PageRank 를 활용한 음악 차트 분석

2015104194 이규호

개 요

구글을 만들어낸 기술인 PageRank 는 광범위한 웹사이트를 분석하여 검색 엔진의 새 시대를 열었다. 이는 단순히 사이트 검색에만 활용되는 것이 아니라 Text, 스포츠, 이미지 등 다양한 분야에서 활용될 수 있다. 본 연구에서는 이를 음악에 접목시켜 차트를 분석하고, 검증해보며 새로운 활용방안을 제시해본다.

1. 서론

1.1 연구 배경

모바일 기기의 보편화로 콘텐츠 소비가 간편해지면서 콘텐츠의 영향력은 더욱 커졌고, 콘텐츠 산업 또한 지속적으로 성장하고 있는 추세이다. 데이터의 중요성 역시 강조됨에 따라 개인이 소비하는 콘텐츠 역시 데이터를 남기며 방대한 양의 데이터를 이용해 시장을 분석할 수 있는 가능성이 생겼다.

특히 음악 분야는 남녀노소 불문하고 소비하는 접근성이 높은 분야이다. 소비자들의 소비 패턴을 예측할 수 있는 음원 차트를 분석한다면 특정 음원에 대한 소비자의 관심을 예측할 수 있다. 분석 기법으로는 랭킹에 특화된 PageRank 알고리즘을 사용한다.

1.2 연구목표

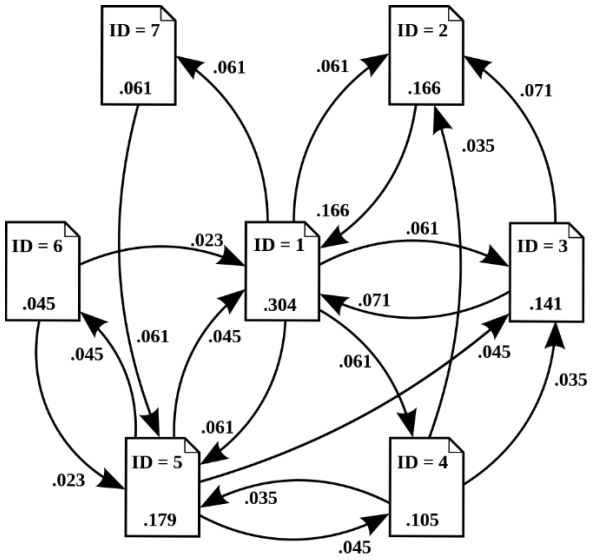
본 연구에서는 이전 음원들의 정보와 순위를 기반으로 새로운 음원의 흥행 정도를 예측한다. 구체적으로 가수의 성별 및 인지도, 피처링 가수의 정보, 노래의 장르, 가사, 음원 사이트의 좋아요 수, 소속사, 앨범의 종류 등을 활용해서 음원의 데이터를 분석한다. 이는 연구를 진행하면서 높은 정확도를 보이는 데이터들만 수집하는 등 바뀔 수 있다.

2020 년은 COVID-19 라는 특수성을 갖고 있기 때문에, 그 이전의 차트를 기반으로 분석한다. 2010 년 ~ 2018 년의 차트를 분석하여 모델을 설계한다. 그리고 2019 년의 차트를 검증 데이터로 활용하여 모델의 정확도를 높이는 것에 중점을 둔다.

2. 기존 연구

2.1 Page Rank

월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법으로, 웹사이트 페이지의 중요도를 측정하기 위해 구글 검색에 쓰이는 알고리즘이다. 이 알고리즘은 서로 간에 인용과 참조로 연결된 임의의 묶음에 적용할 수 있다.

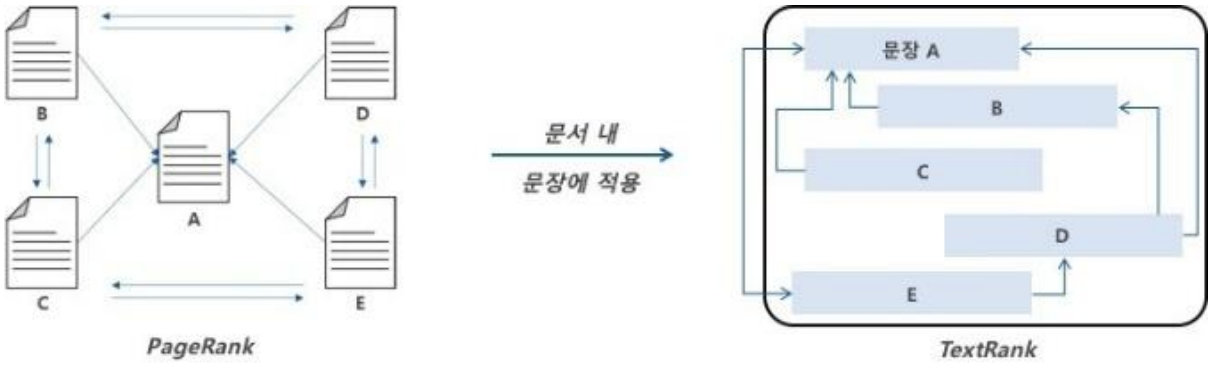


[그림 1] PageRank Algorithm 개념도

Page Rank 는 더 중요한 페이지는 더 많은 사이트로부터 링크를 받는다는 관찰에 기초하고 있다. 또한 Random Surfer 라는 페이지를 임의로 방문하며 탐색하는 모델을 가정한다. 이 모델에서는 A 페이지를 방문한 서퍼는 A 페이지를 보고 만족하여 탐색을 중단하거나, 혹은 A 페이지에서 만족하지 못하여 다른 페이지를 방문할 것이다. 이러한 확률을 α 라 한다면, B 페이지는 α * 1 / 3 만큼 Page Rank 를 받게 된다. 이와 같은 방법을 통해 페이지간 랭크 값을 주고 받는 것을 반복하다 보면, 전체 웹페이지가 특정한 랭크 값으로 수렴한다는 사실을 통해 각 페이지의 최종 랭크를 계산한다.

2.2 TextRank

PageRank 를 기반으로 글의 키워드와 핵심문장을 골라내기 위해 가장 많이 사용되는 알고리즘 중 하나인 TextRank 알고리즘이다. TextRank 알고리즘은 단어 graph 나 문장 graph 를 구축하고 구글의 PageRank 알고리즘을 이용하여 키워드와 핵심문장을 골라낸다. 그리고 이러한 방식으로 골라낸 키워드와 핵심문장을 이용하여 글을 요약하는 방식이다.



[그림 2] PageRank -> TextRank 적용

TextRank 을 위한 식은 다음과 같다.

$$TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TR(V_j)$$

- $TR(V_i)$: 문장 또는 단어(V)i 에 대한 TextRank값
- w_{ij}: 문장 또는 단어 i 와 j 사이의 가중치
- d : dampingfactor PageRank에서 웹 서핑을 하는 사람이 해당 페이지를 만족하지 못하고 다른페이지로 이동하는 확률로써, TextRank에서도 그 값을 그대로 사용(0.85로 설정)
- TextRank $TR(V_i)$ 를 계산 한 뒤 높은 순으로 정렬

[그림 3] TextRank Algorithm

2.3 기존 연구의 문제점

기존에 PageRank 알고리즘을 활용하여 음악 차트를 분석한 연구는 없지만, 딥러닝의 일반적 방법인 DNN 을 적용시킨 선례가 존재한다. 이는 DNN 의 본질적인 단점을 그대로 갖고 있다. 분석 단계에서 변수들을 일정한 순서나 방식으로 넣는 것이 아니기 때문에 결과가 일정하지 않고, 가중치의 의미를 정확히 해석하기가 어렵기 때문에 모델 학습을 통한 결과 해석이 어렵다.

3. 프로젝트

3.1 기존 연구와 차이점 및 해결방안

본 연구에서는 선례와 다르게 PageRank, TextRank 알고리즘을 활용한다. 이는 노드간 연결성을 의미하는 Edge 를 갖기 때문에, 음악 간의 관계를 보다 면밀하게 분석할 수 있다. 또 가사에도 TextRank 를 적용하여 주요 키워드들을 추출해내기 때문에 기존 연구에서 넣지 못했던 가사에 대한 분석도 가능하다. 음악을 들을 때 가사를 고려하는 소비자들이 많다는 것을 생각해보면, 더 높은 정확도를 기대할 수 있다.

3.2 프로젝트 내용

가사를 분석하는 TextRank 에 사용되는 형태소 분석은 python 의 Komoran 을 활용한다. 각 단어간의 co-occurrence 를 계산하여 TextRank 를 적용, 주요 키워드를 추출해내서 노드내 가사의 정보로 사용할 수 있다.

이후 1.2 에서 정의한 음악의 다양한 정보들을 사용해서 음악 노드 내 데이터로 입력한다. 노드 사이의 엣지는 노드간의 코사인 유사도를 측정하여 생성한다. 각 데이터의 중요도는 연구를 진행하면서 높은 정확도를 갖는 방향으로 수치를 조정한다.

4. 결론 및 기대효과

본 연구에서는 PageRank 기법을 음원 차트 예측에 적용한다. 다만, 탐색적인 수준에서 진행되는 연구이기에 실제 음악에 적용했을 때, 높은 정확도를 보일 수 있을지 예상하기 힘들다. 다만 컨텐츠 분야에서 본 접근방식과 유사한 연구가 거의 없기 때문에, 본 연구는 PageRank 를 활용한 문화 산업에 대한 연구에 기반이 될 수 있을 것이다.

본 연구를 통해 음악의 어떤 정보가 차트인에 영향을 주는지 대략적으로 알 수 있다. 이는 음악 산업 프로세스 전략의 기반이 될 수 있다. 실제 음악 제작과정에 어떠한 요인들을 고려해야 될 지 결정할 수 있을 것이다. 보다 방대한 데이터를 활용해서 모델을 설계한다면 트렌드의 움직임을 실시간으로 파악할 수 있을 것이다.

5. 참고 문헌

[1] Google, The PageRank Citation Ranking: Bringing Order to the Web, 1998

[2] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출", 2009

[3] 이도연, 장병희, "딥러닝을 이용한 음악흥행 예측모델 개발 연구", 2020

[4] 조은혜, 박재현, 최창환, "PageRank 알고리즘을 활용한 국내 배드민턴 선수들의 랭킹산정", 2018