# PageRank를 활용한 음악 차트 분석

이규호° 한치근
경희대학교 컴퓨터공학과

# Analyzing music charts using PageRank algorithm

GyuHo Lee° ChiGeun Han
School of Computer Science and Engineering, Kyung Hee University

**요 약**

구글을 만들어낸 기술인 PageRank는 광범위한 웹사이트를 분석하여 검색 엔진의 새 시대를 열었다. 이는 단순히 사이트 검색애만 활용되는 것이 아니라, Text, 스포츠, 이미지 등 그래프로 만들 수 없다고 생각했던 다양한 분에서 활용되어 분석 가능하다. 이에 본 논문에서는 PageRank를 음악에 접목시켜 차트를 분석하고, 검증해보며 흥미로운 결과를 분석해보고, 새로운 음원이 들어왔을 때 흥행 정도를 예측하여 검증한다.

## 1. 서 론

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 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.

본 연구에서는 이전 음원들의 정보와 순위를 기반으로 새로운 음원의 흥행 정도를 예측한다. 구체적으로 가수의 성별 및 인지도, 노래의 장르, 가사, 음원 사이트의 좋아요 수, 앨범의 종류 등을 활용해서 음원의 데이터를 분석한다.

각 Column에 해당하는 제목 및 가사는 예측 모델을 구현할 때 숫자 값이 아니라, 정제하지 않고 사용한다면 정확한 결과를 예측하기가 어렵다. 따라서 이 문제를 해결하기 위해 TextRank 기법을 사용해서 Word2Vec을 적용한다. 이는 음악 차트 순위를 가중치로 활용하여 더 높은 순위의 노래에서 나온 단어가 더 높은 값을 갖도록 구현한다.

## 2. 기존 연구

### 2.1 PageRank

웹드 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.

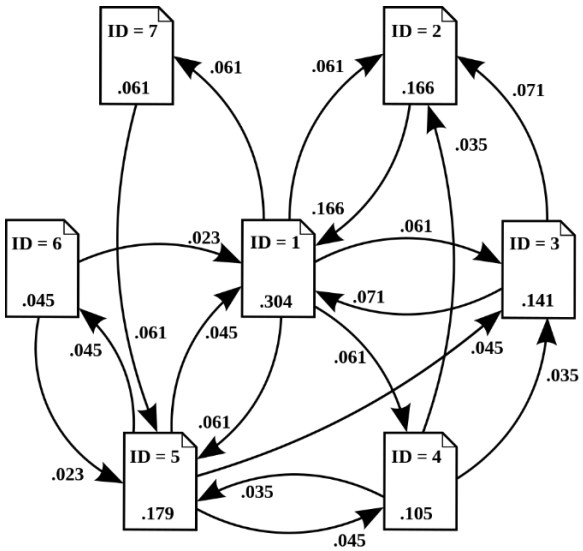


그림1 PageRank Algorithm 개념도

Page Rank는 더 중요한 페이지는 더 많은 사이트로부터 링크를 받는다는 관찰에 기초하고 있다. 또한 Random Surfer라는 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 α라고 한다면, B 페이지는 α / 3 만큼 PageRank를 받게 된다. 이와

같은 방법을 통해 페이지간 랭크 값을 주고받는 것을 반복하다 보면, 전체 웹페이지가 특정한 랭크 값으로 수렴한다는 사실을 통해 각 페이지의 최종 랭크를 계산한다.

### 2.2 TextRank

PageRank를 기반으로 글의 키워드와 핵심문장을 골라내기 위해 가장 많이 사용되는 알고리즘 중 하나이다. TextRank 알고리즘은 단어 graph나 문장 graph를 구축하고 PageRank 알고리즘을 이용하여 키워드와 핵심 문장을 골라낸다. 그리고 이러한 방식으로 골라낸 키워드와 핵심문장을 이용하여 글을 요약하는 방식이다.

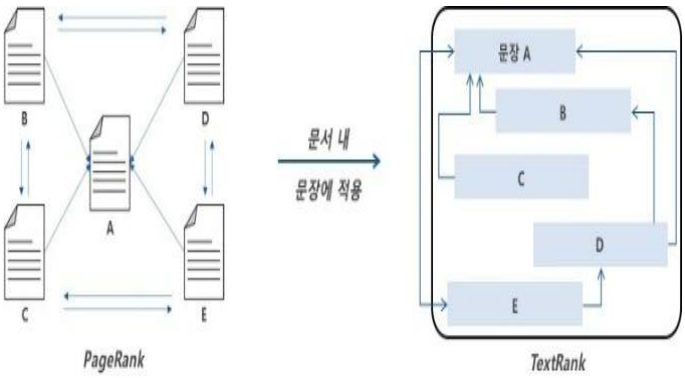


그림2 PageRank → TextRank 적용

## 3. 문제 정의

기존에 PageRank 알고리즘을 활용하여 음악 차트를 분석한 연구는 없지만, 딥러닝의 일반적인 방법인 DNN을 적용시킨 선례가 존재한다. 이는 DNN의 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.

본 연구에서는 PageRank, TextRank 알고리즘을 활용하기에 노드간 연결성을 의미하는 Edge를 갖는다. 때문에 음악 간의 관계를 보다 면밀하게 분석할 수 있다. 또 가사에도 TextRank를 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 있다고 판단된다.

## 4. 연구 내용

### 4.1 데이터 수집

본 연구에서는 예측 모델을 구축하기 위하여 멜론의 차트 파인더를 기반으로 데이터 수집을 진행한다. 멜론의 차트파인더는 1990년대부터의 월간 Top 100을 지원하며 그 곡과 가수에 대한 정보도 알려주면서, 사용자들이 가장 많이 사용하는 음원 플랫폼이기 때문에 신뢰할 수 있는 데이터라고 판단했기 때문이다.

구체적으로 2011년 1월 ~ 2021년 4월까지의 월간 차트 Top 100에 있는 총 12400개의 곡을 사용하며, 총 838팀의 가수가 사용되었다. 곡의 정보에 대한 테이블과 가수의 정보에 대한 테이블을 따로 생성하여 곡의 정보는 연, 월, 순위, 노래 제목, 가수, 앨범, 장르, 출시일, 좋아요 수, 가사를 변수로 갖는다. 그리고 가수의 정보에 대한 테이블은 가수명, 남/여/혼성, 그룹/솔로, 좋아요 수를 변수로 갖는다.

### 4.2 예측 모델

예측 모델을 구현하기 위해 추출한 데이터의 정제가 필요하다. DNN(Deep Neural Network)를 통해 rank 값을 나타내기 위해서, 기존 feature에 해당하는 값 중 텍스트 형태인 제목, 장르, 가사를 숫자 값으로 바꿀 필요가 있다. 장르는 한정되어 있기에 category 형태로 one-hot vector로 표현이 가능하다. 그리고 제목과 가사 부분에 textrank를 적용하여 일종의 워드 임베딩을 진행하여 가중치를 부여한다. 이렇게 전처리된 데이터를 기존 딥러닝 모델에 넣으면, y값에 해당하는 순위 값을 얻을 수 있다. 자세한 모델과 optimizer, cost function 등은 다양한 기법을 사용해보고 가장 정확도가 높은 모델로 선택할 예정이다.

예측 모델을 구현하기 위해 추출한 데이터의 정제가 필요하다. DNN(Deep Neural Network)를 통해 rank 값을 나타내기

## 5. 결론 및 향후 연구

본 연구에서는 PageRank 기법을 음원 차트 예측에 적용한다. 다만, 탐색적인 수준에서 진행되는 연구이기에 실제 음악에 적용했을 때, 높은 정확도를 보일 수 있을 지 예상하기 힘들다. 다만 컨텐츠 분야에서 본 접근방식과 유사한 연구가 거의 없기 때문에, 본 연구는 PageRank를 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 제작과정에서 어떠한 요인을 고려해야 될 지 결정할 수 있을 것이다. 보다 방대한 데이터를 활용해서 모델을 설계한다면 트렌드의 움직임을 실시간으로 파악할 수 있을 것이다.

## 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 랭킹 산정”, 2018