산업공학 시스템 설계

토픽 모델링을 활용한 콘텐츠 기반 추천시스템

Content-based music recommendation system by using topic modeling

임연우 | 이진아 | 최승원

홍익대학교 산업공학과

2018. 12

목 차

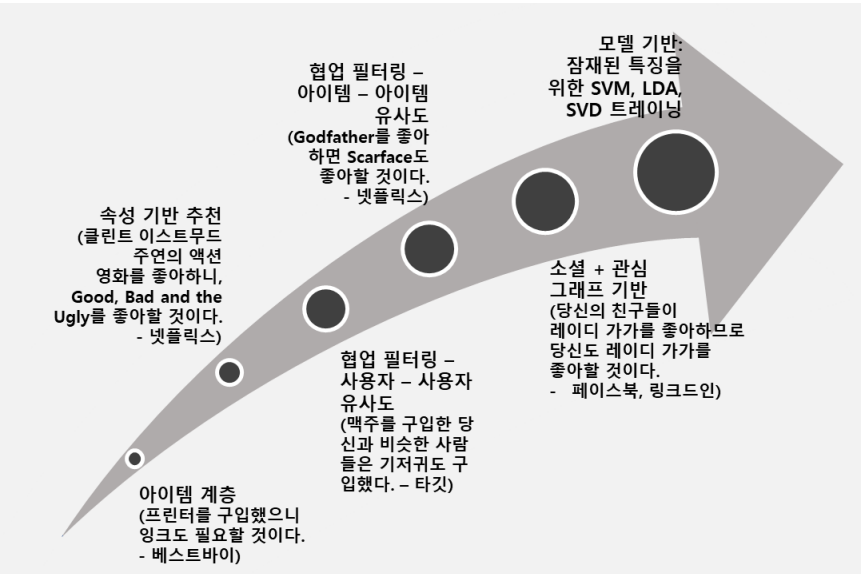
1장. Introduction(서론)

본 논문에서는 사용자가 원하는 정보를 예측하고 찾아주는 개인화 수준의 추천시스템을 구현해보았다. 1장에서는 추천시스템 선정배경 및 목적, 추천시스템에 사용한 기법과 기대효과를 다루어 보겠다.

1.1 주제 선정배경 및 목적

현재는 개인화된 상품 및 서비스의 시대이다. 상품, 서비스, 미디어 등이 모두 충분히 넘치는 과잉의 시대이다. 이러한 개인화 시대는 추천시스템 분야에서도 적용된다. 추천시스템이란 사용자가 선호할 만한 아이템을 추측함으로써 여러 가지 항목 중 사용자에게 적합한 특정 항목을 선택(information filtering)하여 제공하는 시스템을 일컫는다 여기서 'filtering' 이란 여러 가지 항목 중 적당한 항목을 선택하는 기술을 말하는 IT용어다. [1]

추천시스템의 발전동향을 살펴보면 그림1과 같다. 초기의 추천시스템은 아이템 계층 기반 추천시스템으로 아이템의 계층을 바탕으로 상위항목을 구입하면 그에 속한 하위항목을 추천해주는 방식이다. 이후 아이템의 특정 속성을 바탕으로 유사한 속성을 가진 아이템을 추천을 해주는 속성 기반 추천이 이루어졌다. 이후 대규모의 기존 사용자 행동 정보를 분석하여 해당 사용자와 비슷한 성향의 사용자들이 기존에 좋아했던 항목을 추천하는 협업 필터링(Collaborative filtering)을 이용하는 추천시스템으로 발전하였다.



[그림 1. 추천시스템의 발전동향]

황영숙(2016)에 따르면 추천 시스템에 사용되는 대표적인 필터링 기법으로는 아이템에 대한 다른 사용자의 평가 정보를 사용하여 사용자들을 비슷한 선호도를 가진 집단으로 나누고 그 집단 내에서 서로에게 추천해주는 방식인 협업 필터링 기법과 아이템을 구성하는 요소들을 사용자 프로파일과 비교하여 아이템을 추천해주는 방식인 콘텐츠 기반 필터링 기법이 있다. [2]

협업 필터링(Collaborative Filtering: CF)은 구매/소비한 제품에 대한 각 소비자의 평가를 받아서 평가 패턴이 비슷한 소비자를 한 집단으로 보고 그 집단에 속한 소비자들의 취향을 활용하는 것이다. 이 기술은 사람들의 취향이 뚜렷이 구분되는 제품(예를 들어 영화나 음악, 패션)의 경우 정확하다고 알려져 있다. 이 기술에는 특정 수 이상의 아이템 에 대한 평가 정보가 없으면 추천할 수 없다는 점과 비슷한 성향을 가지는 일부 사용자 정보에 근거하여 추천함으로써 나머지 사용자 정보가 무시된다는 한계가 있다. [2,3]

콘텐츠 기반 필터링(Content-based Filtering: CF)은 아이템의 특징을 기술하는 정보와 사용자의 기호를 가지고 있는 프로파일을 비교하여 사용자에게 필요한 정보를 추천하는 방법이다. 사용자의 명시적인 기호 정보를 직접적으로 반영하여 다른 사용자의 정보나 평가/행동 이력이 필요하지 않는다는 장점이 있다. 하지만 사용자 선호도와 취향을 특정 단어로 표현하기 어려운 단점이 있다. [2]

또한 소셜 네트워크가 활성화되며 SNS 상에서 해당 사용자와 네트워크를 이루고 있는 관련된 사용자들의 선호도를 이용하여 해당 사용자에게 추천해주는 소셜 기반 추천시스템이 사용되고 있다. 이러한 사용자 기반 추천시스템은 현재 가장 많이 상용화된 시스템이다. 하지만, 최근에는 이러한 커뮤니티 수준에서 추천해주는 사용자 기반 추천시스템에서 더 나아가 개인 수준에서 추천해주는 개인화 추천 시스템을 원하고 있다. 이에 따라 더 정교하게 SVM, LDA 등의 모델을 이용해 잠재된 특징을 분석하는 모델 기반 추천 시스템으로 발전해가고 있다.

모델 기반 필터링(Model-based Collaborative Filtering)은 기존 아이템 간 유사성을 단순하게 비교하는 것에서 벗어나 데이터 안에 내재한 패턴을 이용하는 기법이다. 데이터에 내재되어 있는 패턴/속성을 알아내는 것이 핵심 기술로 LSA(Latent Sematic Analysis), LDA(Latent Dirichlet Allocation)등의 기법을 사용한다.[2]

본 논문에서는 정교화 된 개인화 수준의 추천을 받고 싶어하는 사용자의 니즈를 바탕으로, 멜론 사이트를 통해 얻은 노래의 다양한 메타 데이터를 이용한 콘텐츠 기반 필터링과 메타 데이터 중 음악을 정교하게 구분 지을 수 있을 것이라 여기는 중요한 속성인 노래 가사 데이터를 모델 기반

필터링을 이용하여 정교화 된 개인화 추천 시스템을 구현해보고자 한다.

1.2 분석기법(Method)

R-studio란 R을 사용하기 편하게 만들어 놓은 개발된 환경이다. R은 프로그래밍 언어로, 오픈 소스로 무료이다. 이미 제작된 패키지를 사용하기에 용이하고, 코드를 한 줄 한 줄 읽어가며 명령을 처리하는 인터프리터 언어로 오류파악과 수정이 용이하다는 장점이 있다. 본 시스템에서 사용하는 기법들은 모두 R을 사용한다.

첫 번째 기법으로 'LDA'라는 토픽모델링 기법을 활용해 노래 데이터의 가사를 분석한다. LDA는 가사를 주제별로 구분해준다. 결과는 주제별 확률분포로 나타나는데 이 확률분포가 유사한 노래일수록 비슷한 노래라고 생각할 수 있다. 두 번째부터 네 번째 기법은 사용자 프로파일과 아이템 프로파일의 유사도를 비교하기 위해 유사도를 계산하는 방법이다. 두 번째 기법으론 명목형 변수와 수치형 변수의 유사도를 한번에 혼합하여 계산해주는 'Daisy 함수의 Gower 기법'이 있다. 이는 본 시스템의 대안2로 명목형 변수를 수치형으로 변환하지 않고 한꺼번에 계산하는 방식을 사용한다.

1.3 기대효과(Contribution) 및 논문 방향

현재 개인화 맞춤 음악추천시스템을 목적으로 하는 여러 음악사이트들은 개인 또는 여러 명의 사용자 분석을 중점으로 해서 음악추천을 해주고 있다. 대부분은 협업 필터링 방식의 추천 시스템으로 풍부한 사용자 정보를 바탕으로 추천을 해주는 강점이 있지만, 음악을 내포하는 다양한 속성들을 추천 시스템에 모두 반영하지는 않고있다. 또한 콘텐츠 기반 음악 추천시스템들에 대한 논문 연구들은 살펴보면, 오디오 신호, 감성·감정 모델 등을 분석한 연구는 많았지만 음악을 구성하는 유의미한 텍스트인 가사 자체에 대해 분석한 연구는 이루어지지 않았다. 주제별로 군집화 해주는 토픽모델링 기법 중 LDA는 매우 정확도가 높고 사용이 편리함에도 노래 가사에는 적용사례가 드물다. 이러한 문제들로, 본 시스템이 노래 분석에 중점을 둔 콘텐츠 기반의 개인화 음악추천시스템이라는 점과 새롭게 LDA기법을 도입해 가사를 분석해본 일종의 모델 기반 음악추천시스템이라는 점은 의미가 깊다.

2장에서는 기존 추천시스템의 현황 및 기존 논문 연구를 소개할 것이다. 3장에서는 2장에서 소개한 기존시스템의 한계를 바탕으로 문제를 정의하여 본 논문에서 구축할 시스템의 필요성을 보일 것이다. 4장에서는 본 논문에서 시스템을 구현하기 위해 사용하는 기법들을 설명하며, 5장에서는 본 추천시스템의 프로세스를 정리하고 추천시스템 구현 결과를 나타낸다. 6장에서는

설문을 통해 추천시스템을 검증하고 결과를 통해 한계점을 찾고자 한다. 7장은 본 논문의 결론으로, 추천시스템의 결과와 유의성을 다시한번 나타낼 것이다.

2장. 기존 연구

본 추천시스템 문제 정의에 앞서, 대표적인 3사 음악 추천 사이트와 기존 추천시스템 연구를 분석하고 비교해보았다.

2.1 기존 음악 추천 사이트 3사비교

본 시스템의 목표는 개인화된 추천시스템이라고 하였다. 따라서, 현재 가장 상용화 된 우리나라 스트리밍 음악사이트의 현황은 어떠한 지 알아보기 위해 이 중 대표 3사의 개인화 음악추천시스템을 분석해보았다.

		멜론			지니			벅스
제공 추천 서비스		①멜론 라디오	② For U	③ 나는 지금	① 감성지능 큐레이션	② 마이스타일	③ 비트런	뮤직 4U & 뮤직 PD
사용자 중심	협업 필터링 (=사용자 여러명 분석)	○			○			
	데이터 분석 (=사용자 개인 1명 분석)	○ (빅데이터)	○ (빅데이터)	○ (TPO)		○ (빅데이터)	○ (TPO)	○ (빅데이터)
노래 중심	콘텐츠 기반 (=노래 중심 분석)							○ (키워드)
강점 Strength		유사도 & 정확성 높음			다양한 추천 형태 제공 & 플레이리스트 및 선곡 좋음			검색 엔진 방향 홀륭 (해시태그, 분위기 검색)
좋은점				추천의 기준이 된 노래도 함께 보여줌		선곡에서 원하는 해시태그와의 결합가능		분위기, 느낌 좋음
한계점		10일 이내 발매된 곡들만 추천				추천 알고리즘 보다는 해시태그에 의한 sorting에 의존		노래 자체의 유사도 & 정확성 떨어짐
*TPO: 시간(Time), 장소(Place), 상황(Occasion)								

[표 1. 음악사이트 3사의 개인화 음악추천시스템]

표 1은 지난 2016년부터 시작된 사용자가 음악을 찾아오는 것이 아니라 음악이 사용자를 찾아오는 '뮤직 4.0시대'에 맞춰 새롭게 정비하여 제공되고 있는 3사의 개인화 음악추천 서비스들이다. 현재 3사는 각각 서로 다른 다양한 음악 추천 서비스들을 제공하고 있으며, 저마다 다른 성격의 강점을 가지고 있다. 좋은점과 한계점은 추천서비스들 중에서도 특별히 좋은점이나 한계점이 있는 경우 그 내용을 적은 것이다.

멜론은 사용자 개인의 개인별 선호 아티스트, 선호 장르 등의 음악취향을 분석하고 다른

사용자들과의 유사도를 이용해 음악을 추천해주는 협업 필터링 기반 추천서비스인 ‘멜론 라디오’와 개인의 감상패턴, 선호 장르, 아티스트 취향 등을 빅데이터 분석해 추천해주는 개인 맞춤형 추천 서비스인 ‘For U’, 그리고 사용자 개인이 지금 이 시간에 어떤 노래를 듣고 싶은지 설정한 TPO에 따라 추천해주는 ‘나는 지금’ 서비스를 제공하고 있다. 이 중 대표 서비스는 ‘멜론 라디오’로 추천 음악의 유사도와 정확성이 높지만, 추천범위가 10일이내 발매된 곡이라는 점에서 음원 선택의 폭이 좁아 만족도가 떨어지는 문제가 있다.

지니는 여러 사용자 간의 유사성을 파악해 음악을 추천해주는 큐레이션 서비스인 ‘감성지능 큐레이션’, 사용자 개인의 스트리밍 및 다운로드 이력과 좋아요 클릭 수 등의 빅데이터를 분석해 맞춤 추천을 제공하는 ‘마이스타일’, 그리고 스마트폰 센서를 통해 측정된 사용자 개인의 걸음 속도에 맞는 음악을 추천해주는 TPO 추천 형식의 ‘비트런’ 서비스를 제공하고 있다. 이 중 ‘마이스타일’은 사용자가 원하는 해시태그로 문맥을 설정하고 음악을 추천 받을 수 있어 사용자 만족도가 가장 높았지만, 추천시스템 자체의 알고리즘보다는 사용자들이 직접 선택한 해시태그를 바탕으로 음악을 소팅(sorting)하여 추천해 주는 방식으로 한계가 있다.

벅스의 경우 자체 알고리즘을 이용한 추천 기술인 ‘뮤직4U’와 사용자가 해시태그를 입력하면 뮤직PD들이 해당 키워드에 대해 직접 선정해 놓은 추천 플레이리스트인 ‘뮤직PD’을 결합한 ‘뮤직 4U & 뮤직 PD’ 서비스를 제공하고 있다. 사람이 직접 선정한 추천음악을 함께 반영하기 때문에 추천 음악의 분위기와 느낌의 감성적인 부분에서는 만족도가 높으나, 노래 자체가 가진 다양한 속성들인 아티스트, 장르 등이 반영되지 않아 유사성과 정확성은 떨어지는 경향이 있다.

현재 3사는 크게 사용자 여러 명을 분석한 후 협업 필터링을 이용한 사용자 기반 추천시스템 또는 빅데이터를 이용해 분석한 사용자 개인에 대한 맞춤 추천 또는 사용자 개인의 현재 TPO에 맞는 추천으로 음악 추천을 제공하고 있다. 다수의 사용자가 사이트를 이용하기에 전반적으로 사용자 데이터가 풍부한 3사는 이를 바탕으로 한 사용자에 초점을 둔 음악 추천시스템을 운영하고 있다.

2.2 콘텐츠 기반 음악 추천시스템 연구

앞서 기존의 대표적인 3사 음악사이트에서는 협업 필터링 방식의 추천 시스템으로 풍부한 사용자 정보를 바탕으로 사용자 중심의 추천을 해주는 강점이 있지만, 음악을 내포하는 다양한 속성들을 추천 시스템에 모두 반영하지는 않고있다는 한계가 있다. 본 장에서는 노래의 속성 자체에 중점을 두고 분석한 콘텐츠 기반 음악 추천시스템에 관련된 연구를 알아보았다. 콘텐츠 기반 음악 추천시스템을 연구한 기존 논문들에는 크게 오디오신호 분석, 감정 분석, 가사 분석의

주제가 있었다.

박태수, 정옥란(2015)은 음악의 감정을 분류하기 위한 감정 모델을 만들고, 감정모델에 따라 Emotion Extractor을 통해 음악의 가사와 태그를 이용하여 음악의 감정을 추출해낸다. API를 이용하여 사용자의 최근 재생목록, 음악에 태깅(tagging)된 정보 등을 불러오고, 트위터 API를 통해 게시된 글들을 불러와 Emotion Extractor를 통해 소셜 네트워크에 나타나는 사용자의 현재 감정을 추출하여 음악추천을 하는 방법을 제안하였다.[4]

최홍구, 황인준(2012)은 트위터 무드 분류기를 사용하여 트윗에서 사용자 감정을 찾아내고, 멀티모달(multi-modal) 음악 무드 분류기를 사용하여 주어진 음악DB를 분석하여 사용자 감정에 적합한 음악을 추천한다. 멀티모달 음악 무드 분류를 위해서 음악에서 오디오 특성과 가사 특성 그리고 태그 정보를 활용하였다. 오디오 특성을 얻기 위해 MARSYAS를, 가사분석을 위해서는 음악 무드 태그와 가사를 TF-IDF 알고리즘을 이용한 무드 분류기를 구현하였다.[5]

공민서 (2016)는 사용자에게 개인화된 음악추천서비스를 제공하기위해 내용기반 필터링에 기반하여 음원의 장르와 음악 특성 값을 이용해 사용자의 음악청취이력을 기계학습기법으로 분석하고 사용자의 음악적 성향에 맞는 음악을 추천하는 기법을 제시하였다. 데이터의 수집은 멜론에서 이용하였으며, 수집한 음원을 이용해 메타데이터의 장르태그를 추출하고 음원을 WAV파일로 변환한 후, jAudio를 이용해 음악 특성값을 추출하여 음원 DB를 구축하였다. 사용자들의 선호음악을 조사한 후, 구축한 음원 DB로 기계학습을 적용하기 위한 데이터를 생성하였고 SVM, 베이지안 네트워크, 랜덤 포레스트 기법을 이용하여 데이터를 학습시켜 각각의 음악청취 분류모형을 생성하였다.[6]

이재환 등(2016)은 음악을 선택하는 맥락 중 주요한 요인인 감정을 이용한 노래간 유사도 측정 방법을 제안하여 새로운 추천 시스템에 대한 가능성을 탐색하였다. 노래의 감정 추출에 가사를 이용하였고 가사에서 노래의 구조도 추출해 노래의 의미적 분석을 시도하였다. 실험을 통해 제안한 모델이 기존의 추천 시스템에 비해 작은 계산 복잡성으로 기존 모델과 유사한 성능을 보일 수 있음을 보였다.[7]

이창준 등 (2015)은 감성 어휘뿐만 아니라 가사의 전체 텍스트를 벡터화해

클러스터링하고 사용자의 클러스터 청취 정보를 음악적 성향으로써 반영하는 추천 시스템을 제안하였다. 기존의 음악 추천 시스템들은 가사를 추천에 활용하지 않거나, 가사의 감성 어휘 같은 일부 요소에만 초점을 맞춰 활용했다. 그러나 음악 추천 시스템이 주로 활용되는 대중가요 분야의 곡들은 가사를 포함하고 있고 가사는 유의미한 텍스트이므로 사용자의 만족도에 영향을 줄 수 있다고 보았다. 가사의 텍스트 분석의 한 방법으로 k-means 클러스터링을 활용해 협업 필터링에 적용하였다.[8]

원재용 (2005)은 대표 선율을 이용한 내용 기반 음악 필터링 기법을 제안한다. 대표 선율은 음악을 음높이와 음길이를 이용해 시계열 데이터 형식으로 표현함으로써 선율의 변화 패턴을 요약한 정보이다. 이러한 대표 선율을 이용하여 하나의 대표 객체를 구성하고 구성된 대표 객체들 간에 클러스터링을 수행하기 때문에 유사한 곡들로 그룹화가 가능하여 사용자의 취향과 유사한 곡을 추천해줄 수 있다고 보았다. 사용자 프로파일에서 사용자가 선호하는 장르와 분위기를 분석하여 사용자의 다양한 기호를 추천 결과에 반영하였다.[9]

논문	저자명	오디오신호 분석	감정분석	가사분석
Social Network Based Music Recommendation System, Journal of Internet Computing and Services(JICS) 2015.: 16(6), 133-141	박태수, 정옥란(2015)		음악을 분류하기 위한 감정 모델을 만들고, 감정모델에 따라 음악을 분류하여 소셜 네트워크에 나타나는 사용자의 현재 감정 상태를 추출하여 음악추천	
Emotion-based Music Recommendation System based on Twitter Document Analysis, Journal of KIISE: 18(11), 762-767	최홍구, 황인준(2012)	오디오 특성을 얻기 위해 MARSYAS 시스템을 사용. MARSYAS는 means and variances of SpectralCentroid, Rolloff, Flux, Mel-Frequency Cepstral Coefficients 등의 63개의 스펙트럼 특성을 사용	음악 가사를 분석하기 해 음악 무드 태그와 가사를 TF-IDF 알고리즘을 이용한 무드 분류기를 구현	
A Music Recommendation Scheme using Analysis of User's Music Listening History and Content-Based Filtering	공민서 (2016)	음악의 성질을 분석하기 위하여 음악으로부터 추출된 MFCC(음파의 성질 중 하나)를 사용, 이를 기반으로 음악 간의 유사도를 측정하여 음악 추천		
Similarity Evaluation of Popular Music based on Emotion and Structure of Lyrics,	이재환 등 (2016)	가사에서 놓칠 수 있는 오디오 정보를 보완하기 위해 빠르기, 높낮이, 음계를 구조적 정보로 사용	음악을 선택하는 맥락 중 주요한 요인인 감정을 이용한 노래간 유사도 측정 방법을 제안, 가사에서 감정 추출, 노래의 구조 추출	
Music Recommendation System Using Text Analysis of Lyrics	이창준, 방한별, 이지형 (2015)			가사의 텍스트 분석의 한 방법으로 k-means 클러스터링을 활용해 협업 필터링에 적용
Content-Based Music Filtering Scheme Using Representative Melodies in Music Recommendation Systems	원재용 (2005)	음악을 음높이와 음길이를 이용해 시계열 데이터 형식으로 표현함으로써 선율의 변화 패턴을 요약한 정보인 대표 선율을 이용한 내용 기반 필터링 기법을 제안		

[표 2. 기존 콘텐츠 기반 음악추천시스템 연구]

기존의 연구를 정리하면 표2와 같다. 기존의 연구들은 대체로 음악의 자체 속성을 분석하는데 있어 오디오신호와 감정 분석을 많이 다루었으며, 가사가 가진 텍스트 자체를 분석한 연구는 드문 것을 확인할 수 있었다.

3장. 문제 정의

본 논문의 1장과 2장을 바탕으로, 현재의 추천시스템은 크게 다음과 같은 3가지의 문제를 가지고 있음을 도출하였다.

① 개인화 된 추천 시스템에 대한 사용자 니즈

서론에서 본 바와 같이, 추천시스템의 동향에 따르면 추천시스템은 초기의 단순한 추천시스템으로부터 협업 필터링, 소셜 기반 추천 등으로 발전해 왔고 최근에는 더 정교한 추천을 위해 SVM, LDA 등의 모델을 이용해 잠재된 특징을 분석하는 모델 기반 추천 시스템으로 발전해가고 있다. 이를 통해 사용자들은 조금 더 개인 수준에서 추천해주는 개인 맞춤형 추천 시스템을 원하고 있음을 알 수 있다.

② 기존 대표 음악사이트들에서의 콘텐츠 기반 추천의 부재

기존 음원 추천 사이트인 멜론, 벅스, 지니는 많은 유저들을 가지고 있다. 따라서 협업 필터링 방식을 이용하여 풍부한 사용자데이터를 바탕으로 사용자 중심의 분석을 통해 음악을 추천해주었다. 하지만, 노래가 가지고 있는 다양한 데이터들을 모두 반영하여 객관화 된 노래 간 유사성만을 가지고 추천해주는 콘텐츠기반 추천에는 초점이 덜하였고, 사용자의 누적 데이터를 바탕으로 추천을 하기 때문에, 사용자가 선택한 단일의 곡에 대해서는 추천 서비스가 잘 구축되어 있지 않는다는 단점이 있었다.

③ 음악의 중요 구성요소인 가사에 대한 분석 연구의 부재.

음악은 다양한 메타 데이터를 가지고 있다. 노래 제목, 아티스트, 장르, 발매연도 등 모두 노래의 메타 데이터라고 할 수 있다. 특히 노래 자체의 구성 요소인 노래 가사는 일종의 스토리텔링으로써 단순한 텍스트가 아니라 작곡가가 스토리를 담아 이야기하는 유의미한 텍스트로 객관성, 풍부함과 같은 장점을 갖고 있는 좋은 데이터라고 할 수 있다. 실제로 사람들의 가사 해석 요청을 바탕으로 외국노래의 가사를 해석해주는 사이트들이 현재 다수 존재한다.[10] 가사에 비중을 두고 음악을 감상하는 사용자들이 많이 존재하는 것이다.

사용자는 음악을 선택할 때 기분과, 상황에 영향을 받고 이에 공감이 가는 음악을 선택한다. 예를 들어, 음악이 '슬픔' 이라는 같은 감정 카테고리에 속하더라도 그 스토리는 '이별', '삶의 버거움', '부모님' 등 다양한 주제로 각기 다르고, 이는 감정만으로는 추출해 낼 수 없는 정보가 가사에 존재한다는 의미이다. 혹은 아이러니하게도 멜로디가 경쾌해도, 내용은 비관적이거나, 슬픔을 경쾌한 멜로디로 승화한 노래의 경우도 존재한다. 가사 속 감정이나

오디오 신호 만으로 노래를 분류하기에는 분명한 한계가 존재하는 것이다.

하지만, 앞서 살펴본 기존 연구들은 오디오 신호를 바탕으로 노래를 분석하거나, 가사의 감성 어휘에만 초점을 둔 감정분석만을 하였고, 가사의 감성만을 추출하여 가사를 구성하고 있는 또다른 유의미한 텍스트 정보는 이용하지 않았음을 확인할 수 있었다.

이를 바탕으로 본 논문에서 제안하는 추천시스템은 사용자의 최근 니즈를 바탕으로 개인화 추천시스템을 목표로 하며, 노래가 가지고 있는 메타 데이터(장르, 분위기, 계절, 발매연도)와 같은 다양한 변수들을 수치적으로 정량화하여 추천에 반영할 것이다. 특히, 음악을 구체적으로 반영하는 중요 요소인 가사를 정교화 된 토픽모델링 기법인 LDA을 사용하여 분석해 추천에 이용할 것이다. 즉, 노래 자체의 콘텐츠에 중점을 두어 이를 최대한 활용하여 본 추천시스템에 반영함으로써 시스템의 분류 성능을 높이고, 사용자의 취향을 다양하게 반영할 수 있도록 한다.

4장. 사용 기법

본 추천시스템에서는 'LDA'와 'Daisy함수의 Gower기법'을 주요 사용기법으로 한다. 각 기법에 대해 살펴보자.

4.1 LDA (Latent Dirichlet Allocation)

텍스트마이닝 기법에는 데이터를 어떤 용도로 분석할 것인가에 따라 다음과 같이 구분할 수 있다. 문서를 주제에 따라 나누는 군집화(clustering) 및 토픽모델링(topic modeling), 문서에서 원하는 정보를 검색하는 정보 검색(information retrieval), 문서로부터 필요한 정보를 얻는 정보 추출(information extraction)이 있고 이외에도 작성자의 감정, 감성, 심리 등을 파악하는 감성 분석(sentiment analysis) 등의 여러 분야가 있다. [11]

이 중에서 가사를 분석하는 용도로는 주제에 따라 문서를 나누는 군집화/토픽 모델링이 적당하나 기존 군집분석은 개별 문서가 하나의 주제에만 해당된다고 가정하는 한계점이 있었다. 이에 비해 토픽분석(토픽모델링)의 LDA는 하나의 문서가 여러가지 주제를 다룰 수 있음을 가정하고 이를 확률로 나타낼 수 때문에 더 섬세한 군집화가 가능하다.

Blei et al.(2003)의 LDA기법은 토픽모델링의 가장 대표적인 기법이다. 이는 주로 신문기사, 논문, 각종 SNS 게시물 등 문서를 주제별로 나눠야 할 때 사용한다. LDA는 여러 문서에서 잠재적으로 의미 있는 토픽(주제)을 찾아주는 확률 분포 모델이다. 하나의 개별 문서가 여러 개의 주제를 가지고 있음을 가정하고 각 문서의 단어들이 자주 출현하는 빈도를 이용해 각

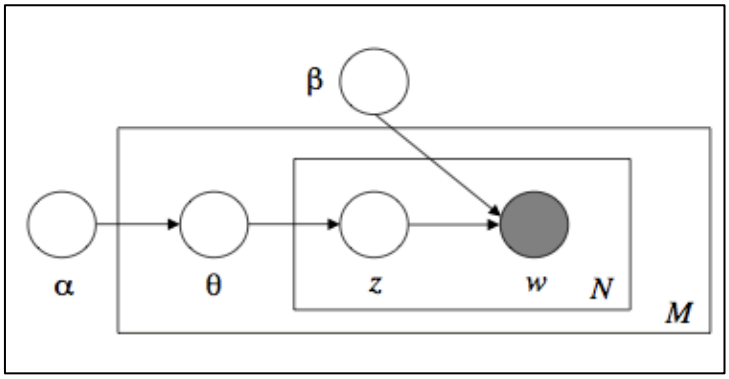
주제에 대한 분포를 확률로 나타낸다.

LDA의 생성과정은 다음과 같다. 먼저 K개의 토픽의 분포를 추출하고 D개의 문서에서 나타나는 개별 토픽의 비율 분포를 추출한다. 그 후 선택된 토픽비율로부터 토픽을 추출하고 개별 토픽으로부터 단어를 추출한다. 그림2는 LDA의 학습 과정을 보여준다.[13]

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

[그림2. LDA 생성과정]

K는 토픽의 수, N과 Θ는 문서 단위, α와 β는 코퍼스(말뭉치) 단위로 정해지는 값이다. α는 k차원 디리클레(Dirichlet) 분포의 매개변수다. N은 문서의 길이고, Θ는 해당 문서에서 각 주제의 가중치를 나타내며 각 엔트리 값을 합치면 1이 된다. β는 각 주제별로 특정 단어가 생성될 확률이 그려진 2차원 테이블이고 즉 β_{ij}는 i번째 주제가 단어집의 j번째 단어를 생성할 확률이다. Z_i는 문서의 i번째 단어에 대한 주제 벡터이며, 하나의 엔트리만 1이고 나머지는 0이다. 이 과정에서 주제의 개수는 k로 고정이고, Θ와 Z_i는 길이가 k인 벡터가 된다.



[그림3. LDA 모델링 과정]

이와 같이 특정 문서에서 주제벡터인 Θ가 있고 앞에서부터 단어를 채울 때마다 Θ로부터 하나의 주제를 선택하고 다시 그 주제로부터 단어를 선택하는 과정을 거친다.

쉽게 예를 들면 100명이 서점에 가서 장바구니에 여러 개의 책을 고른다고 했을 때 각자의 장바구니는 문서가 되고 장바구니에 들어간 책들은 단어가 된다. 100명의 장바구니를 학습시키면 자주 같이 등장하는 책들끼리 K개의 군집이 생긴다. 이때 다시 장바구니들을 학습시켜서 각자의 책들이 각 군집에 포함되는 확률을 구한다. 따라서 결과는 각 100명의 장바구니의 확률분포로 나오고 이 분포가 유사할수록 선호하는 책이 비슷한 장바구니라고 할 수 있다.

본 알고리즘에서는 노래 가사를 LDA를 통해 학습시켜 K개의 군집으로 묶는다. LDA는 가사의 단어들이 각 군집에 포함되는 비율을 수치로 나타내준다. 즉 결과로 나타나는 확률 분포가 유사한 것은 가사끼리의 유사도가 높은 노래가 된다.

4.2 Daisy함수의 Gower기법

유클리드 거리, 피어슨 상관계수, 코사인 유사도 등 유사도를 계산해주는 식은 많지만 이들은 수치형 데이터를 비교할 때 쓰는 식이므로 수치형과 범주형 변수가 혼합된 데이터에는 사용할 수 없다. Daisy함수의 Gower 유사도 계수는 혼합형 변수의 거리를 측정하여 유사도를 비교해준다.[13] 아래는 Gower유사도의 계산식이다.

$$d_{ij} = d(i,j) = \frac{\sum_{k=1}^{p} w\_kdelta(ij;k)d(ij,k)}{\sum_{k=1}^{p} w\_kdelta(ij;k)}$$

d (ij, k) 는 총 거리에 대한 k 번째 변수 기여도이며 x [i, k]와 x [j, k] 사이의 거리이다.

d_ij는 d (ij, k)의 가중 평균이고, d (ij, k)는 w_k delta (ij, k)를 가중치로 갖는다. 이 때 w_k는 weigths [k]고 delta (ij, k)는 0또는 1 이다.

본 알고리즘에선 대안 2 로 Gower 유사도를 이용하여 수치형과 범주형의 혼합형 변수의 유사도를 비교한다.

5장. 문제해결 프로세스

본 연구에서는 Raw Data 수집 및 전처리를 시작으로 R 을 이용한 가사 텍스트마이닝 전처리 단계, 기준노래와의 유사도를 계산하는 시스템 구현, 그리고 검증 순서로 프로세스가 진행된다.

5.1 Raw Data 수집 및 전처리

사람들이 가장 많이 이용하고 음악 수가 많은 멜론으로부터 ID, 제목, 아티스트, 장르, 좋아요 수, 가사, 분위기, 계절과 발매연도까지 9가지의 노래 속성을 가진 3만여개의 노래데이터를 얻었다. 데이터 전처리를 통해 결측치가 존재하거나 가사가 너무 짧은 곡, 좋아요 수가 100개 이하인 곡, 또 텍스트마이닝을 하기 위해 외국노래까지 제거하여 34345개의 곡이 남았다.

이 중 데이터를 분석했을 때 의미 있을 속성을 추려냈다. 고유한 숫자인 ID와 의미가 없는 단어가 많고 영어가 많은 제목은 분석 대상에서 제외했고 좋아요 수는 대중성을 나타내므로 우리의 목적인 개인화 추천과 맞지 않다고 생각하여 제외했다. 또 아티스트는 특성 별로 분류하기가 힘들고 개인이 좋아하는 아티스트가 있으면 검색해서 듣는 경우가 많으므로 제외하였다. 그 결과 장르, 계절, 발매연도, 가사, 분위기의 5가지 속성이 남았다. 장르는 힙합, R&B, 발라드 등으로 나눠져 있고 계절은 봄 여름 가을 겨울로, 발매연도는 1990년대 이전, 1990년대, 2000년대, 2010년대로 4개의 범주로 나눠져 있다. 분위기는 카페, 이별, 사랑 등 19개의 column을 가지고 있고 노래마다 각 분위기에 해당하는 확률 분포로 나타나있다. 가사는 텍스트마이닝을 활용하여 분위기와 같은 확률분포, 즉 수치형 변수로 만든다.

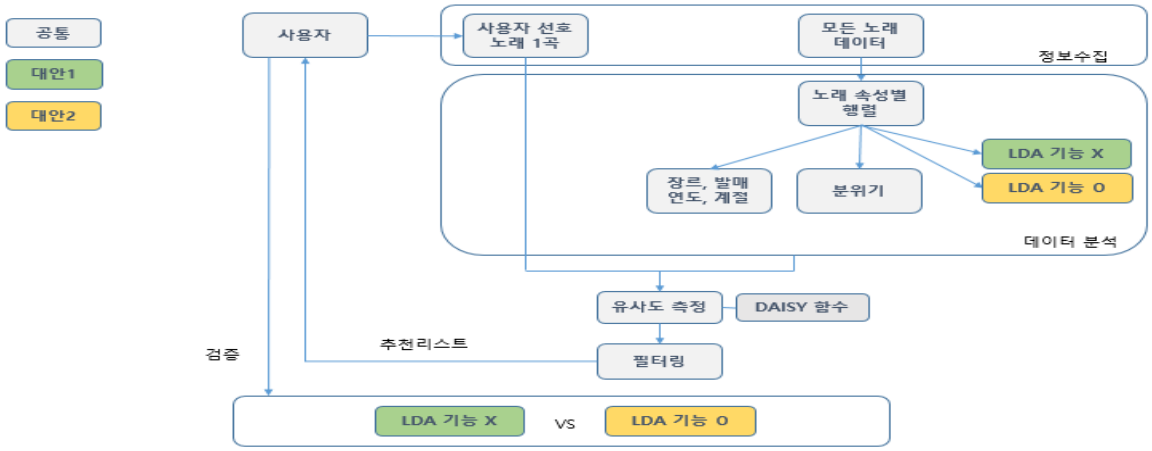
다음 알고리즘에선 전처리 된 노래의 범주형 데이터인 장르, 계절, 발매연도와 수치형 데이터인 분위기, 가사의 속성을 여러 기법을 통해 분석한다. 그런데 컴퓨팅문제가 있어 추천시스템에 입력될 노래 수를 줄이는 과정을 거쳐야했다. 시스템에 쓰일 노래 수는 재현율을 고려해 정하였다. 먼저 좋아요 30000 개 이상인 노래 994 개로 추천해본 결과, 대중성이 높은 노래들만 추천되어 10 곡 모두 사용자가 아는 노래가 나와 검증이 무의미할 것으로 예상됐다. 다음으로 좋아요 7000 개 이상으로 필터링하여 5383 개의 노래로 추천해본 결과, 반대로 사용자가 아는 노래가 거의 없어 사용자가 원래 좋아하던 곡 수가 적어 재현율이 매우 낮게 나오고, 이는 무의미한 검증으로 이어질 것으로 예상되었다. 따라서 좋아요 수 17800 개 이상인 2000 개의 노래로 추천시스템에 입력될 노래 수를 최종 선정하였다.

5.2 추천 시스템 프로세스

본 추천시스템의 프로세스는 다음과 같다. 먼저, 텍스트마이닝 기법 LDA를 통해 가사를 군집화하여 수치화한다. 가사들을 학습시키면 가사의 단어들이 학습되어 유사한 단어끼리 묶인 k개의 단어군집이 생성되고 다시 각 가사들의 단어가 각 군집에 얼마나 들어가 있는지를 수치로 나타내준다. 따라서 3만여개의 가사를 학습시켜 한 노래에 자주 같이 등장하고 유사한 단어끼리 모인 8개의 군집이 나왔다고 했을 때, A노래의 첫 번째 단어군집에 해당하는 단어들이 몇 개 있는지, 두 번째 단어군집의 단어들은 몇 개 있는지 등 8개 단어 군집에 대해 다시 학습시켜 A노래 가사에 대한 결과는 0.4/0/0/0/0.1/0.3/0.2/0 와 같은 식의 확률 분포로 나타난다. 이 분포가 비슷할수록 가사의 유사도가 높은 노래가 된다.

장르, 발매연도, 계절을 그대로 범주형 데이터로 둔 후 daisy함수의 Gower기법을 사용하여 혼합형 데이터의 유사도를 측정하여 사용자가 선호한 노래와 가장 유사한 노래 Top N개를 추출한다.

최종검증은 가사분석 기능의 타당성을 검증하기 위해 대안을 둘로 나눠 대안1은 LDA학습시킨 모델, 대안2는 LDA학습시키지 않은 모델로 나누어 추후 정확도와 재현율을 지표로 결과를 비교 검증해보았다. 이는 6장에서 다룬다. 아래는 이를 그림으로 나타낸 추천시스템 모델 흐름도이다.



[그림4. 추천 시스템 모델 흐름도]

5.3 추천시스템 구현 및 결과

R을 이용해 노래 2000개를 입력 받아 가사 분석 텍스트마이닝을 거친 '가사 군집', '테마'의 수치형 속성과 '장르', '발매연도', '계절'의 범주형 속성의 유사도를 계산해 유사도가 높은 상위 N개의 노래를 추천 결과로 출력하는 시스템을 구현해보았다.

먼저, 보다 정확도를 높이기 위해 LDA패키지를 이용해 가사를 텍스트마이닝 전처리하였다. 구두점 제거, 숫자제거, 띄어쓰기를 한 번으로 만들기, 명사추출의 과정을 거쳐 가사의 토픽(주제)로 선정될 어휘로 의미 없는 어휘가 선정되지 않게 전처리 과정을 거쳤다. 이후에는 'lexicalize()'라는 함수를 이용해, 앞서 언급한 4.1의 'LDA의 학습과정'을 거친 corpus를 생성하였다.

아래는 k = 8으로 설정해 8개의 군집기준으로 corpus를 생성한 결과이다. 그리고 이어서 8 군집별 빈도수를 계산하고, 이로부터 8개의 각 군집에 속하는 확률을 구했다.

1	그대	좋아	으르렁	비가	너를	이건	바래	너를
2	그댈	내가	보다	바람이	너와	우린	라라라	니가
3	그대가	너무	울었다	시선	싫어	오늘	슬로이럽	다시
4	사랑	니가	물러	너의	순간	떨어	아래	혼자
5	말아요	예뻐	여자	작은	너의	노력	위로	네가
6	그대를	물라	아니야	입술	없어	연달고리	내가	이형계
7	그댄	자꾸	랄애	속에	나의	안의	대로	없어
8	그대는	그래	진짜	아름다운	나를	그날	무대	내가
9	그대의	정말	해업	곳에	모든	누구	없어	정말
10	사람	남자	큰데	눈이	내게	때문에	그래	눈물이
11	사랑해	싫어	다운	어느새	네게	소리	위해	너무
12	나의	너는	이게	내리는	오직	최고	보여줄게	아파
13	그대여	너의	보단	여전히	거야	발에	싫의	동아와
14	눈물이	분이야	너의	크리스마스	함께	랄리	순응	우린
15	나를	안돼	호응	너를	눈응	나와	꼴이	보고
16	봐요	너만	그녀는	날아	네가	우리	마안해	이젠
17	그대와	그런	그만	봄이	너만	제미없어	거야	사랑
18	마요	달라	미친	우산	꼴응	통통	꼴은	아직도
19	언제나	그날	망할	별이	수가	없어	외쳐	너의
20	이형계	이런	슈가	없이	있게	왔어	년	눈물
21	사랑아	뭐야	편히	차가운	있다면	홀려	꼴을	수가
22	다시	보여	슈가프리	그대가	있어	점핑	라라	좋고
23	있죠	어떡해	아니	닮은	결에	홀려가	다응	없는
24	결에	네가	진	바람에	준응	난리	나랄	너와
25	사랑이	기분	랩은	밤이	내가	미친	겁이	사랑을
26	사랑을	좋은	등지마	없는	이미	관해	가드	사람
27	마음	이형계	그녀	우리의	지금	가요	예아호	내게
28	사랑은	다른	뷔	내려	다시	춤응	등고	이별
29	그대만	말아	사랑이었다	눈물	같은	친구	맞게	있응까
30	사랑해요	말해	제이홍	뭐예요	너에게	리듬	위에	너에게
31	그대에게	하지마	반해	오는	향해	떡볶이	때까지	나를
32	있나요	요즘	내가	함께	빛나는	째끈해	걱정은	너만
33	우리	말고	하드케리해	기억	매일	너뭔	이젠	그런
34	영원히	아예	너는	않는	속에	완전	지금의	그땐
35	아픈	싫어	너가	그날	다가와	전화해	먹어	안녕
36	멀리	보면	해줄라고	날을	눈이	너와	뒤로	미안해
37	이젠	맘을	거기	남은	이대로	마마	울리고	마지막
38	너무	여자	지금	드는	없는	우아	라랑라	너도
39	가슴이	없어	좀지	추억이	맘을	커지는	하기나	같아
40	있어요	장난	예뻐서	할가가	이형계	한말	영원히	거짓말

[그림 5. LDA학습을 통해 생성된 Corpus 결과]

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
[1,]	7	22	13	3	63	0	46	44
[2,]	33	29	27	5	48	13	31	27
[3,]	8	28	6	0	32	0	42	24
[4,]	16	39	44	6	62	4	36	43
[5,]	16	28	40	6	68	2	48	54
[6,]	26	40	16	7	70	5	46	76
[7,]	32	33	46	5	83	8	77	55
[8,]	20	31	29	17	56	4	34	41
[9,]	56	54	38	25	87	2	78	93
[10,]	31	55	29	13	114	11	73	66
[11,]	13	48	61	13	118	3	78	68
[12,]	18	43	61	9	136	6	83	78
[13,]	10	27	37	19	83	3	49	49
[14,]	74	59	33	20	59	28	84	42
[15,]	17	30	36	9	56	1	42	42
[16,]	25	49	55	12	115	4	78	66
[17,]	18	52	41	18	107	2	75	63
[18,]	14	40	136	8	82	0	59	48
[19,]	42	61	57	17	126	7	59	97
[20,]	12	48	84	9	117	32	89	89
[21,]	40	63	59	13	141	10	102	88
[22,]	14	17	79	10	112	4	57	71
[23,]	24	37	39	15	112	2	49	81
[24,]	14	44	60	31	102	2	64	55
[25,]	34	71	52	34	97	13	58	64
[26,]	116	85	30	8	45	14	85	25
[27,]	21	42	74	16	136	8	81	64
[28,]	37	75	66	26	155	15	102	101
[29,]	79	88	50	21	102	25	95	74
[30,]	17	28	86	10	143	5	101	97
[31,]	21	53	56	9	126	5	96	65
[32,]	20	64	40	7	152	10	81	94
[33,]	6	37	60	8	101	7	71	64
[34,]	39	72	72	34	171	8	84	96

[그림 6. 군집 별 빈도수 계산]

```
> divide.num <- apply(mat.result,1,sum)
> divide.num[which(divide.num==0)] <- 1
> mat.final <- mat.result[1:dim(mat.result)[1],]/divide.num[1:dim(mat.result)[1]]
> mat.final[1,]
 [1] 0.02641509 0.08301887 0.04905660 0.01132075 0.23773585 0.00000000 0.17358491 0.16603774
> mat.final[3,]
 [1] 0.04395604 0.15384615 0.03296703 0.00000000 0.17582418 0.00000000 0.23076923 0.13186813
> mat.final[100,]
 [1] 0.08830846 0.12437811 0.12810945 0.03855721 0.20398010 0.02736318 0.13681592 0.15547264
```

[그림 7. 각 군집에 속하는 확률 예시]

다음으로 앞서 4.2에서 명목형 변수와 수치형 변수를 함께 가지고 있는 데이터의 특성상 이 두가지를 함께 고려한 유사도를 계산해내기 위해 Daisy함수의 Gower 기법을 사용하겠다고 하였다. 이를 위해 Cluster 패키지의 Daisy함수을 이용해 실제 유사도를 계산해냈다. 아래는 장범준의 '사랑에 빠졌죠'를 예로 들어, 다른 노래들과의 유사도 거리를 계산한 예시이다.

```
> #ex)장범준-사랑에 빠졌죠 ('
> calc.dist(8104509)
          distance
1  3753304 0.09858718
2  4543502 0.13711581
3  9631530 0.20862921
4  5719286 0.27285966
5  7861392 0.18918047
6  5398990 0.21182866
7  3910821 0.16737936
8  9620469 0.19237420
```

[그림 8. Daisy함수를 이용한 유사도 거리계산 예시]

이어서 유사한 노래를 추천 받고자 하는 노래의 'songid'와 추천 받고 싶은 노래의 개수 'N'을 입력하면 가장 가까운 노래를 추천해주는 함수 'recommender'을 만들었다. recommender함수를 거쳐 추천되는 유사노래리스트가 곧 추천시스템의 출력결과이다. 출력결과 중 가장 유사한 노래는 자기 자신이므로 TOP 10개의 노래를 알고자 할 때는 11개의 노래를 출력하여 첫 번째로 나온 자기 자신 songid는 제외한다.

아래는 recommender 함수를 이용해 기준 노래를 입력했을 시 기준노래와 유사한 노래로 추천되는 리스트를 구해본 예이다. 예를 들어 아이유 '좋은 날'의 'songid' 3051244와 본 노래를 포함하여 출력될 노래 수인 11을 입력하면 'songid'가 3053259, 1944399, 4027904, 2566458, 2537575, 370133, 429656, 672010, 732373, 1568562 값으로 출력되고 사용자에겐 첫 번째 songid를 제외한 나머지 10개의 노래를 추천해준다. 그리고 앞서 Daisy함수의 Gower기법을 이용해 계산된 유사로부터 기준 노래와 유사도가 높은 순서로 추천되는 'Top N' 방식으로, 1번째로 추천되는 노래일수록 기준 노래와 유사도가 높으며 N번째로 갈수록 유사도가 점점 낮아진다.

```
> #아이유 - 좋은 날 3051244 Ballad
> recommender(3051244,11)
  18401    1982    1734    1777    1814    1699    1970    1646    1828    1721    1731
3051244 3053259 1944399 4027904 2566458 2537575 3701333  429656  672010  732373 1568562
> #522852   너에게 쓰는 편지 (Feat. 린)   MC 몽   Rap/Hip-hop
> recommender(522852,11)
   2791     153     213     506     338     456    1976     358     471     109     437
 522852 2109953 1637914  313002 2612772 1114595 2596161 1556553 2733249 1177472 2139251
> #Ballerino-리쌍
> recommender(1637914,11)
   2131     506     456     279     338     225     387     457     256     530     358
1637914  313002 1114595  522852 2612772  971441 1932604  496377 1850253 1583465 1556553
```

[그림 9. recommender함수 적용 예시]

이렇게 구현한 추천시스템을 이어서 6장에서 설문조사를 통해 시스템의 성능을 검증해본다.

6장. 추천시스템 검증

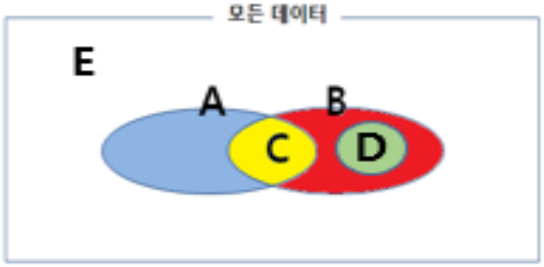
5.2에서 정의한 대안1과 대안2를 비교검증하기 위해, 블라인드 설문조사를 받아 정확도, 재현율, 그리고 F-Measure의 성능지표를 측정해보았다. 이후 F-Measure값에 대해 t-검정: 쌍체 비교를 실시하여 유의수준 α=0.05에서의 유의성을 살펴보았다.

6.1 추천시스템 모델 검증방법

본 추천시스템을 검증하기 위해 사용한 성능지표는 정확도(Precision), 재현율(Recall)을 이용한 F-Measure다. 추천목록에 대한 선호도로만 시스템을 검증하면 결과가 주관적이기 때문에 객관성이 부족하다. 따라서 정확도와 trade-off관계인 재현율을 구하여 정확도와 재현율의 조화평균인 F-Measure값을 최종지표로 사용한다. 각 측정 지표 식은 다음과 같다.

- 정확도(Precision) = 사용자가 실제 선택한 아이템의 수/전체 추천된 아이템의 수
- 재현율(Recall) = 맞는 추천 아이템 수/사용자가 선택한 전체 아이템 수

검증에 쓰일 정확도와 재현율을 본 시스템에 맞게 수정하여 사용하였다. 예를 들어 정확도는 10곡을 추천해 주었을 때 사용자가 그 중에 맘에 들어 한 노래의 비율로 계산한다. 재현율은 사용자가 마음에 들어 한 노래 중 원래 좋아하던 노래 비율로 계산한다. 재현율을 높이기 위해 아이템 수를 늘리면 정확도가 낮아지고 정확도를 높이기 위해 추천하는 아이템 수를 줄이면 재현율이 낮아진다. Trade-off관계인 정확도와 재현율을 고려한 F-Measure측정치를 최종 검증에 사용한다.[3] 수정된 정확도, 재현율과 F-Measure의 공식은 다음과 같다.



A : 사용자가 원래 좋아하던 노래
B : 추천된 10 개 노래
C : 사용자가 원래 좋아했던 노래 중 추천된 노래
D : 추천된 노래 중 새로 알게 된 좋은 노래
E : 2000 개의 모든 노래

[그림 10. 정확도와 재현도의 관계]

- 정확도(Precision) = $\frac{C+D}{B} * 100 = \frac{\text{추천 노래 중 사용자가 마음에 든 아이템의 수}}{\text{추천된 아이템의 수}} * 100$
- 재현율(Recall) = $\frac{C}{C+D} * 100 = \frac{\text{추천 노래 중 사용자가 원래 좋아했던 노래}}{\text{추천 노래 중 사용자 마음에 든 아이템의 수}} * 100$
- F-Measure = (2 x 정확도 x 재현율)/(정확도+재현율)

F-Measure을 구하기 위한 설문으로 30명의 설문 대상자가 좋아하는 노래 1곡을 선택하면 그룹1엔 LDA가 포함된 추천시스템을 사용하여 추천한 10곡, 그룹2엔 LDA가 포함되지 않은 추천시스템을 사용하여 추천한 10곡 노래를 추천해주었고 각 그룹에 대한 정보는 블라인드다. 추천된 각 10개의 노래 중에 마음에 드는 그 중 n개와 원래 좋아했던 m개 노래 수를 조사한 후 정확도, 재현율을 구하여 그룹1과 그룹2의 F-Measure 차이에 대한 t-검정: 쌍체 비교를 실시하였다. t-검정: 쌍체 비교란 대응되는 두 표본집단의 평균을 비교하는 것으로 주로 동일한 n명에 대해 실험 전 후 값을 비교할 때 사용 한다.[14]

6.2 추천시스템 모델 검증결과

	그룹 1			그룹 2		
응답자	정확도	재현율	F-measure	정확도	재현율	F-measure
1	0.700	0.571	0.629	0.800	0.500	0.615
2	0.700	0.714	0.707	0.900	0.333	0.486
3	0.700	0.714	0.707	0.800	0.500	0.615
...	...	...	...	...	...	...
28	0.800	0.750	0.774	0.900	0.778	0.834
29	0.700	0.571	0.629	0.500	0.800	0.615
30	0.800	0.875	0.836	0.900	0.556	0.687
	평균		68.93%	평균		65.29%

[표 2. 1차 설문 결과 - 정확도, 재현율, F-Measure]

위 결과는 설문을 통해 30명의 정확도, 재현율, F-Measure을 구한 값이다. 그룹1은 LDA 기능을 포함, 그룹2는 LDA기능을 포함하지 않은 결과이다. LDA의 타당성을 검증하기 위해, 즉 그룹1의 F-Measure값이 그룹2보다 크다는 것을 검증하기 위해 그룹1의 F-Measure평균이 그룹2보다 작거나 같다는 귀무가설과 그룹1의 값이 더 크다는 대립가설을 세워 유의수준 α=0.05에서의 30명에 대해 각 그룹의 F-Measure 비교를 위한 단측 검정을 실시하였다.

	F-Measure		
응답자	**그룹 1**	**그룹2**	**d(그룹 1-그룹 2)**
1	0.629	0.615	0.014
2	0.707	0.486	0.221
3	0.707	0.615	0.092
···	···	···	···
27	0.771	0.750	0.021
28	0.774	0.834	-0.060
29	0.629	0.615	0.014
30	0.836	0.687	0.149
평균 $\bar{d}$			**0.036**
표준편차 σ			**0.088**

[표 3. F-Measure 차의 평균, 표준편차]

t-검정: 쌍체 비교

	변수 1	변수 2
평균	0.6892743	0.65290316
분산	0.0339058	0.03474008
관측수	30	30
피어슨 상관 계수	0.8868467	
가설 평균차	0	
자유도	29	
t 통계량	2.2597037	
P(T<=t) 단측 검정	0.0157631	
t 기각치 단측 검정	1.699127	
P(T<=t) 양측 검정	0.0315261	
t 기각치 양측 검정	2.0452296	

[표 4. F-Measure에 대한 t-검정 결과]

t-검정 결과, p-value=0.0158가 유의수준 0.05보다 작으므로 귀무가설을 기각하고 대립가설을 채택한다. 따라서 유의수준 α=0.05에서 그룹1의 F-Measure평균이 그룹2보다 크다고 할 수 있다. 즉, LDA기능을 포함한 시스템의 추천시스템 성능이 더 좋다고 할 수 있다.

6.3 추천시스템 모델 한계

본 시스템의 문제점을 파악하기 위해 그룹1보다 그룹2의 결과가 높은 대상자 5명에게 본 추천시스템의 목적을 알리고 피드백을 받는 설문을 진행하였다.

[그룹2의 결과가 높은 대상자]
대상자1 “가사는 중요하지 않고 멜로디를 중점적으로 노래 듣는다.”
대상자2 “가사보단 해당 가수 특유의 감성을 좋아해서 노래를 듣는다.”
대상자3 “그룹2의 노래가 고른 노래의 분위기와 더 비슷했다. 가사는 보지 않는다.”
대상자4 “처음 들었을 때의 느낌이 중요한 것 같다. 노래의 느낌이 좋으면 그 다음에 가사를 본다.”
대상자5 “가사를 중요하게 생각해서 그런지 그룹1의 노래가 전부 마음에 들었지만 원래 알던 노래가 별로 없었다.”

피드백 결과, 5명중 4명은 가사보다 분위기, 가수 등 다른 요소를 중요하게 생각한다는 것을 알 수 있었다. 따라서 가사를 중요시하면서도 그룹1의 값이 그룹2보다 낮게 나온 나머지 1명을 대상으로, 가사를 보면서 다시 노래를 듣고 그룹1에서 마음에 드는 곡 수를 다시 묻는 2차 설문을 실시할 예정이었다. 이를 1차 설문과 비교하여 그룹1의 마음에 드는 곡 수 변화가 있는지 결과를 수치로 보이려 했지만 이미 그룹1에 마음에 드는 곡이 10곡이었기 때문에 더 이상 나은 결과를 기대할 수 없었고 대상자가 원래 좋아하던 노래를 추천해주지 못하여 재현율이 낮다는 한계점으로 남았다. 하지만 재현율은 시스템 검증 목적으로 활용한 값이고 상대적으로 정확도는 높은 값을 가지기 때문에 실제로 추천시스템을 사용할 때 사용자의 만족도는 더 높을 것으로 기대된다.

또한, 데이터 전처리 과정에서 한국어 노래가사를 더 정확하게 분석하기 위해 가사 속의 영어단어를 제거한 점과 외국 노래, 비주류 장르와 대중성이 매우 낮은 노래를 포함하지 못한 점으로 인해 완성도 있는 추천 시스템을 구현하지 못했다는 한계가 있다. 해당 문제는 컴퓨팅문제를 해결하고 데이터베이스의 규모를 늘려, 다양한 사람들의 노래 취향을 반영하게 된다면 본 추천시스템의 완성도는 더 높아질 것으로 기대할 수 있다.

6.4 추천 시스템 모델 결과해석

연구의 유의성을 판단하기 위하여 본 알고리즘을 통해 나온 결과가 신뢰할 만 한 결과인지 시각화해 보았다. 다음은 랜덤으로 선택한 노래를 선택하고 본 알고리즘을 통해 추천 결과로 나온 대안1과 대안2의 top 10개의 노래 가사를 각각 분석 및 비교해 본 결과이다.

기준 노래는 <<강승윤, 비가 온다, Rock, song id=4175139>>로 가사는 다음과 같다. 가사에서 각기 다른 색으로 표시한 부분은 뒤에 나올 대안1과 대안2의 가사내용과 비교할 시, 이해를 돕기 위해 유사한 스토리가 담긴 부분을 각각의 색으로 구분해 놓은 것이다. 본 가사의 전반적인 스토리는 비 바람 부는 날 헤어진 옛 연인을 그리워하는 내용이다.

< 4175139, Rock, 강승윤 – 비가온다 가사>

창문너머엔 슬픈 비가 내리고 문 뒤로 부는 바람은 창가의 눈물을 훔쳐내고 널 위해 준비했던 선물들 고백이 담긴 편지는 갈 곳 없이 먼지만 품고 있어 누가 내 맘을 알아주려나 누가 내 말을 들어주려나 날 녹여주던 그 손길도 부드러운 목소리도 이젠 내 것이 아니라는 게 오늘따라 더 힘들다 맘에 비가 온다 비가 온다 차갑게 맘에 비가 온다 비가 온다 오늘도 방안에 어두운 그림자가 내리고 달빛에 비친 시계는 헤어지던 날에 멈춰있고 널 위해 살아왔던 많은 시간들 너를 안던 가슴은 주인 없이 바람만 품고 있어 누가 내 맘을 알아주려나 누가 내 말을 들어주려나 내 거칠어진 입술마저 차가워진 가슴마저 여전히 너만 찾고 있어서 오늘따라 더 그립다 맘에 비가 온다 비가 온다 차갑게 맘에 비가 온다 비가 온다 오늘도 이젠 볼 수 없다는 걸 난 잘 알고 있는데 다시 여기에 올 것만 같애 비바람이 몰아치는데 피할 곳이 없어 가슴이 언다 그만 맘에 비가 온다 비가 온다 차갑게 맘에 비가 온다 비가 온다 오늘도 널 위해 준비했던 선물들 고백이 담긴 편지는 갈 곳 없이 먼지만 품고 있어

⇨ **비 바람 부는 날 헤어진 연인을 그리워하는 내용**

기준 노래를 바탕으로 본 추천시스템을 통해 나온 두 대안의 결과는 다음과 같다.

LDA 기능 O

Song-Id	Title	Artist	Genre
3832098	그댈 마주하는건 힘들어	버스커 버스커	Rock
4711065	우산	윤하 (YOUNHA)	Ballad
5760115	IF YOU	BIGBANG	Folk/Blues
4677133	이젠 아니야	비스트	Ballad
3832101	정말로 사랑한다면	버스커 버스커	Rock
9614761	그대라는 사치	한동근	Ballad

LDA 기능 X

Song-Id	Title	Artist	Genre
3832098	그댈 마주하는건 힘들어	버스커 버스커	Rock
4711065	우산	윤하 (YOUNHA)	Ballad
4144408	All Right	김예림(투개월)	Rock
5760115	IF YOU	BIGBANG	Folk/Blues
3893169	홀로 (Feat. 김나영)	정키	Ballad
4709050	가끔 내가	김나영	Ballad

4824737	Rain	방탄소년단	Rap/Hip-hop
4806686	공허해	WINNER	Ballad
3834908	한사람	허각	OST

4677133	이젠 아니야	비스트	Ballad
3832101	정말로 사랑한다면	버스커 버스커	Rock
4824737	Rain	방탄소년단	Rap/Hip-hop

대안1과 대안2의 Top1과 Top2의 노래는 <<3831098, 버스커버스커, 그댈 마주하는 건 힘들어, Rock, 3831098>>와 <<4711065, 윤하, 우산, Ballad>>로 동일한 결과가 나왔다. LDA기능이 있는 대안 1에서의 상위 5개의 노래는 LDA기능이 없는 대안2에서도 존재하며, LDA기능이 없는 대안2에서 3위에 추천 된 <<4144408, 김예림(투개월), All right, Rock>>이 예외적으로 존재함을 볼 수 있다. 본 논문에서는 이에 주목하여 왜 이런 결과가 나왔으며 그 차이는 무엇인지 알아보았다.

- **대안 1 (LDA 기능 O) - 3 위 <<5760115, 빅뱅, IF YOU, Folk>>**

< 5760115, Folk,Blues, 빅뱅 - IF YOU >

그녀가 떠나가요 나는 아무것도 할 수 없어요 사랑이 떠나가요 나는 바보처럼 멍하니 서 있네요 멀어지는 그 뒷모습만을 바라보다 작은 점이 되어 사라진다 시간이 지나면 또 무뎌질까 옛 생각이 나 니 생각이 나 아직 너무 늦지 않았다면 우리 다시 돌아갈 수는 없을까 너도 나와 같이 힘들다면 우리 조금 쉽게 갈 수는 없을까 있을 때 잘할 걸 그랬어 그대는 어떤가요 정말 아무렇지 않은 건가요 이별이 지나봐요 그댈 잊어야 하지만 쉽지가 않네요 멀어지는 그 뒷모습만을 바라보다 작은 점이 되어 사라진다 누군갈 만나면 위로가 될까 옛 생각이 나 니 생각이 나 아직 너무 늦지 않았다면 우리 다시 돌아갈 수는 없을까 너도 나와 같이 힘들다면 우리 조금 쉽게 갈 수는 없을까 있을 때 잘할 걸 그랬어 오늘같이 가녀린 비가 내리는 날이면 너의 그림자가 떠오르고 서랍 속에 몰래 넣어둔 우리의 추억을 다시 꺼내 홀로 회상하고 헤어짐이란 슬픔의 무게를 난 왜 몰랐을까 아직 너무 늦지 않았다면 우리 다시 돌아갈 수는 없을까 너도 나와 같이 힘들다면 우리 조금 쉽게 갈 수는 없을까 있을 때 잘할 걸 그랬어

⇨ **헤어지는 당일을 회상하며, 헤어진 연인을 그리워하는 내용**

다음은 대안1의 3위에 추천된 노래 가사이다. 기준 노래와 마찬가지로 유사성을 쉽게 파악할 수 있도록 유사한 부분을 색으로 표시하였다. <<5760115, 빅뱅, IF YOU, Folk>>의 가사를 살펴보면 “비가 내리는 날이면 너의 그림자가 떠오르고”와 같은 구절은 기준노래의 “ 창문너머엔 슬픈 비가 내리고”, “오늘따라 더 힘들다”와 맥락이 같음을 알 수 있으며 “그댈 잊어야 하지만 쉽지가 않네요 멀어지는 그 뒷모습만을 바라보다”는 기준노래 “헤어지던 날에 멈춰 있고”와 그 내용이 유사하다고 판단하였다. 정리하면, 두 화자 모두 어느 비가 내리는 날, 헤어지던 당일을 회상하며 헤어진 연인을 그리워하고 있음을 알 수 있다.

- 대안 1 (LDA 기능 0) - 4 위 <<4677133, 비스트, 이젠 아니야, Ballad>>

<4677133, Ballad, 비스트-이젠 아니야>

왜 바보같이 날 못 잊고 그러고 있니 그러고 있니 왜 아직까지 난 네게 좋은 사람인 거니 좋은 사람인 거니 우리 헤어진 지가 벌써 몇 달이 지났는데 왜 아직도 넌 지난 추억에 살고 있니 좋은 사람 곁에 많잖아 새로운 사랑 시작해도 괜찮아 웃으며 네가 정말 행복하길 바랄게 헤매이는 네가 눈에 밟혀서 새로운 사랑 시작할 수 없잖아 이렇게 늦은 시간에 넌 왜 또 찾아왔니 널 떠나보낸 비겁한 나인데 문 너머로 들려오는 슬픈 울음소리 왜 또 찾아왔니 이미 차갑게 식어버린 내게 너에게 나눠 줄 온기가 더는 없는데 이젠 아니야 이젠 아니야 네가 기댈 사람 그래 마음껏 울어 그렇게 날 씻어낼 수 있다면 네 맘 속 미련 다 지워낼 수 있다면 네가 아플 만큼 가치 있는 사람 아니야 예전처럼 너랑 같이 있는 사람 아니야 널 사랑해서 보낸단 그런 거짓말 같은 건 하기 싫어 난 오늘만 같이 있어줄게 얼른 일어나 손 내밀어 줄 수 있지만 이것도 오늘까지만 돌아갈 순 없어 알잖아 내 곁에선 행복할 수 없잖아 널 웃게 해줄 그런 사람 찾아 떠나 가 헤매이는 네가 눈에 밟혀서 단 하루도 맘이 편하질 않아 이렇게 늦은 시간에 넌 왜 또 찾아왔니 널 떠나보낸 비겁한 나인데 문 너머로 들려오는 슬픈 울음소리 왜 또 찾아왔니 이미 차갑게 식어버린 내게 너에게 나눠 줄 온기가 더는 없는데 그만해줄래 이젠 시간이 갈수록 냉정해지는 내 모습에 네가 상처받을까 난 너무 두려워 내가 보란 듯이 잘 살아가면 돼 이젠 더 이상 아니야 네가 기댈 사람 왜 또 찾아왔니 널 떠나보낸 비겁한 나인데 문 너머로 들려오는 슬픈 울음소리 왜 또 찾아왔니 이미 차갑게 식어버린 내게 너에게 나눠 줄 온기가 더는 없는데 이젠 아니야 이젠 아니야 네가 기댈 사람

⇨ **헤어지고 그리움에 힘들어하는 전 연인에게 하는 말**

<<4677133, 비스트, 이젠 아니야, Ballad>>의 가사는 헤어지고 그리움에 힘들어하는 전 연인에게 건네는 말과 같다. “헤어진 지가 벌써 몇 달이 지났는데 왜 아직도 넌 지난 추억에 살고 있니”, “돌아갈 순 없어, 알잖아”와 같은 내용은 기준 노래의 “이젠 볼 수 없다는 걸 난 잘 알고 있는데”와 유사하며 “이미 차갑게 식어버린 내게 너에게 나눠 줄 온기가 더는 없는데”와 같은 구절은 기준 노래의 “차갑게 맘에 비가 온다”와 같은 상황을 마주한 상대방과 화자의 입장에서 그 맥락이 유사하다.

- 대안 1 (LDA 기능 0) - 5 위 <<3832101, 버스커버스커, 정말로 사랑한다면, Rock>>

<3832101 , 버스커버스커 - 정말로 사랑한다면>

사랑한단 말로는 사랑할 순 없군요 그대 상처 주네요 나의 뻔한 그 말이 너무 쉽게 뱉은 말 너무 쉬운 사랑은 다 거짓말이죠 그래 다 거짓말이죠 무엇을 원하는지 얼마나 힘든 건지 신경 쓰지 않죠 또 쉽게 넘어 갔나요 많이 힘들었나요 그대가 오늘은 헤어지자 말해요 정말로 사랑한담 기다려 주세요 사랑한단 그 말들도 당신의 행동 하나 진심만을 원하죠 정말로 사랑한담 기다려 주세요 그댈 위해 참아줘요 당신의 행동 하나 아픈 추억 돼가요 정말로 사랑한다면 정말로 사랑한다면 사랑한단 말로는 사랑할 순 없군요 그대 기억 하나요 나의 뻔한 그 말이 그댈 웃게 했던 밤 너무 깊은 그 밤은 다 지나 가네요 모두 다 지나 갔군요 무엇을 말 했는지 얼마나 원했는지 기억 하지 않죠 그대만 지쳐 가나요 많이 힘들었나요 그대가 오늘은 헤어지자 말해요 정말로 사랑한담 기다려 주세요 사랑한단 그 말들도 당신의 행동 하나 진심 만을 원하죠 정말로 사랑한담 기다려 주세요 그댈 위해 참아줘요 당신의 행동 하나 아픈 추억 돼가요 그때 또 하필 잡으려 건네는 이 뻔한 한마디 쉽게 뱉은 말 사랑 한단 말로는 사랑할 순 없군요 허 허우워 아아아 허어어 아아아 이젠 사랑할 순 없군요 사랑한다 기다려 주세요 사랑한단 그 말들도 당신의 행동 하나 진심 만을 원하죠 정말로 사랑한담 기다려 주세요 그댈 위해 참아줘요 당신의 행동 하나 아픈 추억 돼가요 정말로 사랑한다면

⇨ **즐거워했던 날들이 헤어짐을 통해 아픈 추억이 됨, 헤어짐이 다가옴을 느낌 => 헤어짐 부정**

<<3832101, 버스커버스커, 정말로 사랑한다면, Rock>>의 경우 또한 화자는 연인과 즐거웠던 날들이 지나가고 있음을 절감하고 결국 이별을 맞는다. “당신의 행동 하나 아픈 추억돼가요”의 가사와 기준 노래의 “ 날 녹여주던 그 목소리도, 부드러운 손길도 이젠 내 것이 아니라는 게 오늘따라 더 힘들다”또한 유사하며, “그댈 웃게 했던 밤 너무 깊은 그 밤은 다 지나 가네요”, “그대가 오늘은 헤어지자 말해요”를 통해 기준노래와 마찬가지로 화자가 이별을 겪고 연인과 보냈던 즐거운 날들이 아픈 추억이 되어가고 있음을 알 수 있다.

대안1과 대안2의 공동 1,2위는 동일하기에 비교의 의미가 없으므로 분석에서 제외하였고 두 대안에 모두 존재하지만 LDA기능이 존재하는 대안1에서 상위권에 위치해 있는 각각의 3위, 4위, 5위의 노래를 살펴본 결과 모두 기준 노래의 스토리와 분위기가 매우 유사함을 파악할 수 있었다. 본 추천시스템은 노래의 장르, 18군집의 분위기, 발매연도, 계절을 반영하여 유사도를 파악하고 그에 덧붙여 가사 토픽의 유사성을 파악하였다. 앞 선 결과를 통해 실질적으로 본 추천 시스템이 가사에서도 상당히 유사한 스토리를 가진 노래를 추천해 줌을 실질적으로 확인할 수 있었다.

다음은, LDA기능이 있는 대안1에서는 추천 된 노래 순위에 아예 존재하지 않는 노래이지만, LDA기능이 없는 대안2에서는 공통 순위 1,2위를 제외하고 가장 높은 위치인 3위로 추천 된 노래이다.

- **대안 2 (LDA 기능 X) - 3 위 <<4144408, 김예림(투개월), ALL Right, Rock>>**

LDA(X) 3번째 LDA(O) X

<4144408 , Rock, 김예림 (투개월)- ALL Right>

요즘 난 너 가도 이별 따위 한땐 니 생각 넌 내게 안갯속의 요즘 난 너 가도 이별 따위 니 생각 넌 내게 안갯속의 기껏 이거야 내 모든 걸 가졌던 너 없는 게 겨우 이거야 걱정 가득한 너의 마지막 굿바이 넌 그 정돈 아냐 난 요즘 난 너 가도 이별 따위 니 생각 넌 내게 안갯속의 기껏 이거야 내 모든 걸 가졌던 너 없는 게 겨우 이거야 걱정 가득한 너의 마지막 굿바이 넌 그 정돈 아냐 난 우리 추억 영원히 잊지 못할거야 부디 좋은 사람 만나길 바랄게 짐작하지 마 걱정하지 마 안부도 묻지 마 진작 그러지 이제 와 뭐지 넌 언제나 그랬지 요즘 난 너 가도 이별 따위 니 생각 넌 내게 안갯속의 걱정하지 마 너의 그 잘난 이미지 내 입에 담길 일 없는 너 걱정 가득한 따뜻한 그 눈빛은 다음 에게나 줘 요즘 난 너 가도 이별 따위 니 생각 넌 내게 안갯속의 짐작하지 마 걱정하지 마 안부도 묻지 마 진작 그러지 이제 와 뭐지 넌 언제나 그랬지 요즘 난 너 가도 이별 따위 니 생각 넌 내게 안갯속의

⇨ **이별상황에서 너 없이도 나는 ALL RIGHT(괜찮다)**

본 연구에서는 LDA기능이 없는 대안2의 3위인 <<4144408, 김예림(투개월), ALL Right, Rock>>의 가사 분석을 통해 LDA기능의 유의성을 보일 정성적 결과를 얻을 수 있었다.

추천시스템의 정확도를 높이기 위해 전처리 과정에서 영어 가사를 제거하였지만 노래의 본래 가사는 “요즘 난 ALL Right, 너 가도 ALL Right, 이별 따위 ALL Right”로, 이것이 중심 후렴구로서 반복되는 노래이다. “내 모든 걸 가졌던 너 없는 게 겨우 이거야”, “넌 그 정돈 아냐”, “안부도 묻지 마 진작 그러지 이제 와 뭐지”와 같은 가사를 보면 헤어진 연인을 그리워한다기보단 이별 상황에서 상대방 없이도 나는 ALL Right(괜찮다)는 의미를 내포하고 있다. 해당 내용은 기준 노래와는 상당히 상반되는 내용으로 헤어진 옛 연인을 그리워하는 스토리와 전혀 맞지 않다. 그리고 해당 노래는 LDA기능이 포함된 대안1에서는 필터링 되었음을 알 수 있다.

7. 결론

본 추천시스템은 추천에 노래의 장르, 계절, 발매연도 그리고 18개 군집의 분위기를 반영하는 알고리즘을 개발하고 이에 더해 가사 분석을 위한 LDA기능을 추가하였다. 그리고 LDA기능 여부에 따라 대안1과 대안2로 나누어 각각의 결과를 해석하였다.

본문의 6.2에서는 추천시스템의 성능지표인 F-Measure과 이를 추가로 분석한 p-value값을 이용하여 시스템 검증을 하였다. 대안1과 대안2의 비교 결과, 대안1의 F-Measure값이 3.64% 높게 나왔으며 p-value<0.05으로 결과의 유의성을 보였다. 또한, 본문의 6.3에서 실제 나온 결과의 가사분석을 통해 본 추천시스템에서 강조하는 가사 반영의 유의성을 보였다. 두 대안의 가사를 가시적으로 직접 비교한 결과, 같은 소재를 가진 노래(6.3에서는 “이별”)일지라도 화자가 가사를 통해 이를 어떻게 풀어내느냐, 즉, 어떤 토픽으로 구성되어 있는가에 따라 노래가 다른 방향으로 흘러가고 다르게 해석 됨을 볼 수 있다. 그리고 이러한 상세한 분석은 대안1과 대안2의 결과 비교를 통해 가사가 가진 텍스트데이터를 통해서만 필터링 될 수 있음을 보였다. 이를 통해, 추천을 받는 이용자 특히, 노래의 상황과 화자의 감정에 이입하는 이용자라면 가사를 분석하고 필터링하여 추천을 제공하는 본 추천시스템은 기존의 다른 시스템과 비교하여 더욱 유의미할 것으로 예상하는 바이다.

8장. 참고 문헌

[1] 서봉원(2016), 콘텐츠 추천 알고리즘의 진화, 한국콘텐츠진흥원

[2] Hwang, Y. S.(2016), A Study on 11st’s Product Recommendation system using Big Data and Natural Language Processing Technology

[3] 임 일 (2015), R을 이용한 추천 시스템, 카오스북, 3-5

[4] Park, T. S. and Jeong O. R. (2015), Social Network Based Music Recommendation System, Journal of Internet Computing and Services(JICS) 2015.: 16(6), 133-141

[5] Choi, H. G. and Hwang E. J. (2012), Emotion-based Music Recommendation System based on Twitter Document Analysis, Journal of KIISE: 18(11), 762-767

[6] Gong, M. S. A (2016) Music Recommendation Scheme using Analysis of User's Music Listening History and Content-Based Filtering, Soongsil Univ

[7] Lee, J. H. and Lim, H. W. (2016), Similarity Evaluation of Popular Music based on Emotion and Structure of Lyrics, KIISE Transactions on Computing Practices, Computer Science Dept, Seoul National Univ

[8] Lee, C. J. and Bang, H. Y. (2015), Music Recommendation System Using Text Analysis of Lyrics, Korean Institute of Intelligent System: 25(2), 99-100.

[9] Won, J. Y. (2005), Content-Based Music Filtering Scheme Using Representative Melodies in Music Recommendation Systems

[10] Boom4u, Genius, hiphople and so on

[11]원중호 외 3인, (2017), 텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류, 2

[12] Lee, H. D. and Kim, J. B. (2017), Issue Keyword Extraction Method Using Document Similarity Method - Focused on Internet Articles - , 385

[13] Song, H, M (2016), 혼합형 데이터에서 유사도 측정을 통한 군집화 방법, 6-7

[13] Kim, S, H, Yoon, J, T and Seo, J, Y. (2014), Semantic Dependency Link Topic Model for Biomedical Acronym Disambiguation, 656

[14] 이정훈, (2018), 품질경영기사 필기, 18