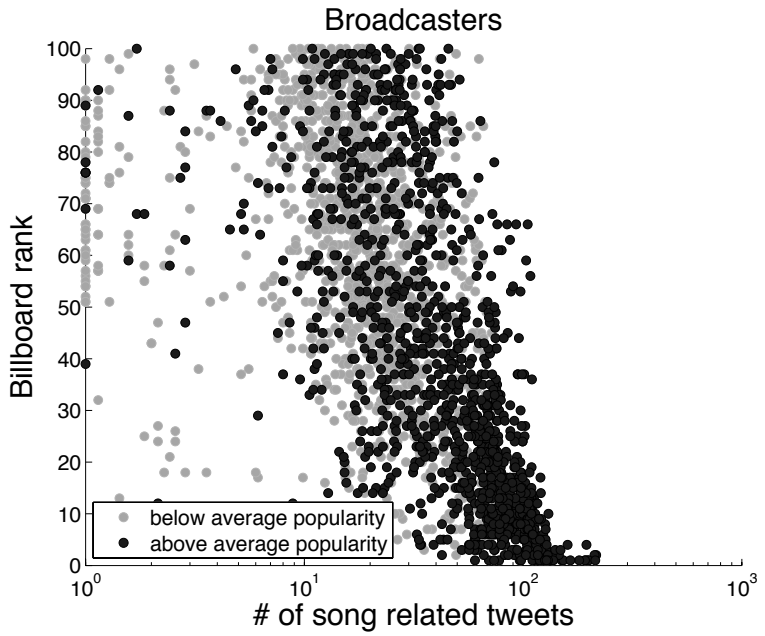

그림 13. Broadcasters 그룹의 Artist popularity에 따른 분포

낮은 순위에 높은 Song popularity를 가진 몇 가지 관찰 값을 통해서 알 수 있었던 사실은 해당 음악들이 유명한 Artist의 앨범 중 Title이 아닌 수록곡 이었다는 것이다. 앨범의 Title은 높은 순위에 안착하지만 그 앨범의 수록곡은 상대적으로 낮은 순위를 가진다. 하지만 높은 순위의 음악과 비슷하게 높은 Song popularity를 가지는 모습을 관찰할 수 있었다. 이와 같은 요인으로 빌보드 순위에 비해 상대적으로 높은 값이 나오게 되는 것으로 추론 가능하다.

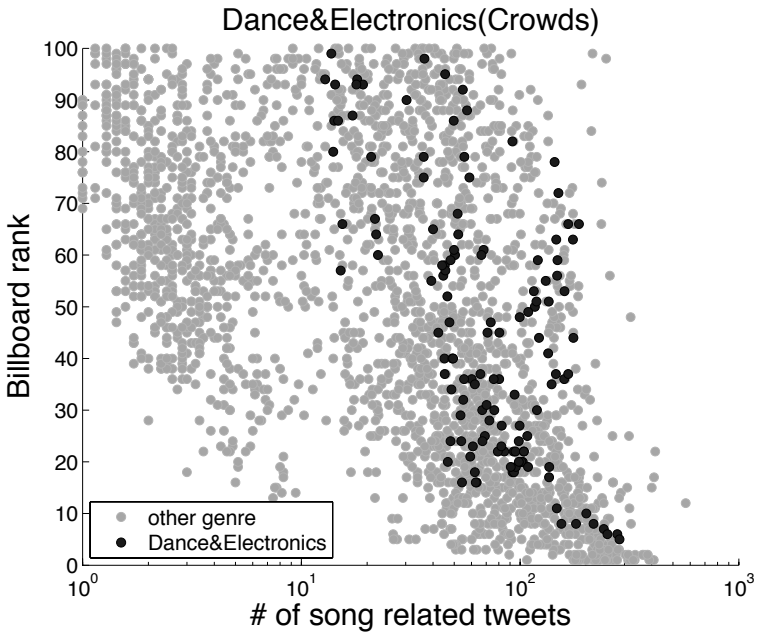

그림 14. Crowds 사용자 그룹의 Dance & Electronics 장르에 대한 분포

Dance & Electronics 장르에서도 비슷한 행태를 관찰할 수 있다. 전체적으로 다른 관찰값들에 비해 높은 song popularity를 가지고 있고, 그 특징이 Broadcaster 사용자 그룹에서 눈에 띄게 나타난다. 즉 Broadcaster 사용자 그룹은 Dance & Electronics 장르의 음악을 집중적으로 재생하는 경향이 있으며 이는 Online Radio Station 특성상 신나는 음악을 중심으로 재생하는 등의 가정이 가능하다.

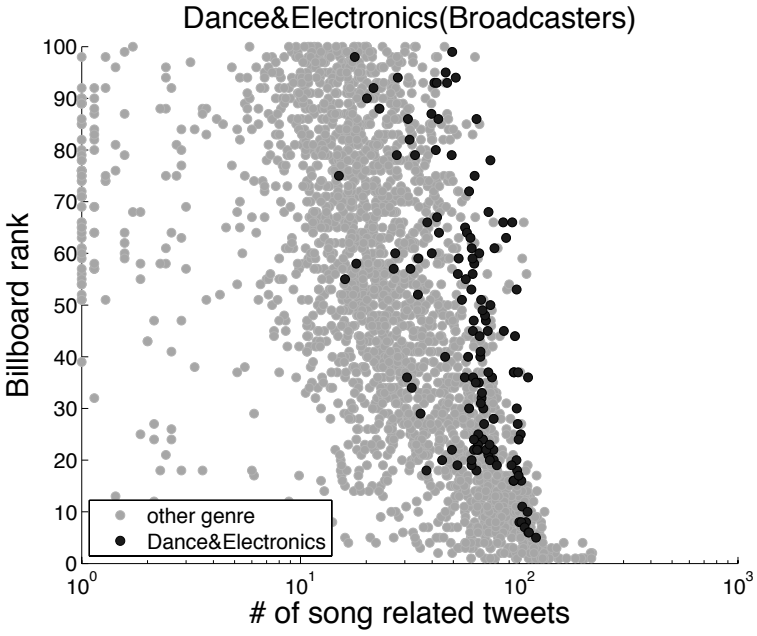

그림 15. Broadcasters 사용자 그룹의 Dance & Electronics 장르에 대한 분포

언급한 pop, country, soundtracks, dance & electronics 이외의 다른 장르에 대한 song popularity, 순위의 평균과 관찰값의 수는 표와 같다.

Genre	Mean Song popularity		Mean rank	# of entry
	Crowds	Broadcasters		
Overall	56.63	36.49	50.49	2399
Pop	90.27	51.3	38.55	808
Hiphop	57.33	31.47	56.17	409
Dance	85.12	63.36	45.64	131
Country	2.25	16.54	63.48	577
Rock	63.37	42.42	47.48	254
R&B	45.9	27.4	61.72	86
Holiday	33.54	23.59	44.22	9
soundtrack	24.7	3.65	52.88	76
other	97.63	35.72	52.98	49
TheVoice	1.93	0.26	77.23	13

표 5. 장르별 평균 song popularity, 순위, 엔트리의 수

전체적으로 Broadcaster들의 평균 song popualarity가 Crowds보다 적다. 분석 결과와 같이 Dance & Electornics 와 Country의 경우는 Broadcaster에서 상대적으로 두각을 드러냈고, Soundtrack은 Crowds에서 우세했다. Others의 경우는 대분류 장르 이외의 다른 latin, folk와 같은 장르를 합친 것으로 사용자 그룹간의 수의 차이가 크지만 의미를 찾기는 힘들었다. The Voice의 경우는 장르가 아닌 미국 NBC에서 발영하고 있는 'The Voice'라는 아마추어 노래 경연 프로그램 참가자의 노래를 뜻한다. 이들은 이미 존

재하는 노래를 리메이크해서 매 경연마다 부르는데, 이 노래가 빌보드에 때때로 등재되었다. 하지만 해당 노래는 모두 장르에 상관없이 대부분 저조한 song popularity와 순위를 보인다. 이는 유명하지 않은 Artist이고 마케팅 등의 외부요소가 적기때문에 그렇다고 유추할 수 있다.

평균 순위에서는 pop장르가 가장 높은 30위권 대를 보였고 Country, R&B장르는 평균 60위권, The voice에서 방영된 노래는 77위의 평균값을 가지고 있다.

4. 발매 이후 시간과 특성간의 관계 분석

음악 발매 이후 시간이 지남에 따라 song popularity에는 어떠한 변화가 있는지 그림 16을 통해 분포를 살펴보면 큰 경향성은 보이지 않는다.

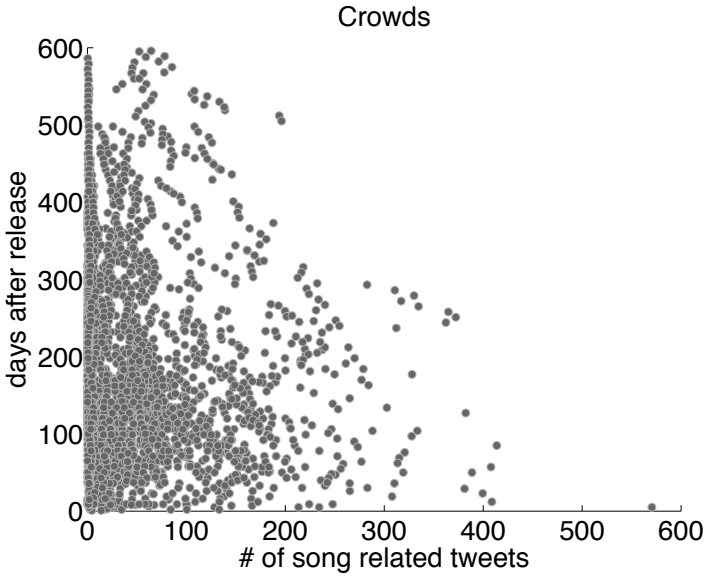

그림 16. Crowds 그룹의 Song populairty 대한 발매 이후 지난 시간의 분포

스피어만 상관 계수도 -0.1185로 약한 상관 관계를 보이지만, 발매 이후 기간이 지날 수록 song popularity가 감소한다는 경향성을 확인할 수 있다. 이에 반해 그림 17을 통해 Broadcasters 사용자 그룹은 그 분포가 다름을 알 수 있다. 역시 큰 경향성은 아니지만 발매 이후 지난 시간과 song popularity가 양의 상관 관계를 가지고 있음을 볼 수 있다. 스피어만 상관 계수는 0.1977로, Crowds와 비교하였을 때 반대의 관계성을 가지고 있음을 알 수 있다.

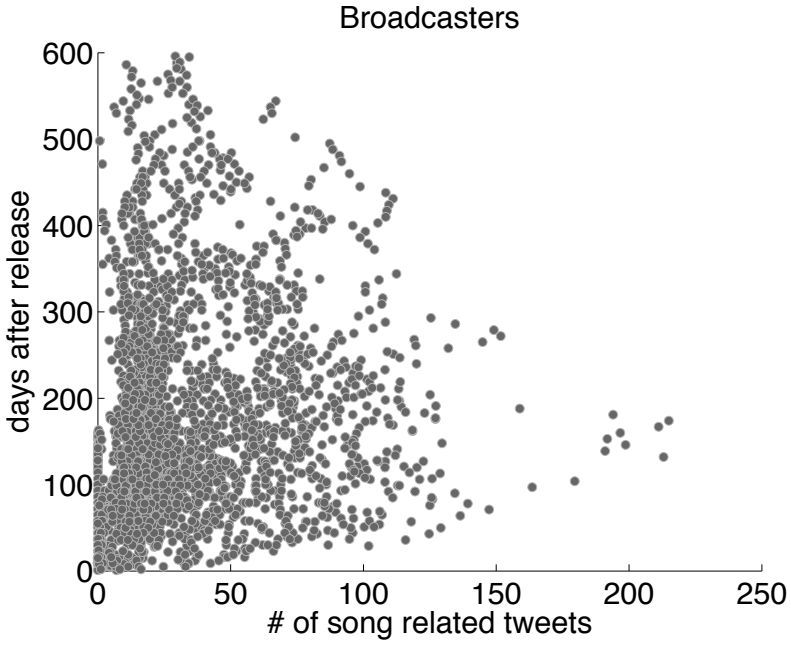

그림 17. Broadcasters 사용자 그룹의 song popularity에 대한 발매 이후 지난 시간의 분포

5. 음악 차트와 특성간의 상관 관계

트위터 사용자의 음악 청취 행태와 음악 차트와의 연관성을 평가하기 위하여 수집한 데이터베이스에서 5.2절을 바탕으로 특성을 추출한다. 5.2절에서 정의한 라디오 사용자와 일반 사용자들의 빌보드 음악에 대한 일별 재생수를 얻었다. 각 사용자 별 song popularity, artist popularity의 특성에서 추가적으로 follower와 Retweet의 수 추출하였다. 이들 특성은 '확산성' 이라는 관점에서 빌보드 순위에 영향을 끼치는지 알아보기 위함이다. 마지막으로 음악의 특성으로 앞서 이야기한 8가지의 장르와 발표 후 시간이 얼마나 흘렀는지를 day기준으로 추가하였다.

다음으로 분류된 사용자 그룹에서 추출한 각각의 특성의 빌보드가 나오기 이전 일주일 데이터와 미래 빌보드 음악 순위와의 연관성을 스피어만 상관계수(spearman correlation coefficient)로 측정하여 나타낸다. 스피어만 상관계수는 어떠한 변수들 사이의 상관 관계를 나타내는 지표로서 사용되며 -1과 1사이의 값으로 나타난다. -1에 가까울 수록 음의 상관관계이며 0이면 상관 관계 없음, 1에 가까울 수록 양의 상관 관계이다. 추출한 특성과 이에 따른 상관 계수는 다음과 같다.

Type	Feature	Group	
		Crowds	Broadcaster
Twitter	Max # of song popularity	-0.5563***	-0.6416***
	Max # of artist popularity	-0.3433***	-0.3649***

	Mean # of song popularity	−0.5667***	−0.6508***
	Mean # of artist popularity	−0.3512***	−0.3682***
	Min # of song popularity	−0.5599***	−0.6429***
	Min # of artist popularity	−0.3512***	−0.3703***
Retweet	Mean RT # of song popularity	−0.4826***	−0.3132***
	Max RT # of song popularity	−0.4525***	−0.2979***
	Min RT # of song popularity	−0.4948***	−0.3155***
User	Mean follower # of song mentioned user	−0.2651	−0.1162
	Mean follower # of artist mentioned user	−0.2091	−0.1135
Music	day after release	−0.0969***	
	pop	−0.2947***	
	hiphop	0.0894***	
	dance	−0.0403	
	Country	0.2534***	
	Rock	−0.0358*	
	R&B	0.0751***	
	Soundtrack	0.015	

표 6. 추출한 특성과 빌보드 순위와의 스피어만 상관 계수

두드러지는 특성은 일반 사용자(Crowds)의 song popularity와 Broadcaster의 song popularity가 차트와의 음의 상관 계수가 높다는 것이다. 이는 재생수가 많으면 숫자가 낮은 순위, 즉 1위에 가깝다는 의미를 가진다. 그중에서도 broadcaster는 crowds보다 상관 계수가 높아, 차트와의 강한 관계성이 있다는 것을 알 수 있다. 아티스트별 음악 청취 행태는 빌보드의 순위와 강한 관계성을 보이지는 않지만 재생수와 마찬가지로 음의 상관관계를 가지고 있어, 아티스트의 음악을 많이 듣는 것으로 나오면 수가 낮은 순위를 가지고 있다는 것을 시사한다.

Retweet의 경우, Crowds사이에서 음악에 대한 Retweet의 수가 많은 것은 차트의 순위가 1에 가깝다는 것을 -0.4826의 상관계수로 알 수 있었고, Broadcaster는 Crowds보다는 낮은 상관관계를 가지고 있지만 상관 간계가 존재한다는 것은 알 수 있다.

음악적 특성의 경우, 발매 이후 많은 시간이 절대적으로 차트의 순위와 상관 관계가 있는 것은 아니며 장르에서 두드러지는 점은 pop의 경우 차트의 순위가 1에 가까우며, country의 경우 100에 가까운 순위를 갖는다는 점이다. 이것은 빌보드 순위에서 인기가 있는 장르와 아닌 장르가 있다는 것을 알 수 있는 관계이다.

추출한 특성과 미래의 빌보드 차트의 상관관계를 분석한 결과, 트위터의 특성은 높은 상관 관계를 가지고 있으며 다른 특성 모두 차트에 영향을 끼치는 요소임을 알 수 있다. 더불어 broadcaster가 crowds보다 차트 순위와 높은 상관관계를 가져 차트를 예측함에 있어서 주요한 역할이라는 것을 도

출할 수 있다.

6. 새로운 feature 제시 및 음악 차트와 상관 관계

관개성 분석을 바탕으로 제시된 독립적인 특성간의 관계를 통해 의미 있는 feature를 도출해내고자 한다. 우선적으로 상관 관계 분석에서 알 수 있었던 것은 장르간에 상관계수가 서로 다르다는 것이었다. 앞서 분석을 통해 트위터 사용자들의 음악 청취 행태가 장르에 따라서 다르기 나타난다는 사실을 직접적으로 알 수 있다. 따라서 트위터 특성에 장르별로 다른 가중치를 준다면 예측에 있어서 보다 유용한 feature로서 사용될 수 있을 것이다. 본 연구에서는 사용자 그룹별로 song popularity의 평균값과 genre에 대한 회귀 모델을 세우고 이를 새로운 feature로서 제시하였다.

$$f = \sum(\beta_x * Song_{user} * genre_x) + \alpha$$

user 변수에는 사용자를 뜻하는 {crowds, Broadcasters} 이며, genre 변수는 장르 {pop, hiphop, country, dance, rock, R&B, soundtrack, holiday, others, the voice}를 뜻한다. Song변수는 사용자 그룹별 일주일간 song popularity 평균의 normalized된 값이다. 빌보드 순위와의 Linear Regression을 통해 얻어진 coefficient는 다음과 같다.

	Intercept	Pop	Dance	Country	R&B	Sound track	The voice
Crowds	49.86	−14.11	−14.23	−17.17	−46.47	−8.48	−19.43

							(0.02)
Broadcast	49.52	−18.04	−11.8	−23.03	−36.6	−6.02	−5.23 (0.30)

표 7. 빌보드 순위에 대한 제시된 feature에서 나타난 장르의 coefficient

더불어 데이터를 분석하면서 얻은 경험적 지식을 바탕으로 Song popularity, Artist popularity, days after release가 각 사용자 그룹별로 관계성을 지니고 있음을 알 수 있었다. Crowds의 경우, Song popularity는 Artist popularity가 클수록 그 수가 더 많아진다는 것을 알 수 있다. 또한 발매 이후 시기가 많이 지났는데 어느정도 song popularity를 유지하고 있다면 높은 순위일 가능성이 높다는 것이다. 반면에 Broadcaster 그룹은 Artist popularity가 크다고 song popularity가 그에 따라서 커지는 경향성이 큰 것은 아니며, 발매 이후 시기가 지날 수록 song popularity가 커는 모습을 보였다. 이러한 부분에서 조정이 필요할 것이라 생각하여 제시한 새로운 feature는 다음과 같다.

$$B_combined = Song * \sqrt{Artist} / \log(t+1)$$
$$C_combined = Song * \log(t+1) / \sqrt{Artist}$$

t는 발매후 지난 시간을 의미한다.

조정된 각 feature들을 통해 회귀 모델을 세우고 이에 따라 다시 새로운 feature로 제시한다. 다음은 새로이 만들어진 feature들의 스피어만 상관계수를 나타낸 것이다. 이들은 이전의 트위터 특성보다 좀 더 상관 계수가 높

은 것을 알 수 있다.

Type	Feature	Correlation coefficient
New	C_song*genre	-0.6008***
	B_song*genre	-0.6714***
	C_combined + B_combined	-0.6644***
	C_combined	-0.5922***
	B_combined	-0,5952***

표 8. 제시한 특성과 빌보드 순위와의 스피어만 상관 계수

제 2 절 음악 순위 예측

연구문제 4의 해결을 위하여 음악 순위 예측을 수행하고 각 특성의 조합별로 서로 다른 모델을 생성하여 모델의 성능을 평가한다. 앞서 추출한 특성과 새로이 만들어낸 특성을 통해서 미래의 빌보드를 시간에 따라 기계학습을 바탕으로 예측하고자 한다. 기계 학습에서 학습 데이터와 테스트 데이터는 5 fold cross validation을 통해 균일하게 평가하였다.

각 특성의 조합에 따라 미래의 음악 순위를 예측하는 회귀 모델을 만들고 이에 대한 평가를 진행하였다. 회귀모델은 Support Vector Regression을 사용하였으며 평가 방식은 R squared를 기준으로 하였으며 1주부터 8주 후까지 미래를 예측하였다. 일주일간의 데이터를 사용해 모델, 시간 별로 예측한 회귀 모델의 R squared값은 다음과 같다.

Model	1	2	3	4	5	6	7	8
Twitter + user	0.63	0.64	0.61	0.55	0.53	0.52	0.48	0.47
Twitter + user + RT	0.64	0.64	0.62	0.57	0.55	0.53	0.5	0.47
Twitter + user+ RT + genre	0.71	0.71	0.68	0.65	0.63	0.63	0.61	0.59
Twitter+ user+ RT + genre+ New	0.74	0.73	0.71	0.67	0.66	0.66	0.63	0.63

표 9. 시간과 모델에 따라 순위를 예측한 R squared 값

표 9를 통해 예측 시간에 따라서 R squared 값이 줄어듦으로써 당연하게도 먼 미래를 예측하는 것이 힘들다는 것을 알 수 있었다. 3주 정도 전까지는 준수하게 예측을 하다가 4주차부터 상당히 정확도가 많이 떨어지는 것으로 보인다. Retweet의 경우 별다른 모델의 진전을 주지 못했으며 Genre와 days after release의 특성은 예측에 많은 도움을 주었다. 새롭게 제시한 New feature들을 통해 모델이 보다 정확해졌음을 알 수 있다. 하지만 많은 특성을 사용하지 않고 트위터의 특성만을 사용하였을 때도 성능이 0.43으로 상당히 준수함을 알 수 있었다. 이를 바탕으로 새로운 feature와 트위터만을 사용한 특성을 통한 예측을 비교하여 보았고 결과는 그림 18과 같다.

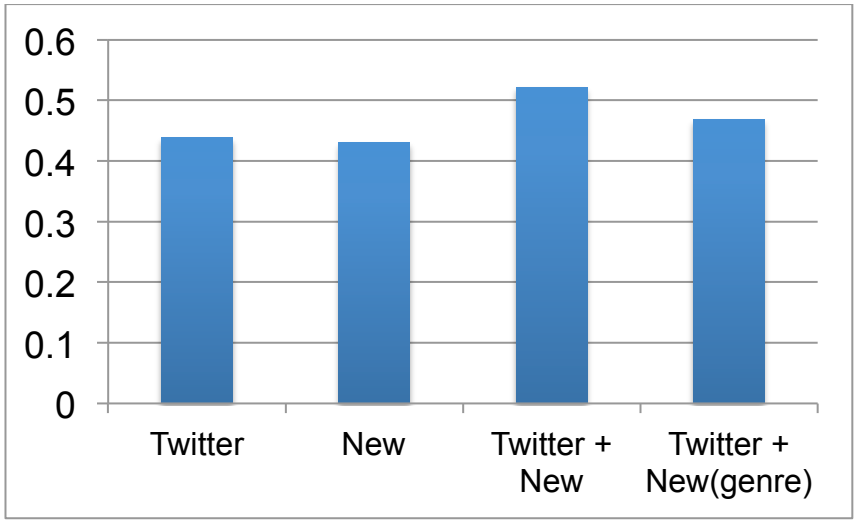

그림 18. 트위터 특성과 새로운 특성에 따른 예측 정확도 (R squared)

트위터와 새로 제안한 특성을 각각 사용하였을 때는 8주 뒤를 예측 할 때 약 0.43의 비슷한 R squared를 보였다. 여기서 트위터의 특성과 새로 제안한 특성이 서로 보강 관계를 가지는지 확인하기 위하여 두 특성 바탕으로 모델을 수립하였을 때 0,51로 상승하는 것을 볼 수 있었다. 따라서 새로 제안한 특성은 기존의 트위터 특성에 보강하여 사용할 수 있는 유용한 특성이라고 할 수 있다.

두 번째로, 초기 모델은 7일간의 데이터를 사용하였지만, 이를 3일과 5일로 추가적으로 더 적은 데이터를 주었을 때 어떠한 변화가 있는지를 살펴보았다. 아래 그래프를 보면 7일간의 데이터를 사용하였을 때 가장 정확도

가 높고, 이후 사용한 날짜가 줄어들 수록 예측 정확도도 낮아지는 것을 볼 수 있다.

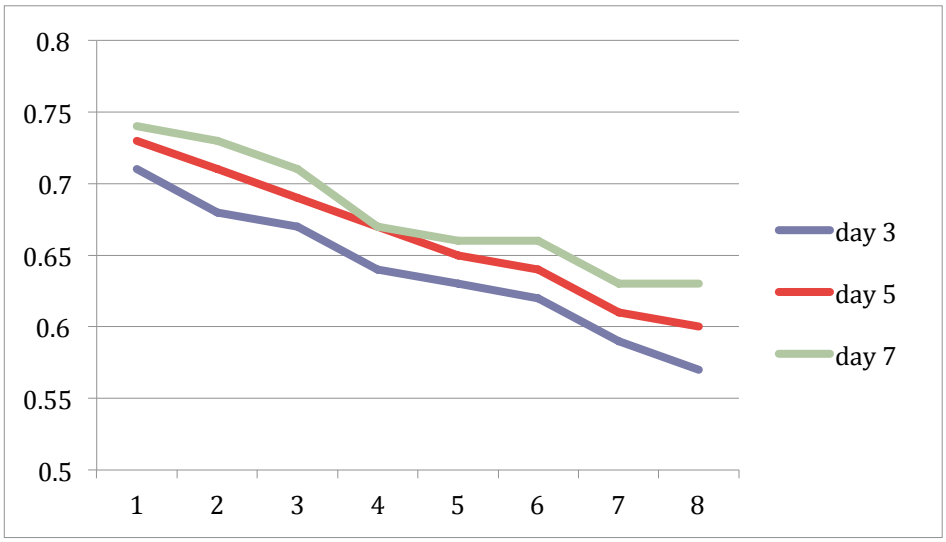

그림 19. 사용한 데이터의 일 수에 따른 예측 정확도 (R squared)

실험을 통해서 트위터의 음악 청취 행태에서 추출된 데이터와 나뉘어진 사용자 그룹이 미래의 음악 순위 예측에 있어서 중요하게 작용한다는 것을 알 수 있었다. 또한 새로이 제시된 특성 역시 보다 정확한 예측 모델을 만드는데 기여하여 의미 있는 특성이라 할 수 있다.

제 5 장 결론

제 1 절 요약

본 논문에서는 트위터의 음악 청취 행태를 수집하고 특성을 추출하여 이들의 관계성을 빌보드 차트를 통해 분석하고 순위를 예측하는 실험을 수행하였다. 이를 위해 트위터의 #nowplaying를 해쉬태그를 이용하여 음악 청취 행태를 수집하였으며 사용자별 트윗 생성 수와 메타데이터를 이용하여 사용자 그룹을 Crowds 와 Broadcasters 두 그룹으로 나누었다.

나뉘어진 그룹의 Song popularity와 빌보드 순위와의 상관 관계를 분석한 것과 더불어, 장르와 Artist popularity 관점에서 두 그룹이 어떠한 차이점과 특징을 가지고 있는지를 나타내었다. 또한 발매 이후 지난 시간이 두 그룹의 song popularity에 어떻게 다른 영향을 주는지를 분석했다.

앞서 분석한 관개성을 바탕으로 각 사용자 그룹에따라 다른 feature를 제안하였으며, 이를 통해서 단순 트위터의 특성 뿐만 아니라 장르와 같은 음악의 특성도 연결시켜 새로운 특성을 제시하였다.

마지막으로 추출된 특성을 통해 미래의 순위를 예측하였는데, 이 과정에서 특성별 모델링을 통해 각 특성들이 얼마나 주요하게 사용될지를 관찰하고, 더불어 트위터의 음악 청취 행태와 빌보드 차트가 가지는 연관성에 대해서도 평가하였다.

제 2 절 의의

기존에는 일반적인 연구자들이 한정적인 정보로만 접근하였던 사용자들의 음악 청취 행태를 마이크로 블로그를 통해 얻어내었다. MIR분야에서 소셜 데이터가 한정적으로 사용되었고, 특히 마이크로 블로그가 타 분야에 비해 상대적으로 연구가 진행되지 않았다. 또한 성격이 다른 사용자 그룹 분류를 통해 각각의 사용자 그룹이 트렌드에 따라 어떠한 음악 청취 행태를 가지고 있는지 분석한 연구는 찾아보지 못하였다. 더불어 장르, 아티스트 선호도, 발매 이후 시간의 요소를 사용하여 관계성을 다양하게 분석했다는 점에서 학문적 의의가 있다. 이 연구는 음악 트렌드에 대한 사용자 음악 청취 행태를 보다 자세히 알 수 있어 사용자 중심 연구에 도움을 줄 수 있을 것이다.

제 3 절 연구의 한계점

본 연구를 통해 트위터 사용자의 음악 청취 행태와 음악 트렌드의 관계성을 도출하였으나, 부족한 데이터, 실험 설계 및 통계적 검증에 대한 어려움 등 연구의 한계점이 존재하였다.

1. 부족한 데이터

6개월의 트위터 수집을 통해 순위 예측에 사용한 데이터는 2400개이며 이는 Machine Learning에서 학습하고 테스트를 하기엔 부족한 데이터의 수라고 할 수 있다. 따라서 결과값에 따라서 절대적인 평가를 내리기는 불분명하며 오랜 기간 관찰을 통해 데이터를 증가시키는 것이 중요하다.

또한 빌보드 순위 예측을 위해 사용된 데이터는 빌보드에 등재되어있는 데이터만을 대상으로 수행된다. 이는 트위터에서 언급된 전체 데이터가 아닌 아주 일부분의 데이터만을 사용한 것으로, 이미 빌보드에 등재되어있는 데이터를 사용하기 때문에 트위터 사용자의 전체 음악 청취 행태와 빌보드 순위의 관계성을 보았다기 보다는 트위터 사용자들의 빌보드 순위에 대한 청취 선호도 정도로 해석할 수 있다. 즉 트위터 사용자들이 빌보드에 올라와있는 음악에 대해 어떠한 선호도를 가지고 있는지를 규명한 것이라고 할 수 있다. 빌보드 차트는 미국만을 대상으로 하기에 이후 여러 가지 차트를 통해서 다양한 지역의 트렌드를 규명 가능할 것이다.

2. 실험 설계 및 통계적 검증

음악 순위 예측에서 사용한 알고리즘인 Support Vector Regression은 어떠한 특성이 중요한지 나타내기 어려우므로 선형 회귀 분석과 같은 사용하여 규명한다면 특성의 중요성을 정량적으로 알 수 있을 것이다. 더불어, 음악 청취 행태와 트렌드의 관계성을 분석을 하는 과정에서 각 특성에 대한 분포를 시각적으로 나타내어 그 차이점을 설명하였는데, 보다 세밀하게 통계적으로 검증하는 정량적인 평가 방식이 필요할 것으로 생각된다.

참고 문헌

[1] A. Oulasvirta, E. Lehtonen, E. Kurvinen, and M. Raento. Making the ordinary visible in microblogs. Personal Ubiquitous Comput., 14(3):237-249, Apr. 2010.

[2] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE, 2010.

[3] Benslimane, Djamal, Schahram Dustdar, and Amit Sheth. "Services Mashups."*The New Generation of Web Applications* (2008): 13-15.

[4] Bhattacharjee, Sudip, et al. "Using P2P sharing activity to improve business decision making: proof of concept for estimating product life-cycle." *Electronic Commerce Research and Applications* 4.1 (2005): 14-20.

[5] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 2009.

[6] Boyd, Danah, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet:

Conversational aspects of retweeting on twitter." *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 2010.

[7] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In Proceedings of NAACL HLT, June 2010

[8] Cunha, Evandro, et al. "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach." *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011.

[9] David A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press (2005)

[10] Dewan, Sanjeev, and Jui Ramaprasad. "Chicken and Egg? Interplay between Music Blog Buzz and album Sales." *PACIS*. 2009.

[11] Dhanaraj, R. and Logan, B. Automatic Prediction of Hit Songs, Proc. of Ismir 2005, London, UK

[12] Friedl, Mark A., and Carla E. Brodley. "Decision tree classification of land cover from remotely sensed data." *Remote sensing of environment* 61.3 (1997): 399-409.

[13] Granger, Clive WJ. "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica: Journal of the Econometric Society*(1969): 424-438.

[14] Gayo-Avello, Daniel. "No, you cannot predict elections with twitter." *Internet Computing, IEEE* 16.6 (2012): 91-94.

[15] Glasgow, Kimberly, and Clayton Fink. "Hashtag lifespan and social networks during the london riots." *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 2013. 311-320.

[16] Hauger, David, and Markus Schedl. "Exploring Geospatial Music Listening Patterns in Microblog Data." *Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR), Copenhagen, Denmark*. 2012.

[17] Hauger, David, et al. "The Million Musical Tweets Dataset: What Can We Learn From Microblogs." *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil*. 2013.

[18] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: A Comparison of Microblog Search and Web Search. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM' 11), Hong Kong, China, Feb 2011.*

[19] Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu Social Knowledge-Driven Music Hit Prediction

[21] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[22] Luo, Zhunchen, et al. "Who will retweet me?: finding retweeters in twitter." *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013.

[23] Ma, Zongyang, Aixin Sun, and Gao Cong. "Will this# hashtag be popular tomorrow?." *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012.

[24] Marvell, T.B., C. E. Moody. 1996. Specification Problems, Police Levels,

and Crime Rates. Criminology 34(4) 609-646

[25] Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello. "How (not) to predict elections." *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*. IEEE, 2011.

[26] M. J. Paul and M. Dredze. You Are What You Tweet : Analyzing Twitter for Public Health. Artificial Intelligence, pages 265-272, 2011. *Conference on World Wide Web (WWW 2010), May 2010.*

[27] M. Schedl and D. Hauger. Mining Microblogs to Infer Music Artist Similarity and Cultural Listening Patterns. *In Proceedings of the 21st International World Wide Web Conference (WWW 2012): 4th International Workshop on Advances in Music* Information Research: "The Web of Music" (AdMIRe 2012), Lyon, France, 2012

[28 Pachet, F. and Roy, P. (2008) Hit Song Science is Not Yet a Science. *Proceedings of Ismir 2008,* pages 355-360, Philadelphia, USA

[29] [25] Pachet, F. (2011) Hit Song Science. [1]In Tao, Tzanetakis & Ogihara, editor, Music Data Mining, CRC Press/Chapman Hall

[30] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real- Time Event Detection by Social Sensors. *In Proceedings of the 19th International*

[31] Weerkamp, Wouter, Manos Tsagkias, and Maarten de Rijke. "Inside the world's playlist." *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013.

[32] Robert E. Vossen Does Chatter Matter? Predicting Music Sales with Social Media

[33] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real- Time Event Detection by Social Sensors. *In Proceedings of the 19th International*

[34] Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10 (2010): 178-185.

[35] Twitter Inc., Twitter API Documentation https://dev.twitter.com/docs

[36] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake

shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[37] Schedl, Markus, and David Hauger. "Mining microblogs to infer music artist similarity and cultural listening patterns." *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012.

[38] Schedl, Markus. "Leveraging microblogs for spatiotemporal music information retrieval." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2013. 796-799.

[39] Shao, Bo, et al. "Music recommendation based on acoustic features and user access patterns." *Audio, Speech, and Language Processing, IEEE Transactions on* 17.8 (2009): 1602-1611.

[40] Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? Large scale analytics on factors impact- ing retweet in twitter network. In *Social Computing (So- cialCom), 2010 IEEE Second International Conference on,* 177–184. IEEE.

[41] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who Says What to Whom on Twitter. *In Proceedings of the 20th International Conference on World Wide Web (WWW 2011),* pages 705–714, New York, NY, USA, 2011. ACM.

[42] Zhao, Dejin, and Mary Beth Rosson. "How and why people Twitter: the role that micro-blogging plays in informal communication at work." *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009.

[43] 황용태, “Social Network Service의 Data 위주 연구를 위한 Platform 설계”, 석사학위논문, *서울대학교 대학원*, 2012

Abstract

Crowds vs. Broadcasters: Predicting Billboard Rank Using Characteristic of Music Listening Behavior in Twitter

Kim Ye-Kyung
Department of Digital Contents Convergence
The Graduate School
Seoul National University

As distribution and consumption platforms of music change from offline to online, music streaming services like Spotify, Last.fm achieve a success in online and log their users' music consumption history. Music consumption log gives us the chance to understand listening behavior of users.

However, there are many services for the music streaming, only Last.fm provides music listening log with their public API. Since the bulk of music data related to users on such online music service is hard to access for general purpose, the source of information about music listening log in MIR area is limited.

In this paper, we extracted Tweets about real-time music listening history posted with hashtag #nowplaying, #np and classify two user groups by their description. To understand the characteristic of music listening behavior in

Twitter about general music trend, we observe the relationship between music listening count and Billboard rank each user groups focusing on genre of music, artist popularity, and days after release. In addition, we propose the feature based on their characteristic and predict the success of Billboard music before two weeks using features of each user group

The different characteristic obtained from classified user group could be used in the social advertisement and inform the improvement of the performance of music hit prediction. Also it gives us the opportunities to understand music consumption patterns of various user groups. We wished to our research crossing the music to Twitter held new possibilities in MIR research area.

Keywords : Social Media Mining, Music Listening behavior, Twitter, Rank Predction.

Student Number : 2013-22403