KOFICE
한국국제문화교류진흥원

2022 빅데이터 활용 한류 시장조사

최종보고서

2023. 7. 31

요 약

1. 개요 및 목적

본 사업은 COVID-19 팬데믹과 엔데믹을 거치며 급변한 동태적인 한류 현상을 이해하고 국내외 한류 정보를 종합적으로 파악하기 위한 기초 자료를 제공할 것을 목적으로 수행되었다. 이에 따라 소셜 미디어 및 영문 언론 등 인터넷 환경에서 발생하는 한류 콘텐츠와 관련된 버즈(Buzz)를 분석하고 적시성이 높은 정보를 제공하여 분야별, 산업별, 국가별 수출시장에 대한 분석을 제공하고자 하였다.

한류 빅데이터 대시보드를 구축함에 있어서, 자체적인 데이터 분석 역량이 부족한 중소 기획사 및 제작사, 관련 연구자 등 관련 산업 종사자의 다양한 계층을 지원할 수 있도록 고려하였다. 사업 기간 및 예산의 범위에서 우선 일반 대중도 접근할 수 있는 수준의 기초 분석과 관련된 데이터를 웹 대시보드를 통해 제공하고, 심층 분석이 필요한 경우에는 별도의 요청하에 관련 데이터를 제공할 수 있도록 하여, 차별적인 수요에 대해 지원할 수 있도록 하였다.

2. 주요 과업 내용

본 사업에서는 소셜 미디어, 리뷰 플랫폼, 언론에서 영어로 작성된 게시글 및 댓글, 국가별 영문 언론에서 작성된 온라인 기사를 수집하여 분석하였으며, 소셜 미디어는 동영상 플랫폼 중 Youtube를 선정하였고 커뮤니티로 Reddit을 선정하여 수집과 분석을 실행하였다.

Youtube의 경우에는 무료로 서비스되고 있는 글로벌 동영상 플랫폼으로서, K-POP 에이전시에서 운영하는 공식 계정을 통해 뮤직비디오 및 음원을 공개할 뿐 아니라, 팬캠(Fancam), 교차 편집 영상, 커버 영상 등이 게시되어 확산되는 플랫폼 특징이 있다. 또, 사용자 경험을 위한 연관 분석 알고리즘을 통한 동영상 추천 기능으로 인해, 해당 영상에 관심이 있고 선호하는 사용자의 시청 빈도를 높이는 특성상 긍정적인 반응이 나타나는 경우 조회 수 및 좋아요, 댓글 수 등 정량적 크기가 급증하여 표현되는 특징이 있다.

Reddit의 경우에는 영어권의 대표적인 종합 커뮤니티 채널로서 주제별로 다양한 서브레딧(Sub Reddit)을 통해 해당 주제에 대한 다양한 의견이 교환되는 특성이 있다. 연관분석에 의한 추천 알고리즘이 없고, 사용자의 반응이 많은 게시물은 메인화면에 게시되는 커뮤니티 채널의 운영 특성상, 화제가 되어 노출되는 게시글에 대해 평소 접해 보지 못하 였거나 부정적인 의견을 가진 사람 등, 다양한 의견을 가진 사람들이 의견을 게시하여 비판적인 의견에 대해 탐색하고 분석하기에 적합한 채널 특징이 있다.

수집한 데이터에서 분석 대상이 되는 키워드는 최근 5년 내 해외 진출 경험이 있는 국내 가수의 대표곡, 당해 연도 해외 수출된 한국 방송프로그램, 국내 제작·방영된 한국 방송프로그램, 당해 연도 해외 수출된 한국 영화, 당해 연도 개봉한 한국 영화, 그리고 기타 이슈가 발생한 한국 방송프로그램 및 한국 영화를 선정하여 분석하였다.

분석 방법은 정량적 분석과 정성적 분석을 통해 각 분야 및 채널별로 빈도 수, 키워드 분석, 감성 분석, 연관어 및 네트워크 분석 등을 심화 분석하였고, 데이터의 수집은 2022년 1월 1일부터 12월 31일까지의 게시글을 기준으로 진행하였다. 수집된 데이터 중 원문과 원문 링크, 수집 키워드 정보 등은 연구 및 분석 목적으로 활용 가능토록 별도로 제공된다.

데이터의 수집을 위해 검색 및 분석에 쓰인 키워드는 아티스트(그룹) 416개, 영화 제목 388개, 드라마 제목 451개, 노래 제목 1,716개를 기준으로 해당 데이터의 수집 및 분석을 위해 파생되는 연관 키워드 사전을 구축하여 사용하였다.

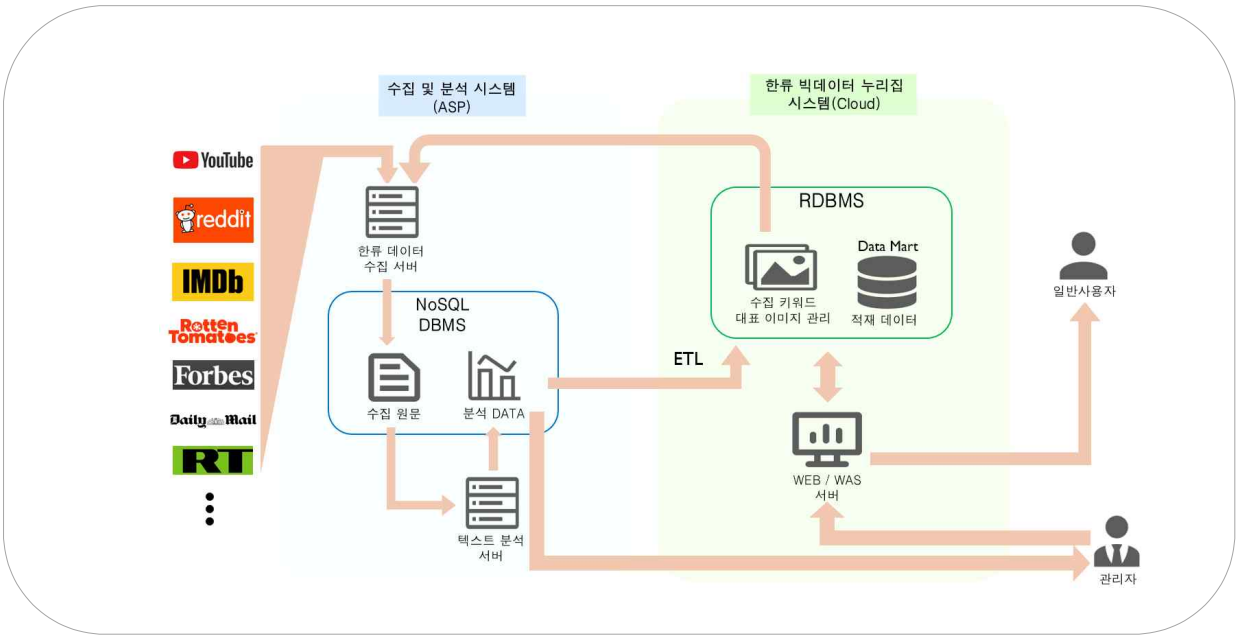
〈 해외 한류 콘텐츠 수집·분석 제공 과업 내용 〉

구 분	내 용
수집언어	■ 단일언어 : 영어
수집채널	■ 게시글, 동영상, 언론 총 3개 분야 - 게시글 : 커뮤니티(Reddit 포함)의 작성 글 기준 (대표성 확보, 지속적·안정적 수집이 가능한 채널 선정) - 동영상 : 관련 동영상 조회 수와 반응 수 (Youtube 포함, 지속적·안정적 수집이 가능한 채널 선정) - 언론 : 국가별 영문 언론에서 작성된 온라인 기사
분석내용	■ 정량분석 - 분야별·채널별 빈도수(조회 수, 댓글 수, 게시일 등 주요 정량 지표 포함) ■ 정성분석 - 분야별·채널별 키워드 분석, 감성 분석, 연관어·네트워크 분석 등 - 기타 심화 분석
조사대상	■ 음악(가수, 곡) - 최근 5년 내 해외 진출 경험이 있는 국내 가수 - 가수의 대표곡 ■ 드라마 - 당해 연도 해외 수출된 한국 방송프로그램 전수 - 당해 연도 국내 제작·방영된 한국 방송프로그램 전수 - 기타 이슈가 발생한 한국 방송프로그램 ■ 영화 - 당해 연도 해외 수출된 한국 영화 전수 - 당해 연도 개봉한 한국 영화(국내 극장, 글로벌 OTT 포함) 전수 - 기타 이슈가 발생한 한국 영화
조사방법	■ 온라인 데이터 수집 - 전문 프로그램으로 수집
수집기간	■ 게시 일자 기준 2022.1.1.~2022.12.31. (댓글 수집 기간은 +4w)
데이터 제공	■ 원문 제공 및 원문 링크, 수집키워드 정보

3. 시스템 구성

한류 빅데이터 대시보드 페이지의 시스템은 ASP로 구축된 수집 및 분석 시스템과 클라우드에 구축된 한류 빅데이터 누리집 시스템의 두 부분으로 구성되어 있다. 수집 및 분석 결과를 누리집 시스템에서 표현할 수 있는 형태의 데이터로 변환하여 연동하는 데이터의 변환 및 적재 시스템이 구성되어 있으며, 해외한류실태조사 결과를 누리집 시스템에서 표현할 수 있도록 정규화하여 적재할 수 있는 RDBMS 데이터베이스 상의 데이터 마트가 구성되어 있다.

〈 한류 빅데이터 수집 · 분석 시스템 및 누리집 시스템 구성도 〉

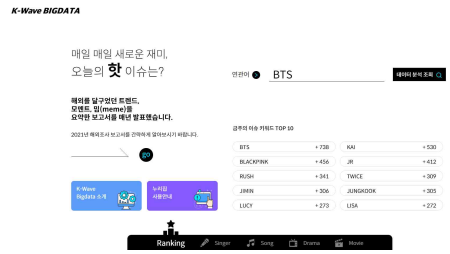
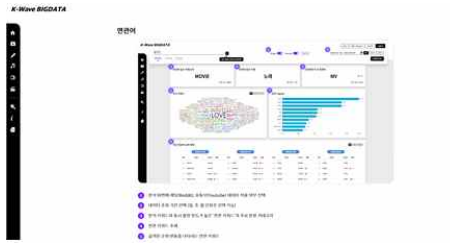
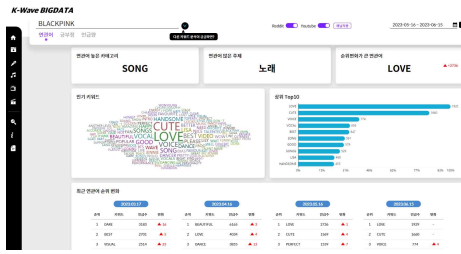
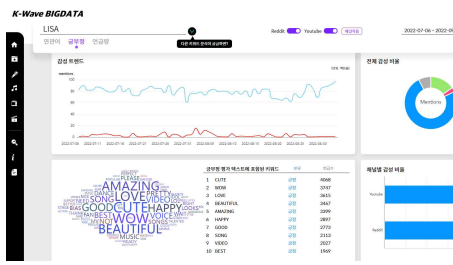


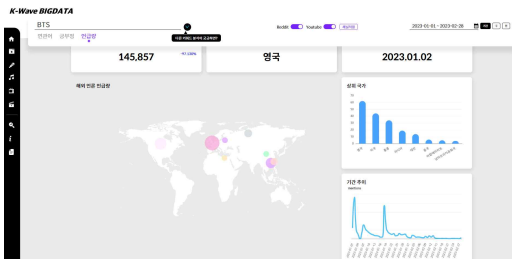
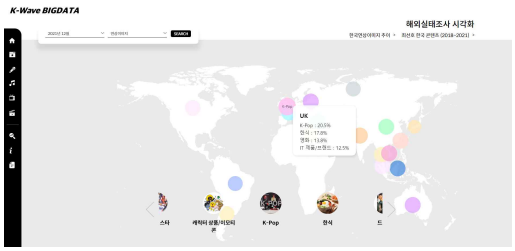
4. 화면 및 기능 구성

누리집 화면은 랭킹, 인트로, 누리집 소개 및 사용 안내와 같은 사용자의 누리집 접근과 이해를 위한 페이지와 연관어 분석, 긍부정 분석, 언급량 분석 및 해외한류실태조사 시각화와 같은 분석 대시보드 화면으로 구성되어 있다.

한정되지 않은 일반 대중에게 공개될 것을 목적으로 한 대시보드 차트는 일반적으로 많이 사용하는 차트로 전문적인 지식이 없이도, 최대한 직관적으로 이해할 수 있는 차트를 주로 사용했다.

〈 누리집 화면 및 기능 구성 〉

구 분	내 용	화면 이미지 예시
랭킹 페이지	■ 주간 이슈키워드 Top 10 제공	
인트로 페이지	■ 가수, 노래, 드라마, 영화 네가지 주제에서 분석을 통해 소셜 미디어에서 버즈가 많이 발생하는 주요 키워드를 대표 이미지와 함께 노출	
누리집 소개	■ 누리집의 목적 및 수집에서 분석에 이르기까지 서비스 구성 개념과 서비스의 특징 등을 소개	
누리집 사용 안내	■ 누리집의 각 페이지 및 대시보드 기능 및 차트에 대한 간략한 설명을 제공	
연관어 분석 대시보드	■ 아티스트 및 콘텐츠 키워드에 연관되어 발생한 키워드에 대한 분석을 제공 - 워드 클라우드(WordCloud) 차트, 막대(Bar) 차트와 함께 기간별 연관어의 언급 수와 순위변화 등을 제공	
긍부정 분석 대시보드	■ 아티스트 및 콘텐츠 키워드에 대한 긍 부정 감성평가를 제공 - 기간별 긍 부정률 추이 및 비율, 긍 부정 감성어의 워드 클라우드 차트, 연관되어 발생한 키워드에 대한 분석을 제공	

언급량 분석 대시보드	■ 기간별 언급량 추이 및 해외 언론 언급량을 국가별 맵차트에 표현하여 제공	
해외한류실태조사 시각화	■ 해외한류실태조사의 주요 결과를 맵차트에 시각화하여 표현	

5. 한계점 및 향후 개선사항

금번에 구축하여 서비스하는 한류 빅데이터 대시보드는 자유도가 매우 높은 소셜 미디어에 게시되는 텍스트 데이터를 수집 및 분석한다. 따라서 관련도가 낮은 데이터가 다량 수집되거나, 분류 처리의 정확도가 높지 않을 수 있으며, 전체 데이터의 규모를 파악하기 어려워 전통적인 통계학적 표본 요건을 충족하지 못하는 등의 빅데이터의 특성에 따른 한계점이 있다. 다만, 이러한 한계점들은 빅데이터 분석 방법론에 따라 인정될 수 있고, 충분히 큰 데이터 셋을 수집 및 분석하여 전체적인 데이터의 경향성을 파악하는 데 주안점을 두고 있다.

〈 한류 빅데이터 대시보드의 한계점 〉

구 분	내 용
분석 목적 이외의 데이터 다량 수집	■ 소셜 미디어 데이터 수집 시 추천 알고리즘이 관련 없는 데이터도 출력될 수 있음 ■ 키워드와 관련 없는 데이터가 다량 수집되어 분석 대상과 무관한 정보가 포함될 수 있음
높은 비정형성에 의한 분석의 어려움	■ 소셜 미디어 데이터는 문법적인 규칙이 통일되지 않아 자연어 처리와 분류가 어려움
데이터의 표본성 문제	■ 전체 데이터 규모를 파악하기 어려워 전통적인 통계학적 표본의 요건 충족이 어려움
키워드 자유 검색 제한	■ 초기 시스템 구축에 예산 제약이 있어 사용자에게 제한된 분석 결과 검색 기능을 제공 ■ 세밀한 분석이 어려워 연구자나 실무자의 개별 요청에 의해 RawData를 제공하는 방식으로 일부 해소

향후 대시보드 시스템은 현재는 제한되어 있는 키워드 자유 검색의 제한을 해소하고, 보다 다양한 콘텐츠 형태 및 한류 산업에 대한 분석이 가능하도록 주제를 확장할 필요가 있다. 다만 이러한 주제의 확장을 위해서는 다른 문화와의 융복합 현상 및 시장경제에서의 기업 간 국제 협업에 따라 형성되는 다양한 콘텐츠에서 어떠한 범위까지를 한류 콘텐츠로 분석할지에 대한 연구가 선행되어야 할 수 있다.

〈 한류 빅데이터 대시보드의 개선사항 〉

구 분	내 용
키워드 자유 검색 제한 해소	■ 현재 시스템은 키워드 자유 검색을 제한하고 상위 키워드를 선택하는 방식을 사용하고 있음 ■ 키워드 자유 검색 기능을 추가하고 검색 엔진 시스템을 확장하는 개선 검토가 필요할 수 있음
분석 주제의 확장	■ 웹툰, 게임 등 다양한 콘텐츠 형태의 소비와 한식, 한국 관광 등 다양한 측면의 한국 문화에 대한 분석 주제를 확장하여 다양한 분야에 대한 데이터 수집 및 분석을 진행 할 필요가 있음
문화 융복합에 따른 구분의 문제	■ 한류 콘텐츠와 다른 문화와의 융복합 현상에 따라 정의하는 문제를 고려해야 함 ■ K-POP 아이돌 등 타국 국적의 연예인이 포함되는 경우나, 국가 경계를 넘나드는 콘텐츠 등에 대한 한류 콘텐츠의 정의를 명확히 하고 구분하는 방법에 대해 연구가 필요할 수 있음

차 례

〈표 차례〉

〈그림 차례〉

KOFICE
한국국제문화교류진흥원

Contents

KOFICE
한국국제문화교류진흥원

1

2022 빅데이터 활용 한류 시장조사
최종보고서

개요 및 목적

1. 사업 개요

Ⅰ. 개요 및 목적

1. 사업개요

본 사업은 COVID-19 팬데믹과 엔데믹으로 급변하는 생활 양식의 변화 아래에서 동태적인 한류 콘텐츠 선호 현상을 이해하고 국내외 한류 현황을 종합적으로 파악하기 위한 지식 정보와 통계 자료에 대한 수요가 증가하는 국내 산업 배경을 바탕으로 수행되었다. 기초적이고, 중립적인 통계 자료를 전문적인 연구자 이외에 관련 산업 종사자 및 일반 대중에게 제공하기 위해 수행되었다.

인류의 생활 양식과 관심사가 급격하게 변화하는 엔데믹 사회의 콘텐츠산업 환경 변화에 선제적으로 대응하고 한류 관련 미래 전략 수립을 지원하기 위해, 각계각층에서 한류와 관련한 체계적인 정보의 필요성이 대두되었다. 관계 부처에서도 정례적으로 일정 수의 표본을 확보하여 조사 및 분석하여 발간하는 설문조사 결과 이외에도, 한국 대중문화에 대한 인식이 부족한 잠재 소비층을 포함해서 세계 속에서 한류의 전반적인 관심과 소비 관여도를 파악하기 위해 데이터 원천을 달리하는 소셜 미디어 빅데이터 분석이 요구되었다.

본 사업의 선행 사업으로 소셜 미디어에서 나타나는 한류 데이터를 분석하여 보고서로 발간한 「2021 빅데이터 활용 한류 시장조사」[1] 사업이 수행되었고, 소셜 미디어에서 시기별로 한류 콘텐츠가 확산하는 정도에 대한 분석을 수행하여 공유하였다.

선행 사업을 통해 획득한 경험으로서 년 주기의 분석 보고서 발간을 통해서는 적시성 높은 정보를 제공하기 어려운 동태적인 한류 현상에 대해 객관적이고 신속한 분석으로 분야별, 산업별, 국가별 수출시장에 대한 정보를 제공하여야 할 필요성이 제기되었다. 그리하여 금번 사업을 통해서는 지속적으로 적시성 높은 정보를 제공할 수 있는 시스템을 구축하고자 하였다.

대시보드를 구성함에 있어서 자체적으로 데이터를 분석해 시장을 개척하기 어려운 중소 기획사 및 제작사 등 관련 산업을 지원할 수 있도록 고려하였다. 그러나 한정된 예산과 구축 기간 동안에 웹사이트에 접근하는 다양한 사람의 심층적인 데이터 분석 수요를 모두 충족시킬 수는 없을 것으로 판단되어, 우선은 일반 대중도 접근할 수 있는 수준의 기초적인 분석과 관련된 수취 데이터를 제공하도록 하였다. 그 외 심층적인 분석이 필요한 경우에는 별도의 요청하에 관련 데이터를 제공할 수 있도록 하여, 웹사이트의 대중성을 확보하여 홍보하고, 산업에 대한 지원은 세밀한 데이터를 제공할 수 있도록 구성되었다.

1) 한국국제문화교류진흥원 (2021). 「2021 빅데이터 활용 한류 시장조사」.

KOFICE
한국국제문화교류진흥원

2

2022 빅데이터 활용 한류 시장조사
최종보고서

주요 과업 내용 및 요구사항

1. 주요 과업 내용
2. 요구사항

Ⅱ. 주요 과업 내용 및 요구사항

1. 주요 과업 내용

가. 해외 한류 콘텐츠 빅데이터 수집 · 분석 제공

본 사업에서는 소셜 미디어, 커뮤니티, 언론에서 영어로 작성된 게시글 및 댓글, 국가별 영문 언론에서 작성된 온라인 기사를 수집하여 분석하였다.

1) 수집대상 소셜 미디어

소셜 미디어는 동영상 플랫폼 중 Youtube를, 커뮤니티로 Reddit을 선정하였다.

Youtube의 경우에는 무료로 서비스되고 있는 글로벌 동영상 플랫폼으로서, K-POP 에이전시에서 운영하는 공식 계정을 통해 뮤직비디오 및 음원을 공개할 뿐 아니라, 팬캠(Fancam)[2], 교차 편집 영상[3], 커버[4] 영상 등이 게시되어 확산되는 플랫폼적 특징이 있다. 사용자 경험을 위한 연관 분석 알고리즘을 통한 동영상 추천 기능으로 인해, 해당 영상에 관심이 있고 선호하는 사용자의 시청 빈도를 높이는 특성상 긍정적인 반응이 나타나는 경우 조회 수 및 좋아요, 댓글 수 등 정량적 크기가 급증하여 표현되는 특징이 있다.

Reddit의 경우에는 영어권의 대표적인 종합 커뮤니티 채널로서 주제별로 다양한 서브레딧(Sub Reddit)[5]을 통해 해당 주제에 대한 다양한 의견이 교환되는 특성이 있다. 연관분석에 의한 추천 알고리즘이 없고, 사용자의 반응이 많은 게시물은 메인 화면에 게시되는 커뮤니티 채널의 운영 특성상, 화제가 되어 노출되는 게시글에 대해 평소 접해 보지 못하였거나 부정적인 의견을 가진 사람 등 다양한 의견을 가진 사람들이 의견을 게시하여 비판적인 의견에 대해 탐색하고 분석하기에 적합한 채널 특징이 있다.

2) 분석대상 키워드

최근 5년 내 해외 진출 경험이 있는 국내 가수의 대표곡, 당해 연도 해외 수출된 한국 방송 프로그램, 국내 제작·방영된 한국 방송 프로그램, 당해 연도 해외 수출된 한국 영화, 당해 연도 개봉한 한국 영화, 그리고 기타 이슈가 발생한 한국 방송 프로그램 및 한국 영화를 분석하였다.

분석 방법은 정량적 분석과 정성적 분석을 통해 각 분야 및 채널별로 빈도 수, 키워드 분석, 감성 분석, 연관어 및 네트워크 분석 등을 심화 분석하였고, 데이터의 수집은 2022년 1월 1일부터 12월 31일까지의 게시글을 기준으로 수집하였다. 수집된 데이터 중 원문과 원문 링크, 수집 키워드 정보 등은 연구 및 분석 목적으로 활용 가능토록 별도로 제공된다.

---

2) 각종 공연이나 방송에 관객으로 참석한 팬에 의해서 촬영된 영상
3) 동일한 아티스트의 곡에 대해 공연이나 방송에서 촬영된 여러 영상들을 번갈아서 잘라 붙여서 하나의 뮤직비디오처럼 편집한 영상
4) 노래 및 댄스에 대해 원곡자가 아닌 다른 인플루언서 또는 팬이 부르거나 춤추는 영상을 게시하는 것
5) Reddit 하위 주제별 게시판

데이터의 수집을 위해 검색 및 분석에 쓰인 키워드는 아티스트(그룹) 416개, 영화 제목 388개, 드라마 제목 451개, 노래 제목 1,716개를 기준으로 해당 데이터의 수집 및 분석을 위해 파생되는 연관 키워드 사전을 구축하여 사용하였다.

표 1. 해외 한류 콘텐츠 수집·분석 제공 과업 내용

구 분	내 용
수집언어	■ 단일언어 : 영어
수집채널	■ 게시글, 동영상, 언론 총 3개 분야 - 게시글 : 커뮤니티(Reddit 포함)의 작성글 기준 (대표성 확보, 지속적·안정적 수집이 가능한 채널 선정) - 동영상 : 관련 동영상 조회 수와 반응 수 (Youtube 포함, 지속적·안정적 수집이 가능한 채널 선정) - 언론 : 국가별 영문 언론에서 작성된 온라인 기사
분석내용	■ 정량분석 - 분야별·채널별 빈도수(조회 수, 댓글 수, 게시일 등 주요 정량 지표 포함) ■ 정성분석 - 분야별·채널별 키워드 분석, 감성 분석, 연관어·네트워크 분석 등 - 기타 심화 분석
조사대상	■ 음악(가수, 곡) - 최근 5년 내 해외 진출 경험이 있는 국내 가수 - 가수의 대표곡 ■ 드라마 - 당해 연도 해외 수출된 한국 방송 프로그램 전수 - 당해 연도 국내 제작·방영된 한국 방송 프로그램 전수 - 기타 이슈가 발생한 한국 방송 프로그램 ■ 영화 - 당해 연도 해외 수출된 한국 영화 전수 - 당해 연도 개봉한 한국 영화(국내 극장, 글로벌 OTT 포함) 전수 - 기타 이슈가 발생한 한국 영화
조사방법	■ 온라인 데이터 수집 - 전문 프로그램으로 수집
수집 기간	■ 게시 일자 기준 2022.1.1.~2022.12.31. (댓글 수집 기간은 +4w)
데이터 제공	■ 원문 제공 및 원문 링크, 수집 키워드 정보

나. 데이터 수집 · 분석 시스템 구축 및 데이터 시각화

데이터 수집 · 분석 시스템은 ASP(Application Service Provider)[6]를 기반으로 데이터 수집 및 분석 시스템이 구축되어 있으며, 분석 결과는 웹 대시보드를 표현하기 위한 누리집 데이터베이스에 매일 주기적으로 전송되어 업데이트된 결과물을 대시보드 웹서비스를 통해 시각화되어 제공하도록 설계되었다.

표 2. 데이터 수집 · 분석 시스템 구축(ASP) 및 데이터 시각화 과업 내용

구 분	내 용
수집·분석 시스템 구축 (ASP)	■ 상기 데이터를 자동으로 수집·분석하여 결과물 적재 ■ 발주처 내부 시스템 연동하여 정형·비정형 데이터 수집·분석 ■ 일/주/월간 단위 자동산출 및 업데이트, 모니터링 고려하여 설계 ■ 계약 종료 시점까지 유지보수 포함
대시보드 웹서비스 구축	■ 상기 데이터의 정형·비정형 분석 결과를 대시보드 형태로 시각화 한 웹서비스 제공 ■ 발주처 보유 데이터의 정형·비정형 분석 결과 제시 ■ 사용자 편의를 고려한 화면설계 및 디자인 ■ 추가 기능에 대한 제안 ■ 계약 종료 시점까지 유지보수 포함

6) 고가의 솔루션 구매 비용 부담을 경감하기 위하여 기간 및 사용량 등을 기준으로 네트워크 상에서 솔루션을 임대하여 사용하는 방식

2. 요구사항

대시보드 누리집은 '2022 빅데이터 활용 한류 시장조사'의 제안요청서에 명시된 전체 24개 주요 요구사항 모두를 충족하도록 구축되었다.

가. 기능 요구사항

기능 요구사항에서는 데이터 수집과 분석, 대시보드 웹서비스 표현을 위한 기능을 충족하여, 일반인이 대시보드 웹서비스를 통해 분석 결과를 조회하고 탐색할 수 있도록 요구되었다. 또한, 조회 결과를 쉽게 다운로드 할 수 있고, 수집, 분석과 관련된 현황을 모니터링 가능하며, 부가 기능과 관리자 페이지도 생성하여 서비스를 원활하게 관리할 수 있도록 시스템을 구축하였다.

표 3. 기능 요구사항 및 적용 여부

구 분	내 용	적용 여부
데이터 자동 수집	구축된 자동화 시스템으로 수집 대상 채널에 대하여 반복적으로 동작하며 신규 데이터를 수집·적재할 수 있는 기능의 구현 및 탑재	적용
데이터 수집 현황 확인	수집 관리 기능으로서 데이터 수집 장애 사항에 대해 모니터링이 가능한 현황 파악 기능 제공 - 각 수집 채널별 수집 데이터의 유효성 검사기능 - 각 수집 채널별 수집기의 동작 현황 파악 기능 - 채널별 수집량에 대한 모니터링 기능 등	적용
데이터 분석 현황 확인	수집 관리 기능으로서 데이터 분석 장애 사항에 대해 모니터링이 가능한 현황 파악 기능 제공 - 각 분석 서버 및 분석 기능의 동작 현황 파악 기능 - 채널별 데이터 분석량에 대한 모니터링 기능 등	적용
발주처 내부 시스템 연동	발주처 홈페이지에 게시된 한류 관련 조사연구 자료의 정형·비정형 데이터 연동 - 한국국제문화교류진흥원 홈페이지(http://kofice.or.kr/)에 게시된 조사연구자료 참조 - 연동 데이터의 범위는 발주처와 협의를 통하여 최종 결정	적용
분석 결과 시각화	• 한류 관련 빅데이터의 흐름 전체를 확인할 수 있는 대시보드 및 분류체계에 따른 각 콘텐츠 분야별 흐름을 확인할 수 있는 대시보드 시각화 설계 및 구축 - 시계열에 따른 트렌드 흐름, 신규 키워드, 급상승 키워드 콘텐츠에 대한 긍 · 부정 감성 반응 등이 직관적으로 확인될 수 있도록 대시보드 시각화 설계 • 국가별 · 지역별 구분이 가능한 데이터에는 map 이미지 적용	적용

구 분	내 용	적용 여부
대시보드 웹서비스 구축	웹서비스로서 일반인이 탐색 가능한 대시보드를 구축하여 분석 결과를 조회하고 탐색할 수 있도록 웹서비스 구축 - 수집·분석 데이터에 대하여 지속적인 반영이 이루어져야 함 - 웹서비스의 부하 등을 고려하여 조회 기능의 범위는 발주처와 협의하여 조정하도록 함	적용
키워드 확장성 확보	한류 관련 외부 데이터의 수집·분석 시 분류 키워드 및 유의어, 불용어, 기존 분류의 변경 등이 가능하도록 텍스트 분석 사전의 관리 기능(CRUD) 탑재 - 신규 콘텐츠, 동의어 발생으로 인한 불용어 처리 및 분류에 대한 추가·변경·삭제가 가능하여야 함	적용
옵션별 조회 및 다운로드	데이터 요구사항(DAR)에 따른 데이터를 다음 옵션에 따라 조회할 수 있는 기능 및 결과값 다운로드(csv 파일) 기능을 제공해야 함- 기간별·분야별·채널별(지역별) 결과값 조회 및 다운로드 - 그 외 발주사에서 요청하는 옵션에 따른 조회 및 다운로드	적용
부가기능 및 관리자 페이지	시스템 이용 현황 파악을 위한 부가기능과 관리자 페이지를 제공 - 로그정보, 이용통계 등 제공 - 부가기능은 발주처와 협의하여 확정	적용

나. 데이터 요구사항

데이터 요구사항에서는 한류 콘텐츠와 관련한 해외 반응을 분석할 수 있도록 주요 소셜 미디어와 해외 언론에 게시된 한류 데이터의 수집을 요구하였다. 또한 수집된 데이터를 분석하기 위한 정형 데이터 및 비정형 텍스트 데이터 셋(Data Set)을 생성하여, 분석 결과를 확인하거나, 심층적인 분석 목적으로 재가공할 수 있는 수집 및 분석 원본 상세 데이터를 함께 제공할 것이 요구되었으며 이를 적용하였다.

표 4. 데이터 요구사항 및 적용 여부

구 분	내 용	적용여부
해외 한류 소셜 미디어 빅데이터 수집	• 조사대상 한류 콘텐츠가 언급된 해외 소셜 미디어 데이터를 수집·저장함 - 해외 소셜 미디어 데이터는 영문으로 작성된 게시글 및 댓글 반응을 대상으로 함 - 조사대상 한류 콘텐츠의 분야는 음악(K-POP 가수, 곡), 드라마, 영화 등 콘텐츠 분야를 기본으로 하며 그 외 수집 분야에 대하여 발주처와 협의를 통하여 최종 결정 • 소셜 미디어 데이터의 수집 매체는 주요 조사대상 한류 콘텐츠의 반응이 잘 나타날 수 있는 소셜 미디어(Youtube 외) 및 리뷰 플랫폼(IMDb 등)으로 발주처와의 협의를 통해 최종 확정 - 주요 정량지표(게시글 수, 댓글 수, 조회 수, 좋아요 수), 비정형 데이터(게시글 본문, 댓글 본문)를 수집·저장 • 소셜 미디어 데이터의 수집은 2022.01.01. 이후 작성된 게시물에 대하여 시스템을 통하여 지속 수집·저장 실시	적용
해외 한류 언론 기사 빅데이터 수집	• 조사대상 한류 콘텐츠가 언급된 해외 언론 데이터를 수집·저장함 - 조사대상 한류 콘텐츠의 분야는 음악(K-POP 가수, 곡), 드라마, 영화 등 콘텐츠 분야를 기본으로 하며 그 외 수집 분야에 대하여 발주처와 협의를 통하여 최종 결정 - 해외 언론은 영문으로 작성되어 온라인에 게시된 기사를 대상으로 함 • 해외 언론 기사 데이터의 수집 매체(언론)는 국가별 주요 영문 언론사를 대상으로 함 - 한류가 확산되고 있는 10개 이상 국가의 언론을 수집·저장하며 대상 언론은 발주처와 협의를 통하여 최종 결정 - 게시된 기사의 제목·본문 텍스트 데이터를 수집 대상으로 함 • 해외 언론 기사 데이터의 수집은 2022.01.01. 이후 작성된 기사에 대하여 시스템을 통하여 지속 수집·저장 실시	적용
정형 데이터 분석	• 분석 데이터 정의 - 과업 수행에 필요한 분석 데이터 정의 • 데이터의 정제, 전처리 - 이상치·비적합 데이터 제거, 대체 등 정형 데이터 정제 및 변환 등 전처리 • 데이터의 분석 - 대시보드 활용이 용이하고 분석 및 유지 관리에 적합한 정형 데이터 셋 생성	적용

구 분	내 용	적용여부
비정형 데이터 분석	• 분석 데이터 정의 - 과업 수행에 필요한 분석 데이터 정의 • 데이터의 정제, 전처리 - 이상치·비적합 데이터 제거, 대체 등 비정형 데이터 정제 • 데이터의 분석 - 자연어 처리(NLP, Natural Language Processing)를 통한 비정형 텍스트 데이터 변환 및 분석 - 키워드 분석, 감성 분석, 연관어·네트워크 분석 등을 제공 - 텍스트 데이터의 분류·분석에 있어서 분석 결과물의 활용성을 높이기 위해, 분야별 분류체계(Taxonomy) 구성 및 주요 키워드의 도출은 발주처에서 각 콘텐츠 분야 전문가로 구성한 위원회와의 유기적인 협업을 통하여 구성 - 향후 지속적으로 변화되는 텍스트 데이터에 대해서도 분석 정확도를 확보하기 위하여 도입되는 텍스트 데이터 분류·분석 솔루션은 발주처의 담당자가 쉽게 키워드 및 분석 규칙을 추가, 수정할 수 있도록 구성하여 제공	적용
상세 데이터 지원	• 분석 결과를 확인하거나 재가공하기 위한 상세 데이터 제공 • 주제별 심층 분석 등 연구 목적의 데이터 지원 • 동영상·게시글의 URL 등 상세정보제공	적용

다. 성능 요구사항

성능 요구사항에는 안정적인 운영 및 지원을 위한 일반적인 성능 요구사항과 시스템 오류에 대한 메시지 응답, 대용량 데이터 질의에 대해 수행 지연이 발생할 수 있는 데에 대한 사전 경고 등의 성능이 요구되었으며 요구사항을 적용하였다.

표 5. 성능 요구사항 및 적용 여부

구 분	내 용	적용 여부
성능 일반 요구사항	• 사업 대상 시스템의 성능을 고려한 구축 방안을 제시하여야 함 • 안정적 운영 지원 및 사용자 지원 방안 제시	적용
시스템 및 오류 응답	• 등록, 오류 등 사용자 확인 메시지 제공 - 정보 요청에서 결과가 조회되는 것에 대한 응답 제시 - 시스템의 각 웹페이지의 경우, 수 초 내에 완전히 디스플레이 되어야 함 - 오류 메시지는 사용자가 인지하여 즉시 조치할 수 있도록 작성되어야 함 ※ 예외사항 : 대량의 데이터에 대한 검색 요청, 다수의 필터 혹은 연산 조건이 적용된 데이터 요청에는 적용되지 않음	적용
수행 지연에 대한 사전 경고	• 대용량 데이터의 질의 등 수행 지연 건에 대한 사용자 사전 경고 - 작업 수행 시간이 오래 걸리는 경우 사용자에게 이를 안내하는 기능 - 팝업 등의 수단으로 사용자에게 작업 시작 전에 경고 메시지를 출력하여 사용자가 작업 취소가 가능할 수 있도록 기능 구현 - 수행 지연 시 Status Bar 등을 활용해 작업 진행 상황 제시	적용

라. 인터페이스 요구사항

사용자 인터페이스 요구사항에서는 다양한 브라우저 환경에서 사용할 수 있는 인터페이스 표준에 대한 준수와 사용자 및 관리자가 쉽고 편하게 사용할 수 있도록 직관적인 사용자 인터페이스가 요구되었고 이에 대해 적용하였다.

표 6. 인터페이스 요구사항 및 적용 여부

구 분	내 용	적용 여부
인터페이스 표준 준수	사용자 인터페이스 표준 준수 - 발주처가 요청하는 다중 브라우저 지원(Microsoft Edge, Chrome, Safari 등)	적용
사용자 인터페이스 제공	사용자 및 관리자가 시스템을 쉽고 편하게 사용할 수 있도록 사용자 인터페이스 제공 - 데이터 활용도에 따라 정보 접근의 편의성을 높인 화면 구성 - 초급자도 쉽게 운영할 수 있는 직관적인 인터페이스 제공	적용

마. 보안 요구사항

보안 요구사항에서는 개인정보보호를 비롯한 사업 수행 과정 및 종료 후 과제에 사업에 대한 보안정책 준수를 요구 받았으며 이에 따른 요구사항을 준수하였다.

표 7. 보안 요구사항 및 적용 여부

구 분	내 용	적용여부
보안 법규 및 지침 준수	• 개인정보보호 및 정보 보안 강화와 관련 발주처가 권장하는 보안체계 준수, 각종 정책, 지침 및 매뉴얼 등 준수 • 사업자는 본 사업 수행 과정에서 취득한 모든 정보를 사업 수행 중 혹은 완료 후 발주처의 승인 없이 외부에 유출 또는 누설할 수 없음 - 용역 사업 결과물 및 용역 수행 중 수집한 데이터, 결과물 일체 (발주처가 외부 공개를 요청한 일부 보고서 제외) - 기관에서 사용하는 내·외부 IP 주소 현황 - 시스템 접근 권한 정보 - 기타 보안이 필요하다 판단되는 개인정보 및 각종 내부 문서	적용

마. 프로젝트 관리 요구사항

프로젝트 관리 요구사항에서는 프로젝트 진행에 대한 보고 및 산출물에 대한 요구사항이 제시되었으며 이에 대해 준수하였다.

표 8. 프로젝트 관리 요구사항 및 적용 여부

구 분	내 용	적용여부
보고 계획 수립 및 관리	사업 관리를 위한 보고 계획 수립 및 이행	적용
산출물 관리	사업관리를 위한 산출물 관리	적용

바. 프로젝트 지원 요구사항

프로젝트 진행 지원에 있어서 업무협조 및 발전방안 컨설팅과 관련한 요구사항이 있었으며 이에 대한 지원을 진행하였다.

표 9. 프로젝트 지원 요구사항 및 적용 여부

구 분	내 용	적용여부
업무협조	• 계약 대상자는 용역을 수행함에 있어 문화체육관광부 등 정부기관 협조 요청 시 적극 지원하여야 함 • 검증 관련 자료 요청 시 적극 지원하여야 함 • 계약 기간 종료 후 제3자가 동일 사업을 수행하게 될 경우 계약자는 제3자에게 사업 수행에 필요한 전반적인 사항에 대한 인계를 성실히 수행해야 함	적용
발전방안 컨설팅	한류 빅데이터 조사 설계·분석의 정교성과 내실 향상을 위한 제안 • 시스템의 기능 확장 및 고도화, UI/UX 기능 제안 등 한류 종합정보시스템(ISP) 구축을 위한 추진전략 • 운영·유지 보수에 관한 제안	적용

KOFICE
한국국제문화교류진흥원

3

2022 빅데이터 활용 한류 시장조사
최종보고서

시스템 설계 및 구성

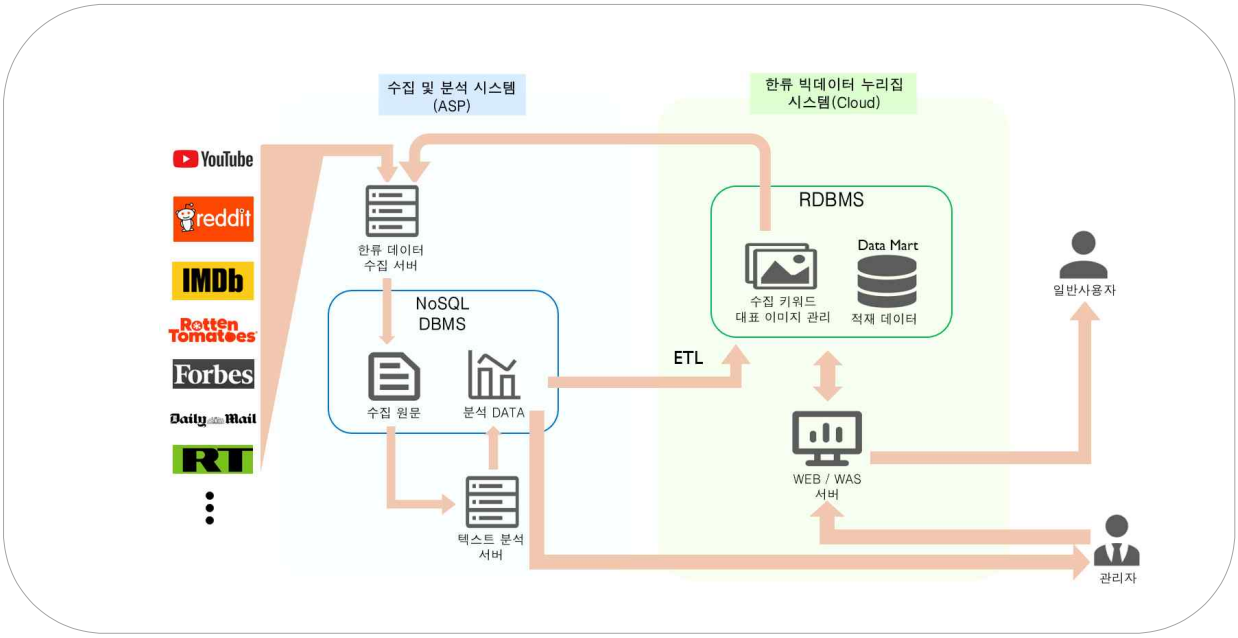
1. 시스템 구성도
2. 각 시스템 특징

Ⅲ. 시스템 설계 및 구성

1. 시스템 구성도

한류 빅데이터 대시보드 페이지의 시스템은 ASP로 구축된 수집 및 분석 시스템과 클라우드에 구축된 한류 빅데이터 누리집 시스템의 두 부분으로 구성되어 있다. 수집 및 분석 결과를 누리집 시스템에서 표현할 수 있는 형태의 데이터로 변환하여 연동하는 데이터의 변환 및 적재 시스템이 구성되어 있으며, 해외한류실태조사 결과를 누리집 시스템에서 표현할 수 있도록 정규화하여 적재할 수 있는 데이터 테이블이 구성되어 있다.

그림 1. 한류 빅데이터 수집 · 분석 시스템 및 누리집 시스템 구성도



2. 각 시스템 특징

가. 한류 빅데이터 누리집 시스템 구성

한류 빅데이터 누리집 시스템은 연구자 이외의 산업 종사자 및 일반 대중에게 한류와 관련된 분석 정보를 공공 데이터로서 웹상에서 제공할 것을 목적으로 하는 BI 대시보드[7] 시스템으로 기획되었다.

대량으로 수집되고 분석된 비정형 텍스트 데이터에서 다양한 차트 형태의 연관 데이터를 제공하기 위해서는 데이터를 다양한 차원[8]을 기준으로 집계하여 원본 데이터를 네트워크를 통하여 전송해야 할 필요성이 있다. 그러나 한정되지 않은 숫자의 사용자가 제한되지 않는 다양한 조건으로 실시간으로 원본 데이터에서 조회를 요청하게 되면 대량의 데이터에 대해 연산을 수행하기 위한 큰 연산 자원(Computing Resources)이 필요하고, 이에 따라 서비스에 장애가 발생하거나, 소요되는 비용이 예산 범위를 초과할 가능성이 제기되었다.

이에 따라 안정적인 서비스를 효율적인 비용으로 제공하기 위하여 사용자가 실시간으로 원본 데이터에 대한 연산을 수행하여 결과물을 받아볼 수 있는 방식은 배제하고, 주로 요청되는 형태의 결과를 쉽게 반환할 수 있도록 매일 가공하여 갱신된 집계 데이터에서 원하는 결과를 조회하여 빠르게 결과를 볼 수 있도록 시스템을 구성하였다.

나. 수집 및 분석 시스템 구성

수집 및 분석 시스템은 ASP로 구성된 H/W 및 S/W를 사용하여 프로젝트의 초기 예산 소요를 줄여 한정된 예산 범위 내에서 전체 시스템을 구성할 수 있도록 기획되었다.

1) 수집 시스템 구성

수집 시스템은 Youtube, Reddit, IMDb, Rotten Tomatoes 및 다양한 언론 페이지에서 한류와 관련된 키워드로 데이터를 검색하고 수집하는 것이다. 다양한 키워드로 해당 데이터를 조회하는 과정에서 요구되는 네트워크 사용량에 대한 고려가 필요하며, 해당 서비스의 서버에 과도한 부하를 주지 않기 위한 최적화, 반복적인 수집을 위한 스케줄링, 정상적으로 수집되고 있음을 확인하기 위해 모니터링 등이 요구되었다.

또한 글로벌 플랫폼에서 연관 조회 알고리즘을 제공하는 것과, 다양한 국가의 사람들이 작성하는 댓글 등으로 인하여 조회되는 데이터는 다국어지만 사업의 수집 및 분석 대상 언어는 영어로 한정되어 있다. 특히 한류 콘텐츠와 관련돼서 조회되는 한국어 버즈의 수량이 많으나 분석 대상이 되지 않는다. 따라서 이러한 영어 이외의 다국어를 수집 단계에서 필터링하여 분석 대상이 아닌 데이터가 과도하게 수집되지 않도록 하고 있다.

또한 여러 대의 서버가 수집한 데이터를 지연 없이 적재하기 위해 데이터베이스의 하드웨어 및 소프트웨어 성능이 요구되었으며, 이를 달성하기 위해 수집 및 분석을 위한 데이터의 저장용 데이터베이스

---

7) Business Intelligence Dashbord : 데이터를 그래프, 차트, 표등 시각적 요소로 표현하여 데이터를 이해하기 쉽도록 하는 시스템
8) 데이터의 특성이나 정보를 나타내기 위해 사용되는 변수(속성)로 데이터를 표현하는 데 사용되는 축(채널, 일자, 연관어, 긍부정 등)

소프트웨어는 NoSQL을 기반으로 구성하였다. NoSQL 데이터베이스는 RDBMS에 비교할 때 대량의 데이터의 조회 및 적재에 유리하고, 다양한 형태의 수집 원천에서 요구되는 속성의 추가로 인한 데이터의 스키마 변경이 유연하다.

2) 분석 시스템 구성

분석 시스템은 NoSQL에 적재된 비정형 텍스트를 분류 분석하기 위해, 문장을 최소 의미를 가지는 형태소 단위로 분리하고, 형태소 간의 빈도, 거리, 순서 및 AND, OR 등 논리연산을 통해 해당 문장이 한류 콘텐츠에 연관되는지와 어떠한 분류에 해당되는지 출현하는 주요 키워드를 분석하고 결과를 다시 NoSQL DB에 적재되도록 하였다.

다. 데이터 변환 및 적재

NoSQL에 수집 및 분석되어 적재된 데이터는 ETL[9] 프로세스에 따라 추출 및 가공된 후 RDBMS[10] 시스템에 적재되며, RDBMS 시스템은 이후 사용자의 다양한 조회요구에 빠르게 응답할 수 있도록 다수의 사실(Fact) 테이블과 연결된 차원(Dimension) 테이블로 구성되어 스타 스키마(Star Schema)[11] 구조에서 확장된 다중 사실 스키마(Multi Fact Schema)로 변환 및 적재되고 야간 및 새벽의 유휴시간대를 이용하여 매일 증분적재[12] 되도록 설계되었다.

라. 해외한류실태조사 데이터 정규화 및 적재

한류 빅데이터 대시보드 누리집에는 기존에 매년 실시하고 있는 해외한류실태조사에서 선정된 주요 결과 데이터를 취합하여 제공하여 탐색 수 있도록 RDBMS에 정규화된 데이터베이스를 구성하고 시각화를 제공하도록 구성되었다. 해당 데이터베이스는 이후 실태조사의 문항과 답변이 추가될 것을 고려하여 확장성을 가질 수 있도록 구성되어, 새로 발간되는 보고서의 최신 데이터 및 추가로 제공하기를 원하는 기존 설문조사 문항에 대한 데이터를 추가하는 것으로 웹상에 데이터가 관련된 데이터 셋을 선택하고 그 결과를 표현할 수 있도록 구성되었다. 이를 통해 기존에 각각의 연례 보고서 문서를 비교하여야 하였던 시계열 데이터[13]를 웹 플랫폼을 통하여 통합 조회할 수 있도록 함으로써 관련 산업 종사자에게 편의를 제공할 기반을 구축하였다.

9) Extract-Transform-Load : 추출-변환-적재를 순차적으로 실행하는 데이터 가공 및 적재 흐름 방식
10) 데이터를 정형화하여 엑셀 시트와 같은 테이블 형태로 적재하여 사용하는 전통적 방식의 데이터베이스 프로그램 대표적으로 Oracle, SQL Server(MicroSoft), MySQL 등이 있다
11) Data Warehouse에 주로 사용되는 데이터 구조로서 간단하고 직관적인 구조, 대량 데이터의 조회 성능 향상, 확장에 대한 유연성, 집계와 리포팅에 적합한 구조
12) Incremental Loading : 이전에 기록된 데이터 이후 추가되거나 변경된 데이터만을 추가로 기록하는 방식
13) 시계열 데이터(Time Series Data) : 시간 경과에 따라 측정된 데이터로 시간의 흐름에 따른 변화를 추적할 수 있도록 구성된 데이터

KOFICE
한국국제문화교류진흥원

4

2022 빅데이터 활용 한류 시장조사
최종보고서

화면 및 기능 구성

1. 대시보드 화면 구성
2. 관리 기능 구성

Ⅳ. 화면 및 기능 구성

1. 대시보드 화면 구성

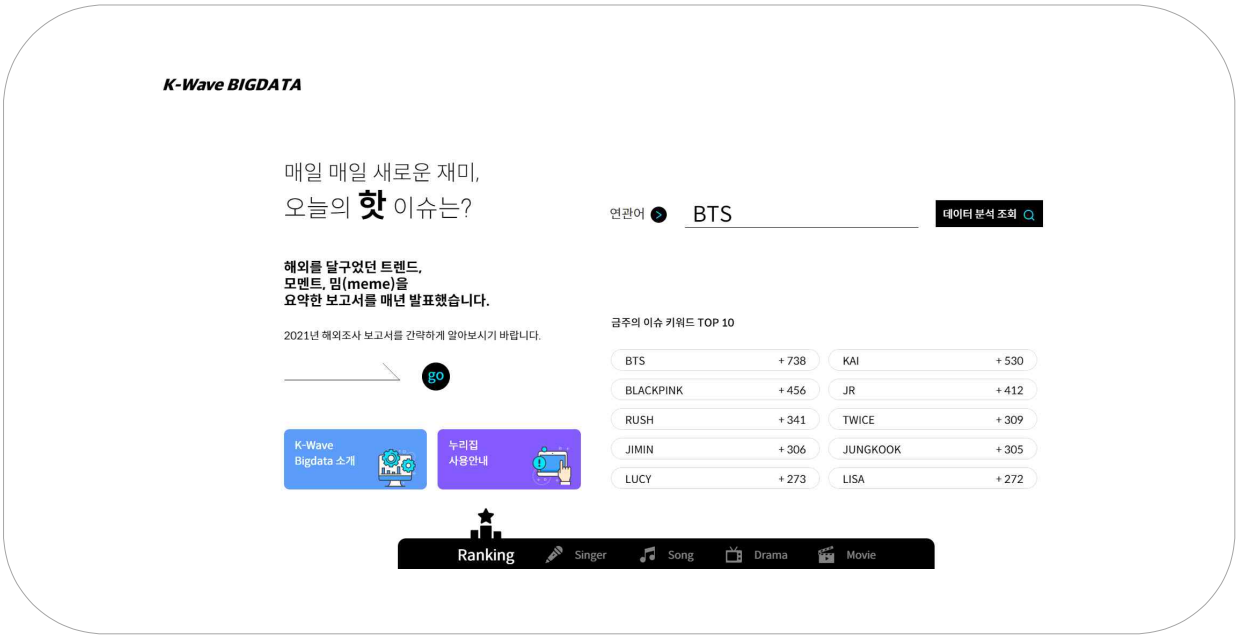
한류 대시보드는 연구자 및 관련 산업 종사자뿐 아니라 일반 대중에게 친숙하게 느껴질 수 있도록 하기 위해, 주제별 상위 콘텐츠 또는 아티스트를 대표하는 키워드 이미지를 노출하여 사용자의 관심을 끌고, 주제 간 전환이 슬라이드로 구성되어 있어 역동감을 느낄 수 있도록 구성하였다.

가. 랭킹 페이지

랭킹 페이지는 주간 이슈 키워드 TOP 10을 제공하고, 해당 키워드에 대한 연관어, 긍부정 지수, 언급량 데이터 분석 조회 화면으로 연결할 수 있는 기능을 제공한다. 이를 통해 사용자들은 주간 이슈 키워드와 관련된 다양한 정보를 확인할 수 있다.

또한, 랭킹 페이지에서는 대시보드 누리집에 대한 소개 및 사용 안내, 해외한류실태조사 페이지를 확인할 수 있는 버튼 링크를 제공한다.

그림 2. 랭킹 페이지



나. 인트로 페이지

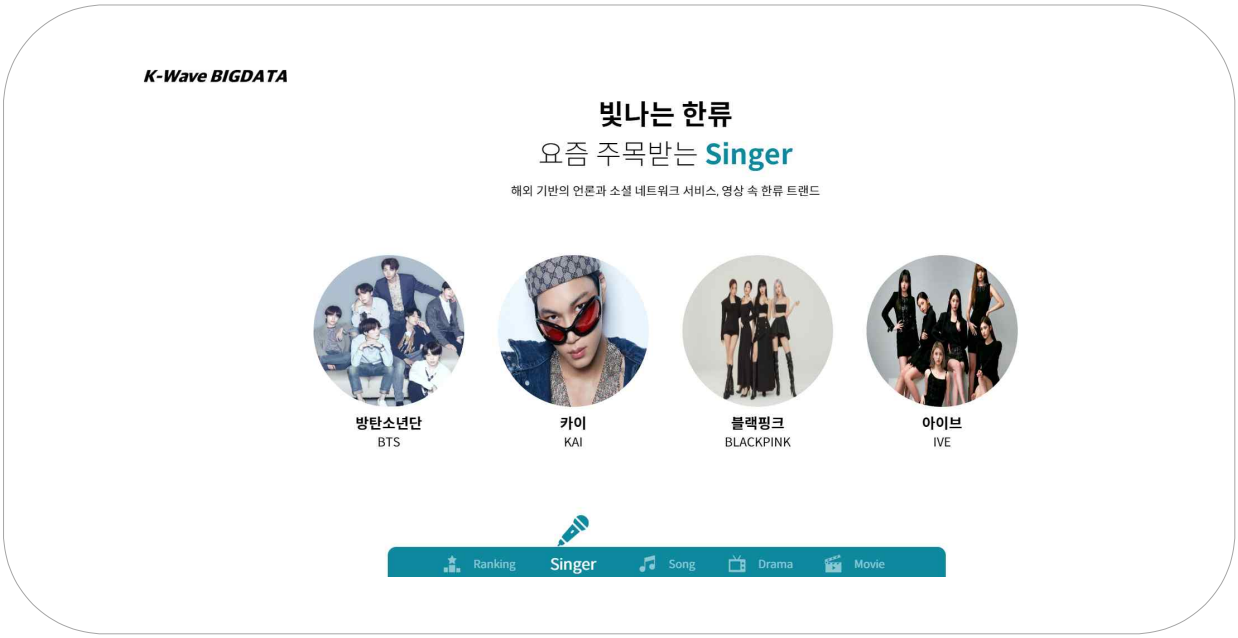
인트로 페이지는 가수(Singer), 노래(Song), 드라마(Drama), 영화(Movie)의 네 가지 주제로 분류되어 있으며 시간에 따라 주제 간 화면전환이 이루어지도록 되어있다. 하단의 내비게이션 바를 통해서도 원하는 주제를 직접 선택할 수 있다.

각 인트로 페이지에서는 분석을 통해 소셜 미디어에서 수집 및 분석된 영문 데이터가 많은 주요 키워드를 이미지와 함께 노출하고 있다. 노출되는 이미지를 클릭하면 해당 키워드에 대한 데이터 결과 페이지로 연결되어 사용자의 흥미를 대시보드로 연결하는 사용자 경험을 제공한다.

1) 가수(Singer)

가수(Singer) 주제에서는 K-POP 아티스트[14]로서 소셜 미디어 및 언론에서 그룹 또는 소속 멤버 개인에 대해 최근 발생한 영문 버즈가 많다고 분석된 키워드를 노출하고 있다. 이때 개인에 대해서만 언급된 버즈는 해당 그룹에 대한 버즈에 합산되지 않고 각각에 대해 텍스트상 명시된 경우에는 각각의 키워드에 중복적으로 합산하도록 하였다. 가수 주제에서 주로 나타나는 버즈는 주요 방송 출연 등 활동과 관련한 동영상, 게시글 등에서 주로 발생하고 있다.

그림 3. Singer 인트로 페이지



14) 국적이 대한민국이거나 대한민국 국적의 소속사를 두고 활동하는 경우를 기준으로 하였음

2) 노래(Song)

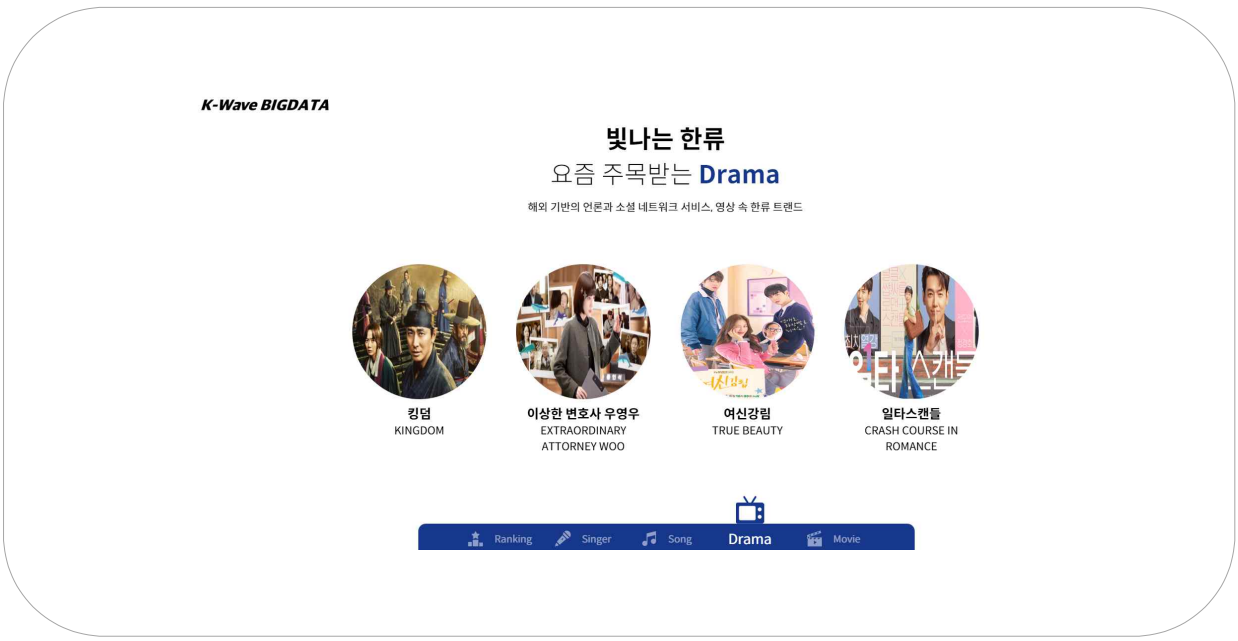
노래(Song) 주제에서는 K-POP 아티스트가 발매한 음원으로 최근에 소셜 미디어 및 언론에서 영문 버즈가 많이 발생했다고 분석된 키워드를 노출하고 있다. 노래와 관련된 버즈는 소속사에서 게시한 공식 뮤직비디오에서 관련한 댓글 버즈가 많이 발생하고 있다.

그림 4. Song 인트로 페이지

3) 드라마(Drama)

드라마(Drama)의 경우는 국내 지상파 및 케이블 방송사, OTT 플랫폼을 통해서 공개된 작품들을 기준으로 최근 영문 버즈가 많다고 분석된 키워드를 노출하고 있다.

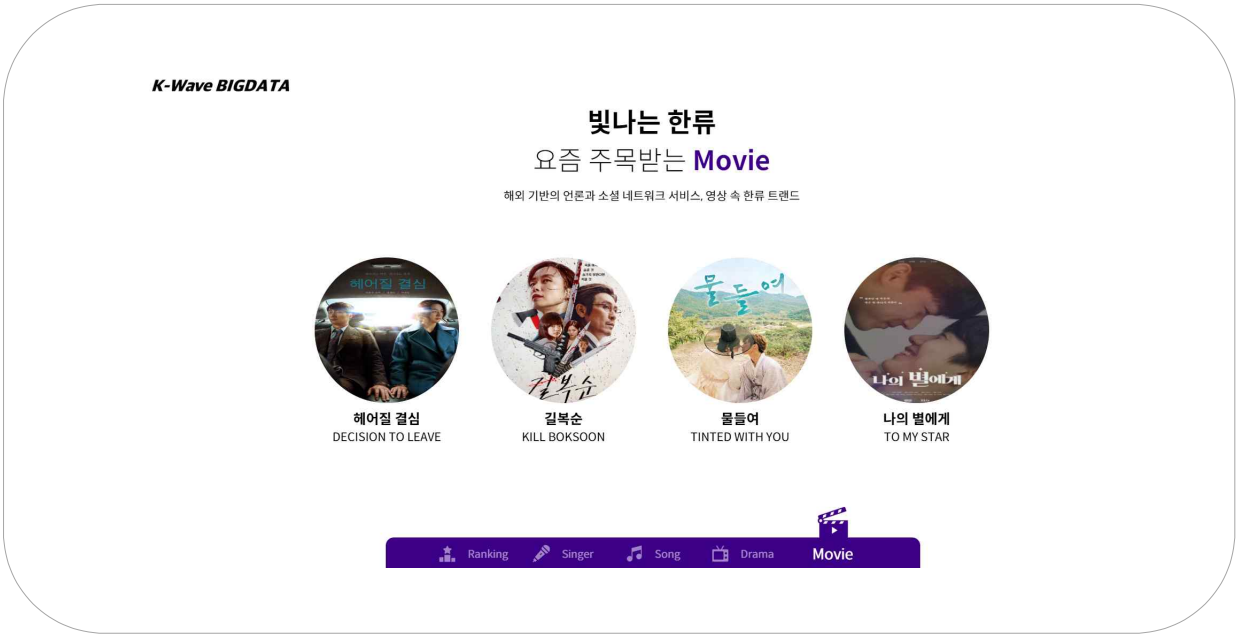
그림 5. Drama 인트로 페이지



4) 영화(Movie)

영화(Movie)의 경우는 국내 상영작 및 상영 예정작으로서, 극장에서 상영되었거나 OTT 플랫폼을 통해서 공개된 작품들을 기준으로 최근 영문 버즈가 많다고 분석된 키워드를 노출하고 있다.

그림 6. Movie 인트로 페이지



다. 대시보드(K-Wave BIGDATA) 소개

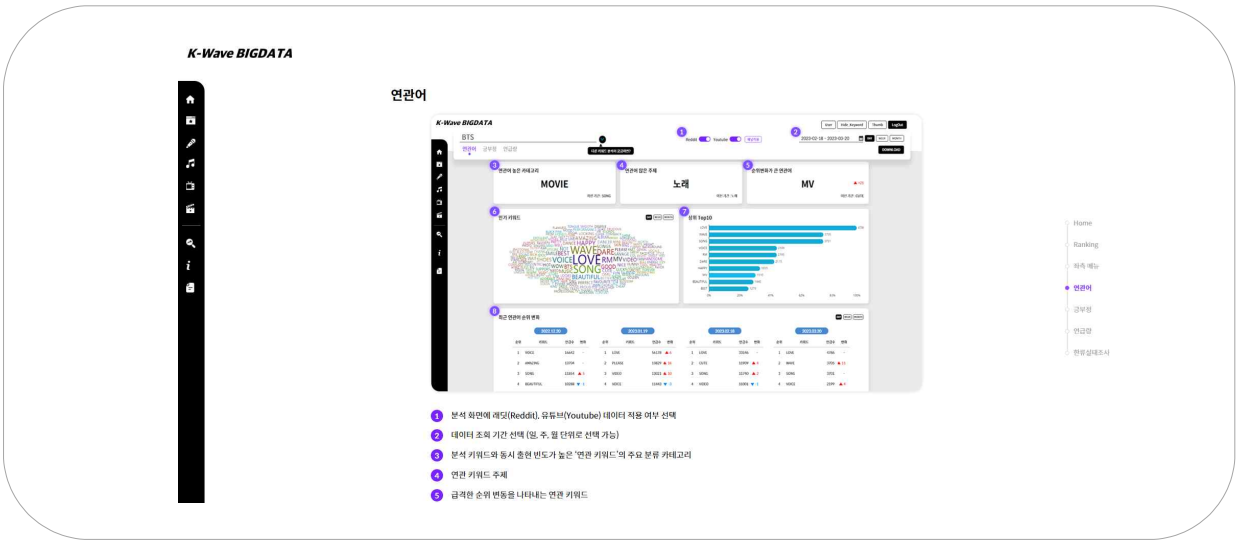
K-Wave BIGDATA 소개에서는 글로벌 소셜 미디어, 해외 언론, 리뷰 사이트 등 한류 데이터에 대한 수집 및 분석을 통해 한류 트렌드, 이슈를 사용자에게 동적으로 제공하고자 하는 누리집의 목적 및 수집에서 분석에 이르기까지 서비스 구성 개념과 특징 등을 소개하고 있다.

그림 7. K-Wave BIGDATA 소개

라. 누리집 사용안내

누리집에 대한 사용 안내 페이지를 두어 누리집의 각 페이지 및 대시보드 기능 및 차트에 대한 간략한 설명을 제공하여 사용자의 대시보드 이해를 돕는다. 우측에는 내비게이션 바를 두어 원하는 위치로 빠르게 스크롤 이동을 돕는다.

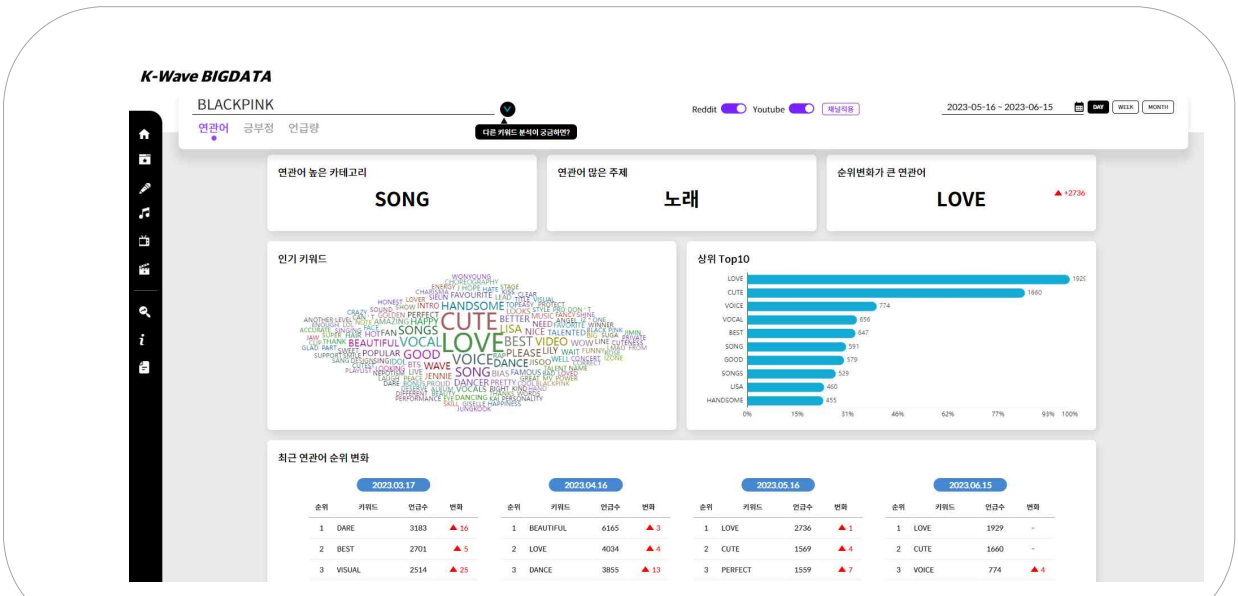
그림 8. 누리집 사용안내



마. 연관어 분석 대시보드

연관어 분석 대시보드에서는 분석의 대상이 되는 아티스트 및 콘텐츠 키워드에 연관되어 발생한 키워드에 대한 분석을 제공한다. 사용자가 선택한 기간에 발생한 연관어에 대해 워드 클라우드(WordCloud) 차트, 막대(Bar) 차트와 함께 기간별 연관어의 언급 수와 순위 변화 등을 제공한다.

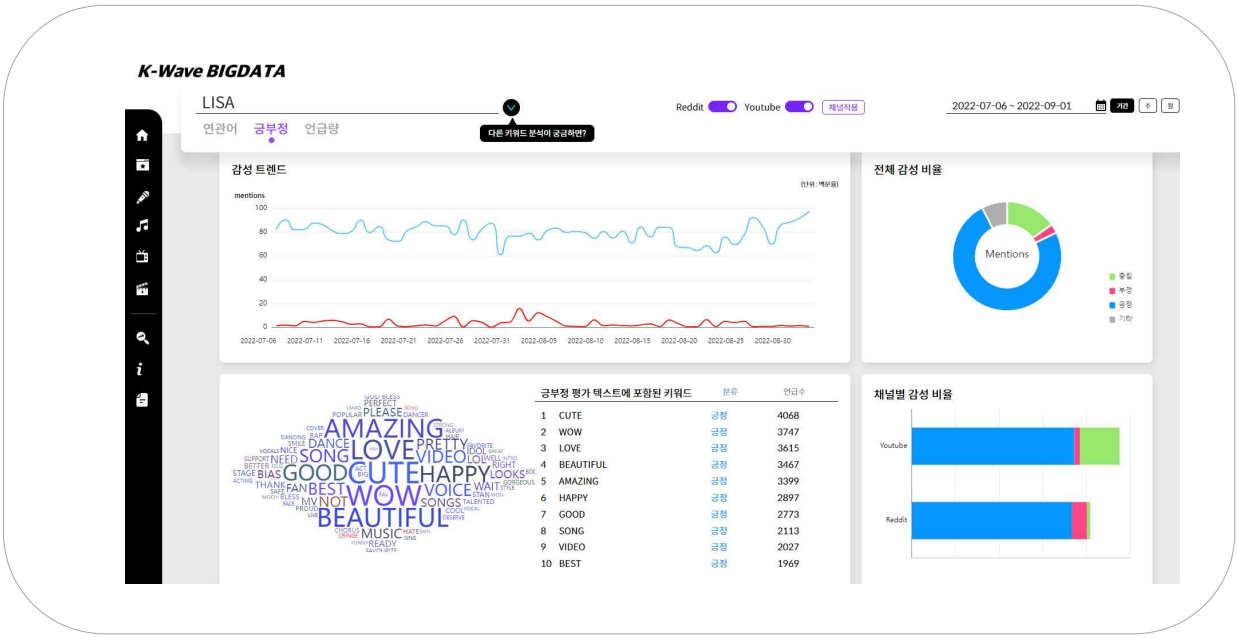
그림 9. 연관어 분석 대시보드



바. 긍부정 분석 대시보드

긍부정 분석 대시보드에서는 분석의 대상이 되는 아티스트 및 콘텐츠 키워드에 대한 긍부정 감성 평가를 제공한다. 기간별 긍부정률 추이 및 비율, 긍부정 감성어의 워드 클라우드 차트, 연관되어 발생한 키워드에 대한 분석을 제공한다. 사용자가 선택한 기간에 발생한 연관어에 대한 긍부정 키워드의 워드 클라우드 차트, 채널별 긍부정 누적 막대(Stacked Bar) 차트를 제공한다.

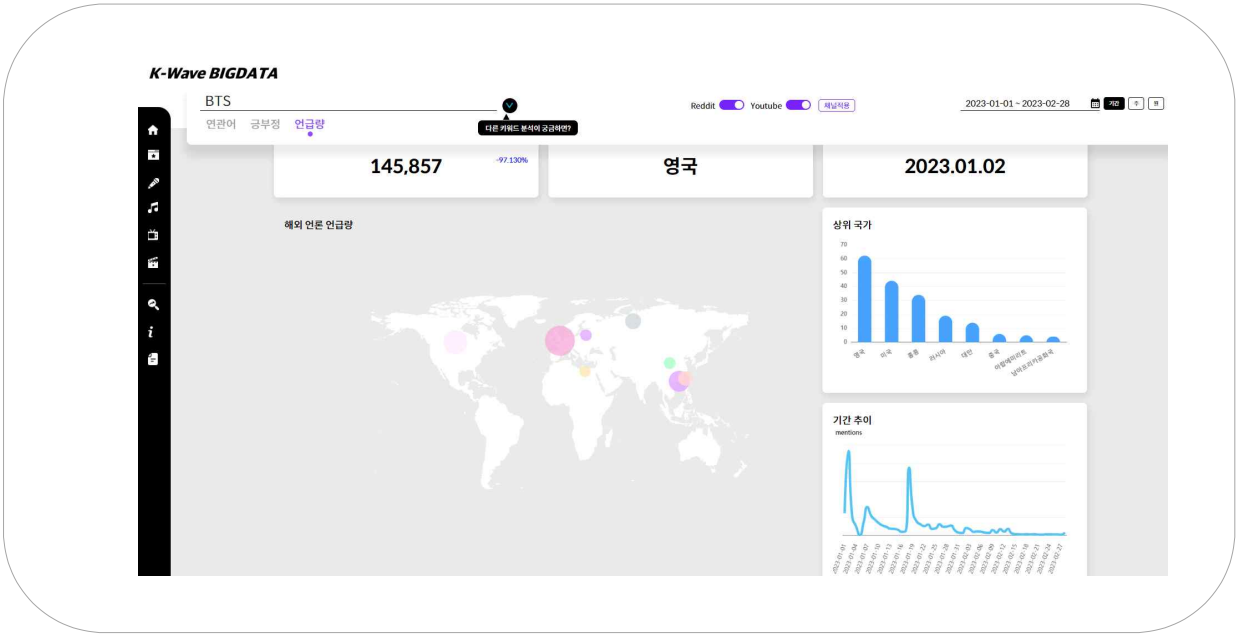
그림 10. 긍부정 분석 대시보드



사. 언급량 분석 대시보드

언급량 분석 대시보드에서는 분석의 대상이 되는 아티스트 및 콘텐츠 키워드에 대한 기간별 언급량 추이 및 해외 언론 언급량을 국가별 맵차트에 표현하여 제공하고 있다.

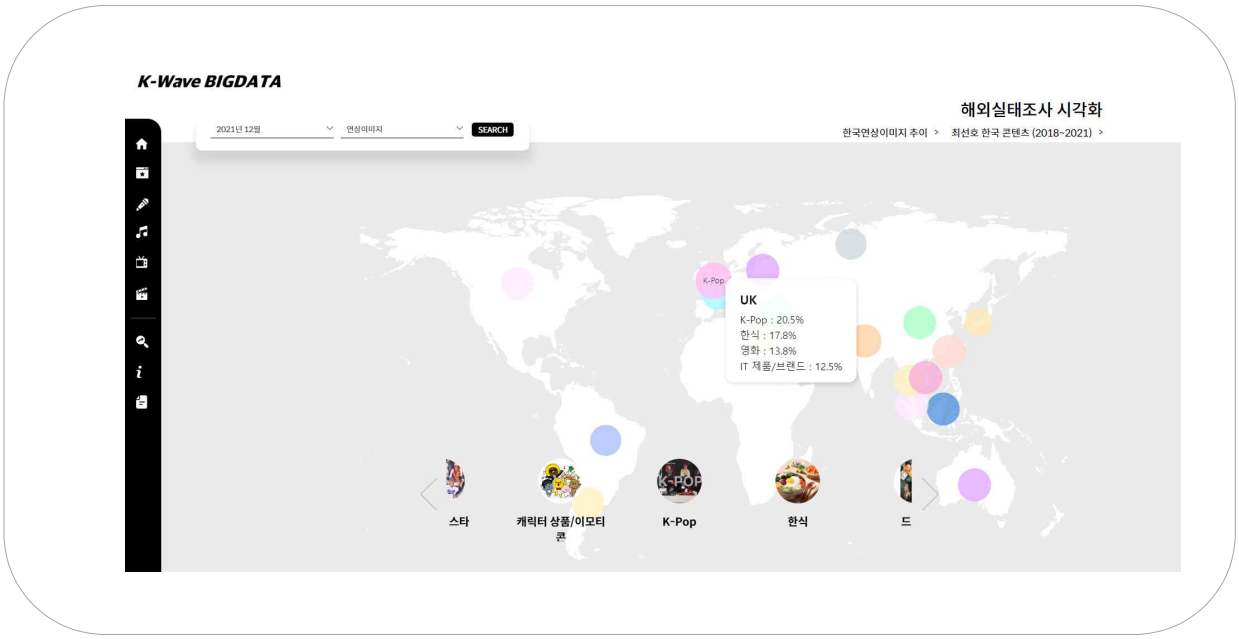
그림 11. 언급량 분석 대시보드



사. 해외한류실태조사 시각화

해외한류실태조사 시각화 화면에서는 2012년부터 진행한 해외한류실태조사의 주요 결과를 맵차트에 시각화하여 표현하였으며, 이후 추가적인 주제에 대한 확장성을 가지도록 설계하였다.

그림 12. 해외한류실태조사 시각화



2. 관리기능 구성

한류 빅데이터 대시보드 누리집의 랭킹 페이지 및 언급량 분석, 연관어 분석, 긍부정 분석의 주요 분석 데이터 조회 페이지는 일반 대중에게 공개되어 별도의 로그인 없이 사용할 수 있도록 하여 접근성을 높인 반면, 누리집의 관리자 페이지는 인트로 이미지 관리와, 불용어(Stopword)[15] 필터링 페이지, 관리기능에 접근할 수 있는 사용자 목록 관리 페이지는 로그인을 통해 접근 및 관리를 할 수 있도록 구성되어 있다.

가. 사용자 목록

사용자 목록 페이지는 관리자 페이지에 접근할 수 있는 사용자 정보를 관리하고, 사용자 계정의 추가 및 삭제 등의 기능이 제공된다.

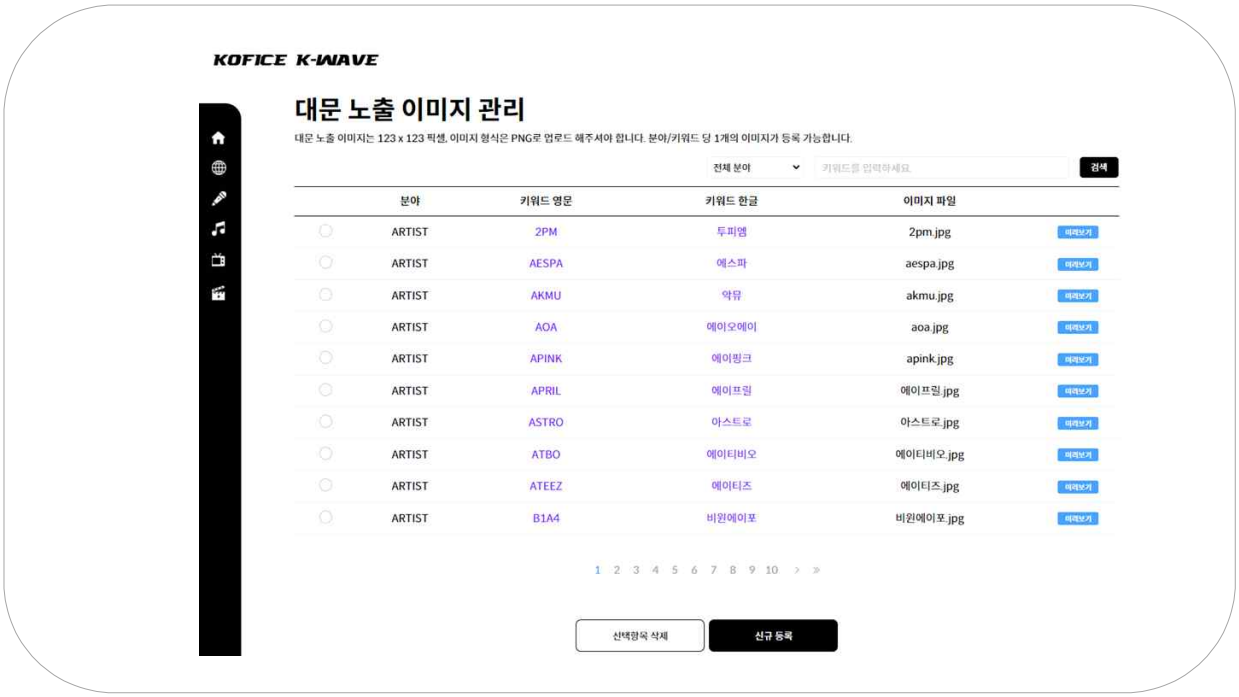
그림 13. 사용자 관리화면

15) 자연어 처리(NLP)를 통해 언어를 분석하는데 의미가 없거나 애매모호한 단어나 조사(postposition)로 불용어를 제거하여 무의미한 정보를 제거하여 중요한 정보에 집중할 수 있도록 하고, 처리해야 할 단어 수를 줄여 연산 성능을 개선하고 처리시간을 줄이는 작업이 필요함

나. 인트로 이미지 등록

인트로 페이지에 노출되는 주요 콘텐츠 및 아티스트 키워드와 관련된 대표 이미지를 등록 및 수정할 수 있는 관리기능이 제공된다.

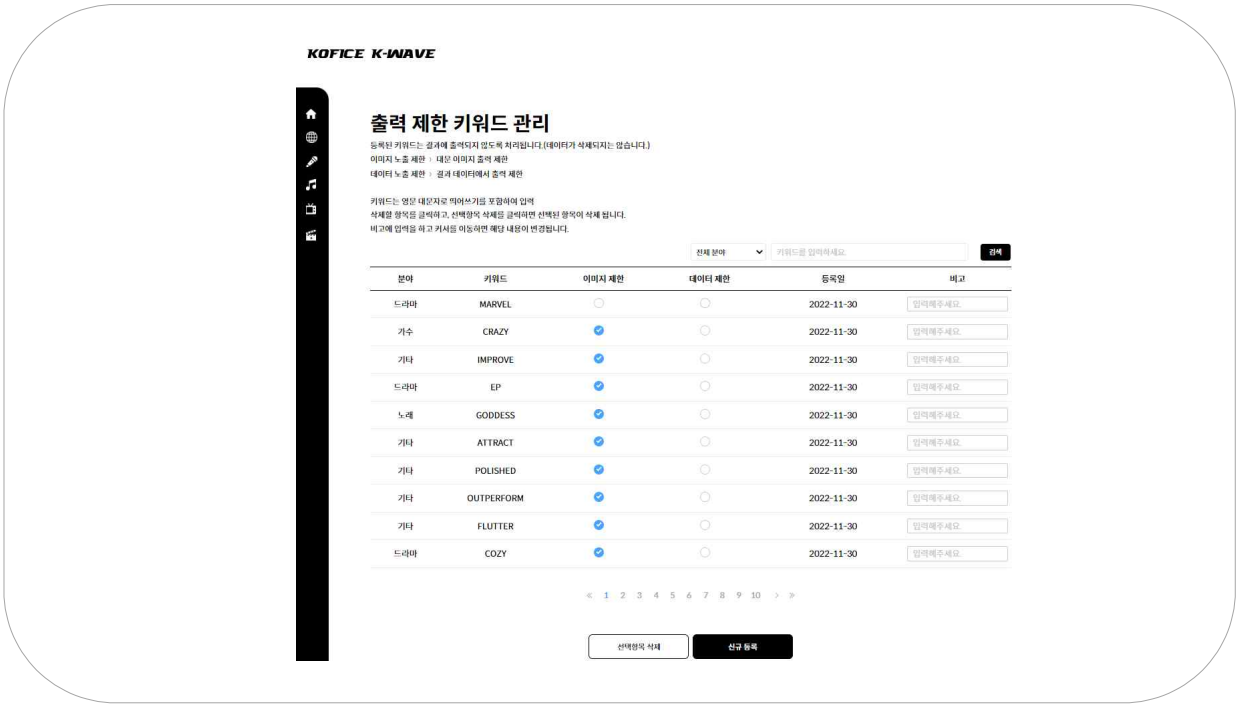
그림 14. 인트로 이미지 관리화면



다. 출력제한 키워드(불용어) 관리

비정형 텍스트 분류 분석 모델에서 사용자 사전을 통하여 불용어(Stopword) 제거 기능을 제공하고 동작하고 있지만, 한류 콘텐츠 및 아티스트, 음악 등과 같은 키워드와 관련된 어휘 등은 지속적으로 생성되거나 의미가 변화할 수 있다. 분류 모델은 야간 및 새벽의 유휴시간을 이용하여 작동하지만, 부적절한 키워드가 홈페이지에 노출되어 있는 경우 빠른 수정이 요구된다. 원인을 파악하고 분류 모델에 불용어를 추가하거나 분류 모델을 수정한 후 데이터를 다시 분석 및 적재하여 노출되어 있는 키워드에 대한 이슈 해결 대응을 하는 것은 적시성을 충족하지 못할 수 있어, 집계되어 적재되어 있는 데이터에서 필터링 기능을 추가하여 적시에 대응할 수 있도록 관리 기능이 제공된다.

그림 15. 불용어 필터링 키워드 관리화면



KOFICE
한국국제문화교류진흥원

5

2022 빅데이터 활용 한류 시장조사
최종보고서

한계점 및 향후 개선사항

1. 한계점

2. 향후 개선사항

Ⅴ. 한계점 및 향후 개선사항

1. 한계점

한류 빅데이터 대시보드는 일반적인 소셜 미디어 데이터를 수집하여 비정형 텍스트 분석 결과를 시각화한 것으로서, 일반적인 소셜 미디어 데이터의 특성에 따른 다음과 같은 한계점이 있을 수 있다.

가. 분석 목적 이외 데이터 다량 수집

소셜 미디어의 수집에 사용하는 API 및 해당 채널의 검색 기능은 각자의 추천 알고리즘을 통해 입력 키워드뿐 아니라 유사하거나 사용자들이 같이 검색하곤 하는 게시글을 출력하도록 되어 있다, 이는 원하는 데이터와 관련이 있는 데이터를 찾아주기도 하지만, 관련이 없는 데이터를 출력하기도 하며 입력 키워드에 따라서는 분석하고자 하는 목적 이외의 데이터가 다량 수집되는 원인이 되고 있다.

또한 검색하고자 하는 콘텐츠의 제목이 영어에서 일상적으로 사용하는 단어, 관용구이거나, 유사한 제목의 기존 콘텐츠 등이 존재하는 경우 검색 시스템에서 다량의 비관련 데이터가 출력되어 수집됨에 따라 수집된 데이터에 분석 대상이 아닌 데이터가 다량 포함되어 품질이 낮을 수 있으며, 다양한 방법을 통해 정제 처리를 하지만 비정형 데이터의 특성상 완전한 분리는 어려운 점이 있다.

나. 자연어 처리 및 분석의 어려움

소셜 미디어 데이터의 경우, 언어나 문장 구조, 철자 등의 오류, 약어, 은어, 이모티콘 등 비표준적인 표현이 많이 사용되어 일반 문서보다 문법적인 규칙 등이 통일되지 않아 자연어 처리를 하는 데에 있어서 신뢰도 높은 분류를 하기 어려운 특성이 있다.

다. 데이터 표본성 문제

소셜 미디어 데이터의 경우 API를 통하는 경우던지 검색 시스템을 통하는 경우던지 전체 데이터의 목록을 제공하지 않으며, 이에 따라 전체 데이터의 규모를 확인할 수 없고, 전체 데이터에서 균일한 품질의 샘플링을 보장하지 못하여, 전통적인 통계학적 표본의 요건을 충족하는 데이터 셋을 확보하기 어려운 문제가 있다.

이러한 특징은 최근의 빅데이터 분석에 수반되는 한계로서 빅데이터 분석 방법론에 따라 일부 한계를 인정하고, 충분히 큰 데이터 셋을 분석함으로 전체적인 데이터의 경향성에 대한 설득력을 확보하는 분석 관점을 따르고 있다.

라. 키워드 자유 검색 제공의 어려움

한류 빅데이터 대시보드는 연간 약 1억 건 이상의 텍스트 데이터를 수집 및 분석하는 시스템으로 초기 시스템 구축에 대한 예산 제약이 있는 상황에서 일반 사용자를 대상으로 웹 서비스를 제공하여야 하는 요구사항이 부여되었다. 이에 따라 시스템의 응답 성능을 향상시키기 위해 전체 키워드에 대한 자유로운 분석 결과 검색 기능을 제한하고, 대신에 사전에 ETL 프로세스를 통해 집계되어 요약된 리스트에서 상위에 해당하는 키워드를 선택하여 분석 결과를 확인하는 방식을 채택했다. 이에 따라 세밀한 분석이 어려운 점은 일반 사용자가 아닌 정밀 분석이 필요한 연구자나 실무자의 개별적인 요청에 의해 RawData를 제공함으로써 일부 해소될 수 있다.

2. 향후 개선사항

가. 키워드 자유 검색

현재 구축된 시스템은 시스템의 응답 성능을 향상시키기 위해 전체 키워드에 대한 자유로운 분석 결과 검색 기능을 제한하였다. 대신 사전에 ETL 프로세스를 통해 집계되어 요약된 리스트에서 상위에 해당하는 키워드를 선택하여 분석 결과를 확인하는 방식을 채택하고, 연구자 또는 현업 실무자의 요구에 따라 원본 데이터를 제공하는 형식으로 분석 요구를 해소하고 있다. 다만 이러한 경우 개별 기업 및 연구자의 분석 역량이 필요한 부분이 있어 향후에는 누리집 시스템에서 키워드 자유 검색 기능이 요구될 것이라고 생각되며, 이러한 자유 검색 제공의 제약은 향후 검색 엔진 시스템의 추가 또는 하드웨어 클라우드 구성의 확장을 통해 해소할 수 있을 것으로 예상한다.

나. 분석 주제 확장

2022년 사업에서는 음악, 영화, 드라마를 주제로 선정하여 한류 콘텐츠와 관련한 버즈를 분석하였다. 하지만 최근 이외에도 웹툰, 게임 등의 다양한 장르가 주목받고 있어 향후 분석 주제를 확장하여 나아가야 할 것으로 생각된다.

다. 문화 융복합에 따른 구분

한류 콘텐츠와 관련한 버즈를 수집 및 분석함에 있어서 문화의 융복합 현상에 따른 타 문화와의 결합이 나타나는 경우, 한류 콘텐츠로 분류할 수 있을지에 대한 문제가 생긴다.

K-POP 그룹에는 대만, 중국, 일본, 태국 등 타국 국적의 연예인이 포함되어 있고, 드라마와 영화, 웹툰 등은 그 시나리오의 저작권이 판매되어 해외에서 제작되거나, 한국인 작가가 해외 플랫폼에 웹툰을 연재하고, 국내에서 개발된 웹툰 플랫폼이 해외에 서비스되는 등, 그 생산 주체적인 측면이나 유통의 권한 측면 등에서 국가, 민족 등의 경계를 넘나드는 경우가 일반화 되어가고 있다.

향후 프로젝트에서 이러한 문화 융복합에 따라 생성되는 다양한 콘텐츠에 대하여 한류 콘텐츠로 정의할 것인지, 한류 이외의 콘텐츠로 분류할 것인지에 대한 고려가 필요하다고 판단된다.

KOFICE
한국국제문화교류진흥원

2022 빅데이터 활용 한류 시장조사
최종보고서

별첨

1. 참고문헌

참고문헌

유원준, 안상준 (2023).「딥 러닝을 이용한 자연어 처리 입문」. https://wikidocs.net/book/2155/

젠스 알브레히트, 싯다르트 라마찬드란, 크리스티안 윙클러 (2022).「파이썬 라이브러리를 활용한 텍스트 분석(Blueprints for Text Analytics Using Python)」.

카토 코타 (2018).「파이썬을 이용한 웹 크롤링과 스크레이핑」.

한국콘텐츠진흥원 (2022).「2022 대한민국 게임백서」.

한국콘텐츠진흥원 (2022).「2022년 상반기 및 연간 콘텐츠 산업 동향분석」.

한국콘텐츠진흥원 (2022).「2022 애니메이션 산업백서」.

한국콘텐츠진흥원 (2022).「2022 해외 콘텐츠 시장 분석」.

2022 빅데이터 활용 한류 시장조사
최종보고서

발 행 인 정길화(한국국제문화교류진흥원 원장)
발 행 처 한국국제문화교류진흥원
조사수행기관 ㈜미소정보기술
발 행 일 2023년 7월 31일
한국국제문화교류진흥원
(03920) 서울특별시 마포구 성암로 330 DMC첨단산업센터 A동 203호
전화 (02) 3153-1779
팩스 (02) 3153-1787
http://www.kofice.or.kr