하둡을 이용한 내용기반 음악 검색 시스템 설계

정형용 김준형 박현민 이정준 ○

한국산업기술대학교 지식기반기술·에너지대학원 컴퓨터공학과

Jung21w@gmail.com rha99@hanmir.com tnfusdlekt@lycos.co.kr jjlee@kpu.ac.kr

The Design of Content-based Music Search System Using Hadoop

Hyoungyong Jung JunHyoung Kim Hyunmin Park Jeongjun Lee

Department of Computer Engineering Graduate School of Knowledge-base Technology and Energy Korea Polytechnic University

요 약

음악은 인류의 대표적인 예술로서 오랜 세월동안 사랑을 받아왔다. 그 오래된 세월만큼이나 인류가 만들어온 음악의 수는 방대하다. 방대한 음악이 IT기술의 발달과 인터넷의 확산을 통하여 온라인 음악시장을 형성하였고 음악은 디지털 음원으로 관리되게 되었다. 이러한 디지털 음원을 효과적으로 검색하기 위한 방법은 많이 연구되었다. 그리고 검색을 도와줄 대량의 디지털 음원 자료들을 저장하고 관리하는 기법에 관한 연구가 필요하다. 본 논문에서는 대용량 자료를 처리하는 기술로 관심 받고 있는 하둡을 통하여 이 문제를 연구하였다. 하둡의 맵리듀스, HDFS 그리고 HBase를 이용하여 음악 내용기반검색을 설계하였다. 본 시스템은 음악 검색 시스템을 관리하고 유지하는데 있어서 컴퓨팅자원을 절약함으로써 비용을 절감 효과를 얻을 수 있다.

1. 서 론

IT기술의 발달과 인터넷의 확산으로 디지털 음원을 기반으로한 온라인 음악시장이 성장하는 추세이다[1]. 국내 주요 음악 서비스 업체들의 음원보유량은 180만곡 이상이며 인터넷 환경에서 다양한 음악 콘텐츠로 제공되고 있다[1]. 이러한 온라인 음악시장은 그 규모가 2009년 4,201백만달러에서 2014년에는 8,102백만 달러로 14.0%정도 성장할 것으로 전망된다[1].

온라인 시장의 발전으로 디지털 음원의 양이 방대해진 만큼 효과적으로 검색하고 관리할 수 있는 요구가 증가하게 되었다.

음악 검색을 효과적으로 검색하기 위해 다양한 연구가 진행중이며 이 연구를 크게 분류하면 메타정보와 태그를 이용하여 검색하는 키워드 검색[2,3]과 내용기반의 검색이다. 키워드 검색은 곡의 정보가 정리되어있고 사용자가 검색하고자 하는 곡의 정보를 잘 아는 경우에 유리할 수 있다. 그러나 음원의 수가 많은 경우에는 곡 정보와 키워드가 많이 겹쳐 중복되는 경우가 있으며 사용자가 반대로 곡의 정보를 원하는 경우에도 부적합하다.

본 논문에서는 이를 검색하기 위한 방법으로 내용기반검색 방법 중에 하나인 Shazam 의 핑거프린팅 알고리즘을 사용한다[4]. 내용기반검색은 라디오에서 나오는 노래, 벨소리, 허밍 등 음성을 통해 검색하는 기법으로 곡의 정보를 몰라도 곡의 일부분을 이용하여 곡을 검색할 수 있다[4-10]. 내용기반검색 기법 중에서도 Shazam 의 핑거프린팅 알고리즘은 여러 플랫폼에서 사용되고 있으며 대표적인 음악 검색 알고리즘이다[4].

위와 같은 음악 검색 알고리즘을 대량의 디지털 음원에 적용시키기 위해서는 높은 컴퓨팅자원과 비용이 요구되며 대량의 음원에 대한 검색과 데이터베이스를 유지하기 위해서도 고성능의 장비가 요구된다. 그에 대한 검색기술과 관련된 연구는 그 동안 많았지만 이를 저장하고 유지하기 위한 연구는 부족하다.

본 논문에서는 음악 검색 시스템을 효율적으로 관리하기 위한 방법으로 클라우드 컴퓨팅 기술 중에 하나인 하둡(Hadoop)을 사용한다. 하둡은 다양한 서브프로젝트들을 가지고 있다. 그 중에서 본 시스템은 맵리듀스, HDFS, HBase를 사용하여 설계하였다. 맵리듀스는 맵과 리듀스로 이루어진 분산 컴퓨팅 기술로 본 시스템에서 음원의 핑거프린트정보를 수정 및 검색하는데 사용하고 HDFS는 하둡 분산 파일시스템으로 음원의 핑거프린트 정보를 저장하며 분산데이터베이스를 지원하는 HBASE를 통하여 관리되게 된다[10-14].

2장에서는 기존의 음악 검색 방법과 대용량 데이터 관리 기술을 살펴본다. 그리고 3장에서는 본 시스템의 요구사항을 나열하고 4장에서는 요구사항을 만족시키는 설계를 기술한다. 마지막으로 5장에서 본 논문의 결론 짓는다.

2. 관련연구

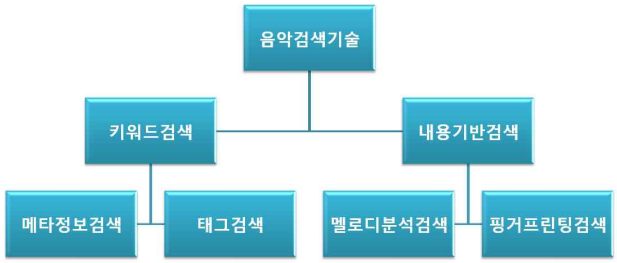
2.1 음악 식별 기술

음악식별기술은 사용자가 원하는 음악을 정확하게 찾아주는데 목적을 두고 있다. 음악식별기술은 대표적으로 크게 두 분류로 나눌 수 있다. 첫 번째 방법은 사용자가 입력한 키워드

를 곡의 태그정보 또는 메타정보와 비교하여 검색하는 방법이다[3,4]. 이는 일반적인 검색 포털사이트 등에서 사용되는 방법과 유사하다[15]. 곡에 메타데이터들을 데이터베이스에 저장하고 사용자의 검색 키워드를 데이터베이스에서 검색하여 그 결과를 사용자에게 보여주는 방식이다. 음원의 양이 많을 경우 메타데이터들의 정보가 겹치는 경우가 많기 때문에 검색하고자 하는 음악을 찾는데 어려움이 있어 본 논문에서는 사용하지 않는다. 다른 하나는 내용기반 검색기술이다. 내용기반 검색기술은 멀티미디어 검색에 사용되는 유용한 기술로 데이터의 속성을 자동으로 추출하고 이를 기반으로 검색하는 기술이다.

음악에서는 소리를 통하여 내용기반 검색을 한다. 멜로디를 형식이 있는 문자열로 변환하여 비교분석하는 방법[6-9]과 소리의 핑거프린트로 비교분석하는 방법[4,5,10]이 있다. 멜로디를 이용하여 검색하는 방법은 허밍으로 검색, 비슷한 느낌의 음악찾기등 다양한 방면에서 사용될 수 있지만 음악표기정보 및 음악 설명정보등을 가지고 있는 MIDI 형식이 아닌 경우에는 제한을 갖는다. 소리의 핑거프린트는 사람의 유일한 핑거프린트처럼 중복되지 않는다. 그렇기 때문에 음악을 정확하게 찾아주고 플랫폼에 영향을 받지 않아 본 시스템에 적합하다.

음악 검색 시스템을 정리하면 다음 (그림 1) 과 같다.



(그림 1) 음악 검색기술 분류

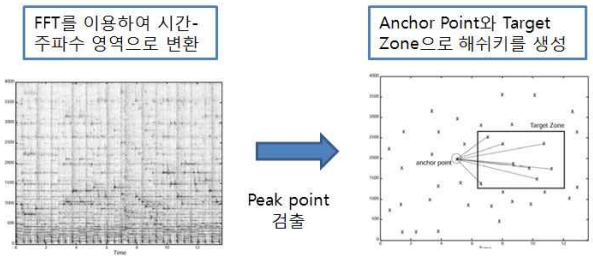
2.2 Shazam 핑거프린팅 검색기술

핑거프린팅 검색기술은 음악인식을 위한 음원의 핑거프린트된 정보를 데이터베이스에 저장하고 이를 기반으로 임의의 음악을 분석하는 기술이다[4]. 다양한 소리들은 사람의 핑거프린트와 같이 주파수 변환 알고리즘인 FFT(Fast Fourier Transform)을 거쳐 유일한 핑거프린트를 갖는다. 이러한 핑거프린팅 기술은 방송모니터링, 음악식별 시스템, 저작권필터링 등에 활용된다[10].

핑거프린트를 이용한 검색 기술은 Shazam, Grace, Google 등이 보유하고 있다. 그 중에서도 본 시스템에서는 150여나라 5천만 명 이상이 사용하고 있는 Shazam의 핑거프린팅 기술을 사용하려고 한다[1].

Shazam의 핑거프린팅 알고리즘[4]은 다음과 같다. 오디오 신호를 주파수 변환알고리즘 중에 하나인 STFT를 이용하여 오디오 신호를 시간-주파수 영역으로 변환하고 매 프레임마다 정점(peak point)을 검출한다. 각 정점을 앵커점(Anchor Point)으로 하여 임의의 타켓지역(Target Zone)을 생성하여 타겟지역 안에 있는 정점과 시간 및 주파수의 차이 정보를 이용하여 해쉬키를 생성하고 시간을 해쉬의 값으로 하여 곡 정보와 함께 데이터베이스에 저장하게 된다. 음악 검색은 해쉬들을 스코어링 해서 가장 높은 스코어를 가지는 음악을 찾아 주게 된다.

(그림 2)는 해쉬키를 생성하는 과정을 보여준다.



(그림 2) Shazam의 핑거프린팅 알고리즘

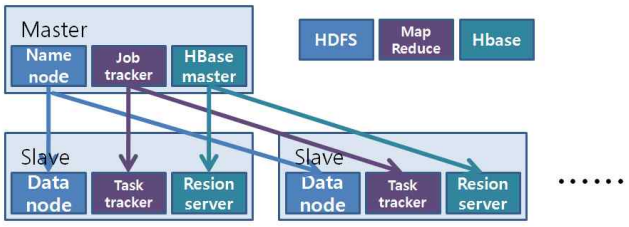
2.3 하둡(Hadoop)

하둡은 클라우드컴퓨팅 기술로 대량의 자료를 처리할 수 있는 컴퓨터에서 동작하는 분산 응용 프로그램을 지원하는 오픈 자바 소프트웨어 프로그램이다.[11,12] 클라우드 컴퓨팅 기술은 인터넷상의 가상의 서버를 통하여 IT서비스를 제공받는 환경으로서 본 시스템과 같이 대량의 음악을 검색하고 정보를 제공받는 서비스에 적합한 환경이다. 그 중에서도 하둡기술은 대용량 데이터를 처리하는 기술로 확장성과 비용절감 측면에서 본 시스템에 적합하다.

하둡은 분산컴퓨팅을 위한 인프라스트럭처 (infrastructure)와 그 산하의 관련 서브프로젝트들의 모음이라고 볼 수 있다. 대표적으로 HDFS와 맵리듀스의 조합으로 알려져 있지만 더 고수준의 추상화를 제공하기 위한 프로젝트들이 구성되어있다.

본 논문에서는 하둡에 서브프로젝트인 HDFS 와 맵리듀스 그리고 HBASE를 사용하여 시스템을 구성한다.

하둡은 Master-Slaves 구조이다. 하둡 프로젝트들은 Master-Slaves 구조를 갖는다. (그림 3)은 하둡에 Master-Slaves 구조와 서브프로젝트들과의 관계이다.



(그림 3) 하둡의 Master-Slaves 구조

2.3.1 HDFS

HDFS는 하둡의 분산파일시스템이다.[11] HDFS는 파일의 메타정보를 관리하는 Namenode와 실제 데이터를 저장하는 여러 대의 Datanode로 구성된다. HDFS는 데이터를 일정한

블록단위로 나누어서 관리하며 일부 Datanode에 장애가 생기더라도 문제가 생기지 않도록 중복되어 관리한다.

2.3.2 맵리듀스

맵리듀스는 데이터 처리를 위한 프로그래밍 모델이다[11,13]. 사용자로부터 맵리듀스에서 처리하도록 받은 일을 잡(job)이라고 하며 이는 분산환경에서 맞게 나누어진다. 나누어진 잡은 태스크라고 불린다. 맵함수와 리듀스함수를 이용하여 잡을 처리한다.

맵리듀스는 잡 수행을 상호조정하는 잡트래커와 해당 잡에 대한 분할된 태스크를 수행하는 태스크트래커로 이루어져 있다.

2.3.3 HBASE

HBASE는 HDFS에 구현한 분산 컬럼-기반 데이터베이스이다[11]. 대규모 데이터셋에 실시간으로 랜덤 엑세스가 필요할 때 사용할 수 있는 하둡 응용프로그램이다. HBASE는 마스터 노드가 하나 이상의 리전 서버를 조율한다. 리전서버는 HBASE의 테이블을 수평으로 분할한 데이터 베이스를 가지고 있는 서버이다. 이러한 HBASE 대량의 데이터를 분산 조회 및 업데이트를 빠르게 실행한다.

3. 연구동기

음악 검색을 위한 방법들은 다양하게 연구되었다. 사용자들이 주어진 환경에 맞게 효과적으로 검색하고자 하는 연구는 오래 전부터 진행되어왔다. 음원의 메타데이터나 태그를 통한 검색, 내용기반검색 등 검색을 위한 다양한 방법을 제공할 수 있게 되었다.

이중에서도 내용기반검색의 핑거프린트를 이용한 기술은 음악의 자료가 방대해진 만큼 새로운 검색기법으로 많은 관심을 받고 있다. 구글, 다음, 네이버 등 주요 포털사이트에서 위와 같은 방식으로 음악 검색을 서비스 하고 있다[1].

이렇게 다양한 음악 검색기법에 비해 검색을 지원하기 위한 자료들을 관리하는 연구는 부족하다. 음악의 수만큼이나 음악 검색을 지원하기 위한 자료들의 양 또한 방대할 수 밖에 없다. 이를 고성능의 장비로 지원하려면 비용적인 측면에서 효율이 떨어질 수 밖에 없다.

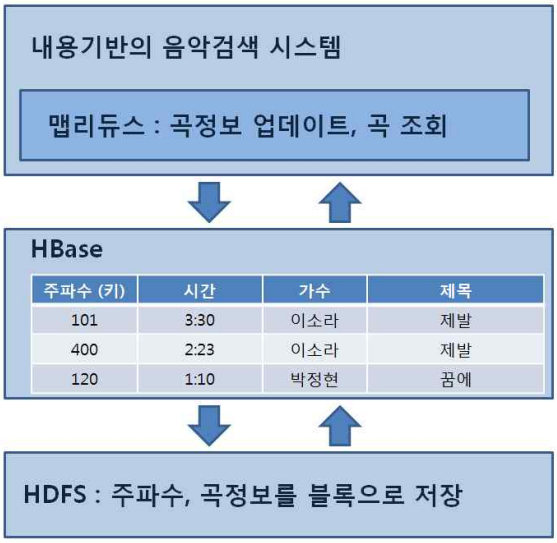
위와 같은 문제를 해결할 수 있는 방법으로 본 논문에서는 하둡을 제안한다. 하둡의 기술은 오픈 프레임워크로 비용이 저렵하며 고성능 장비를 값이 싼 여러 대의 범용장비로 대체한다는 개념에서 효과적으로 비용절감을 할 있다[11].

4. 구현

4.1 시스템 구조도

본 시스템은 검색 기술로 Shazam 핑거프린트 알고리즘을 응용하여 설계하였다. Shazam 의 핑거프린트 알고리즘에서 보편화 되어있는 cooleytukey FFT 방식으로 변형하여 설계하였다.

그리고 전체 시스템은 하둡을 이용하여 설계하였으며 하둡의 서브프로젝트인 맵리듀스, HDFS 그리고 HBase 를 사용하였다. 맵리듀스는 곡 정보 업데이트, 곡 조회 등에 사용되고 매칭을 위한 데이터베이스는 HBase에 저장되어 있다. (그림 4) 전체적인 시스템 구조를 보여준다.

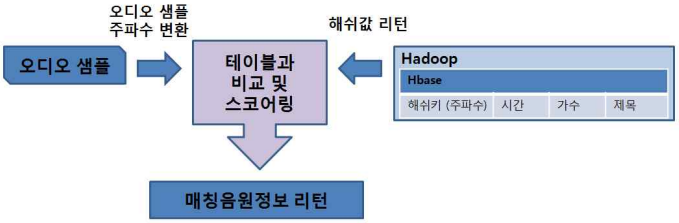


(그림 4) 시스템 구조도

4.1 오디오 검색

오디오 검색 절차는 다음과 같다. 오디오 샘플은 cooleytukey FFT방식을 통해 주파수가 강한 부분의 특징점들을 추출한다. 이 특징점들은 주파수의 크기와 곡의 시간정보를 가지게 된다.

아래 (그림 5)은 오디오 검색 과정을 보여준다. 곡 검색에 필요한 정보를 가지고 있는 HBase 에 오디오 샘플의 변환된 주파수 값을 키로 검색하여 그에 해당하는 시간 값들을 추출하게 된다. 이 추출된 값들의 시간 값과 오디오 샘플의 시간 값의 차이를 가지고 스코어링을 하게 된다. 같은 곡일 경우 이 차이는 일정하게 나타나게 될 것이며 일정 스코어링 이상이면 그 곡의 정보가 오디오 샘플과 일치하는 음원이 된다.



(그림 5) 오디오매칭 구성도

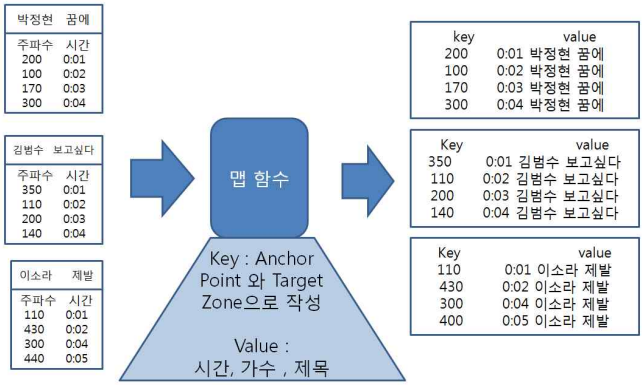
4.2 하둡을 이용한 업데이트

전 세계적으로 하루에도 수십 수많은 곡들은 업데이트 된다. 따라서 새로운 곡에 대한 데이터베이스 업데이트도 중요하다.

이 때도 맵리듀스를 통하여 HBase에 업데이트 시켜주게 된다.

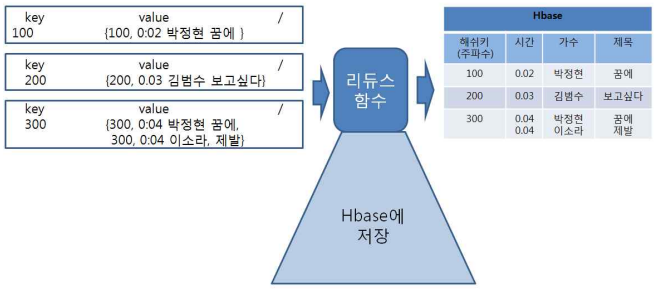
업데이트될 곡들은 Cooleytukey FFT를 사용하여 주파수 변환하여 HDFS에 파일로 저장한다. 각 주파수파일들은 맵을 거쳐 키와 값으로 쌍을 이루게 되는데, 여기서 키는 Anchor Point와 Target Zone 두 가지를 가지고 주파수와 시간차의 조합으로 해쉬키를 생성한다. 이 값이 맵리듀스에서의 키 값으로 설정된다. 맵 값은 주파수의 시간정보와 곡의 정보를 함께 갖게 된다.

아래 (그림 6)는 맵 과정에서 이루어지는 절차를 보여주고 있다.



(그림 6) 맵 과정

맵을 통하여 나온 값들은 리듀스 함수를 통하여 HBase에 업데이트 된다. (그림 7)은 이와 같은 과정을 보여준다.



(그림 7) 리듀스 과정

5. 결론

본 논문에서는 최근 주목 받고 있는 음악 내용기반 검색시스템을 하둡을 이용하여 설계하였다. 검색기법으로는 Shazam의 핑거프린팅 기술을 이용하였고 전체 시스템은 하둡기반으로 설계하였다.

핑거프린팅 기술을 사용함으로써 오디오를 통한 정확한 음악 검색이 가능하며 이를 하둡기반으로 설계함으로써 하둡이 갖는 장점을 본 시스템에도 가지도록 하였다. 즉 음악 검색에서도 사용할 수 있는 설계방법을 개발 하였다.

하둡을 사용하기 위해서는 맵리듀스 형태나 연산 구성이 필요하므로, Shazam 알고리즘을 이용하기 위한 시스템 구성방법을 제안하여 대용량 음악 검색 및 색인이 보다 활용적으로 가능하다.

추후 본 논문은 구현될 예정이며 구현을 통하여 구현 결과 및 특징등에 관한 내용을 보강할 예정이다

6. 참고문헌

[1] 이승재, 김성민, 김정현, 유원형 , “음악 서비스 및 관련 기술 동향”, 전자통신동향분석 제 26권 제2호 , 2011

[2] 김룡, 이광동, 성민선, 김영국 , “모바일 시스템에서 메타정보를 이용한 멀티미디어 콘텐츠 검색 기법”, 한국콘텐츠학회 , 2007

[3] Zhendong Zhao, Xinxi Wang, Qiaoliang Xiang, Andy M Sarroff, Zhonghua Li, Ye Wang, “Large-scale Music Tag Recommendation with Explicit Multiple Attributes”, MM’10 , 2010

[4] Li-Chun Wang , “An Industrial-Strength Audio Search Algorithm”, ShazamEntertainment,Ltd , ISMIR . 2003

[5] 윤원중, 박규식 , “잡음에 강인한 내용기반 음악 검색 시스템에 대한 연구”, 전자공학회논문지 , 2008

[6] 노승민, 황인준 , “사용자 질의 기반의 빠른 오디오 검색 기법”, 정보처리학회논문지 A , 2003

[7] 노승민, 이수철, 황인준 , “대용량 오디오 데이터베이스에서의 효율적인 오디오 검색”, SIGDB2003 , 2003

[8] 정명범, 고일주 , “오디오의 파형과 FFT분석을 이용한 대표 선율 검색”, 정보과학회논문지 소프트웨어 및 응용 제 34권 제 12호 , 2007

[9] 지정규, 오해석 . “디지털 음악정보 검색 시스템의 설계”. 97ICMDI . 1997

[10] Jaap Haitsma, Ton Kalker , “A Highly Robust Audio Fingerprinting System”, ISMIR , Oct. 2003

[11] White Tom, Cutting Doug , “Hadoop:The Definitive Guide”, O’REILLY , 2009

[12] 한재선 , “검색 플랫폼의 진화 : Google에서 Hadoop까지”, Search Day 2008 Spring , 2008

[13] Jeffrey Dean, Sanjay Ghemawat , “MapReduce: Simplified Data Processing on Large Clusters”, Google.Inc , Communications of the ACM, Vol.51, No. 1, pp. 107-113, 2008

[14] 조성환, 이승하, 이광진, 김양우 , “대용량 스펨 메일 처리를 위한 Hadoop 기반 분산 필터링 서비스 모델”, 한국인터넷 정보학회 학술발표대회 논문집 . pp. 165~168 , 2009

[15] 박소연 , “주요 포털들의 멀티미디어 검색 서비스 분석”, 한국문헌정보학회지 , 2010