내용 기반 이미지 검색을 위한 효율적인 분산 고차원 색인 기법

An Efficient Distributed High-Dimensional Index Structure for Contents-Based Image Retrieval

김 민 수*, 김 기 훈*, 송 희 섭**, 한 진 수*, 유 승 훈*, 안 지 환*, 박 주 영*, 복 경 수*, 유 재 수*†
*충북대학교 정보통신공학부,
**충북대학교 빅데이터학과

Minsoo Kim*, Gihoon Kim*, Heesub Song**, Jinsu Han*, Seunghun Yoo*, Jihwan Ahn*, Juyoung Park*, Kyoungsoo Bok*, Jaesoo Yoo*†
*Department of Information & Communication Engineering, Chungbuk National University
**Department of Big Data, Chungbuk National University

요약

다양한 디지털 기기 활용의 증가로 인해 멀티미디어 데이터가 증가됨에 따라 내용 기반으로 검색하는 기술이 연구되고 있다. 내용 기반 검색을 위해 멀티미디어에서 추출된 고차원 특징 벡터가 대용량이 되면서 고차원 데이터를 분산해서 관리하는 색인 기법이 필요하다. 본 논문에서는 대용량 멀티미디어 데이터에서 유사한 이미지를 검출하기 위한 분산 고차원 색인 기법을 제안한다. 제안하는 기법은 마스터/슬레이브 구조로 되어 있다. 마스터 서버의 색인 구조는 그리드 방식을 사용하여 검색 요청 시 탐색하는 노드를 감소시킨다. 슬레이브 서버의 색인 구조는 구 형태로 색인하여 범위 질의와 최근접 질의를 효율적으로 검색한다.

I. 서론

최근 핸드폰, CCTV 같은 디지털 기기 활용의 증가로 인해 멀티미디어 데이터가 급증하고 있다. 대용량의 멀티미디어 데이터에서 사용자가 원하는 데이터를 검색하기 위해 내용 기반으로 검색하는 기술이 활용되고 있다. 내용 기반 검색은 멀티미디어 내용을 대표하는 특징 벡터를 데이터베이스에 저장하고 사용자가 원하는 내용을 포함하는 멀티미디어 데이터를 검색하는 방법이다. 내용 기반 검색의 높은 정확성을 위해 멀티미디어 데이터에서 추출하는 특징 벡터가 증가함에 따라 이를 효과적으로 검색하기 위한 고차원 색인 기법이 요구된다. 그러나 대용량의 고차원 데이터를 단일 서버에서 저장할 경우 검색 성능이 저하되는 문제점이 있다. 이러한 문제점을 해결하기 위해 고차원 색인 구조를 분산으로 처리하기 위한 연구들이 진행되고 있다.

[1]에서는 공항 비디오 모니터 시스템에서 유사 이미지를 검색하기 위해 Distributed MVP-tree를 제안하였다. [2]에서는 동영상 데이터의 내용 기반 검색을 지원하기 위해 Hybrid Spill-Tree를 기반으로 확장된 색인 기법을 제안하였다. 기존 기법은 마스터 서버의 색인 구조가 M-Tree 기반의 기법을 사용하기 때문에 데이터가 밀집되면 영역이 중복되어 색인된다. 따라서 데이터가 밀집된 영역을 포함하는 질의가 요청되면 색인 구조에서 방문하는 노드들이 증가하게 된다. 또한, 질의 요청 시 마스터 서버의 색인 구조에서는 거리 계산을 수행하므로 색인 구조가 커질수록 마스터 서버의 부하가 증가한다.

본 논문에서는 대용량 멀티미디어 데이터에서 유사한 이미지를 검출하기 위해 고차원 데이터를 효율적으로 관리하고 검색하기 위한 분산 고차원 색인 기법을 제안한다. 제안하는 기법은 마스터 서버의 색인 구조를 그리드 기반의 기법을 사용하여 검색 요청 시 거리 계산량이 적기 때문에 질의를 빠르게 슬레이브 서버로 전달한다. 슬레이브 서버의 색인 구조는 참조점과 데이터의 거리를 B+-tree로 색인하기 때문에 임의의 위치에서 순차접근이 필요한 범위 질의와 최근접 질의를 효율적으로 처리한다.

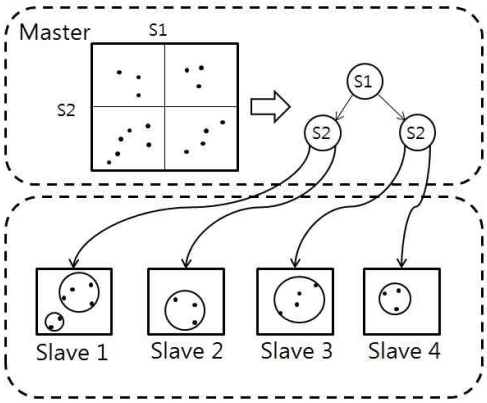
II. 제안하는 분산 고차원 색인 기법

2.1 전체적인 구조

제안하는 기법은 질의 요청 시 마스터 서버의 색인 구조에서 다수의 질의를 처리하기 위해 색인 구조에서의 계산량을 줄이고 슬레이브 서버의 색인 구조에서는 중복된 영역을 감소시켜 질의 처리 속도를 향상시킨다. 그림 1은 제안하는 기법의 전체적인 구조를 보여준다. 제안하

† 교신 저자 : yjs@cbungbuk.ac.kr
본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2016-H8501-16-1013, IITP-2017-2013-0-00680)과 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환[B0101-15-0266, (딥뷰-1세부) 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발]으로 수행하였음.

는 분산 고차원 색인 구조는 마스터(master) 서버와 슬레이브(slave) 서버로 구성된다. 마스터 색인에서는 거리 계산량을 감소시키기 위해 그리드 기법을 사용하였다. 슬레이브 색인에서는 영역을 중복해서 색인하지 않기 위해 데이터 분포를 고려하여 효율적으로 탐색한다.



▶▶ 그림 1. 전체적인 구조

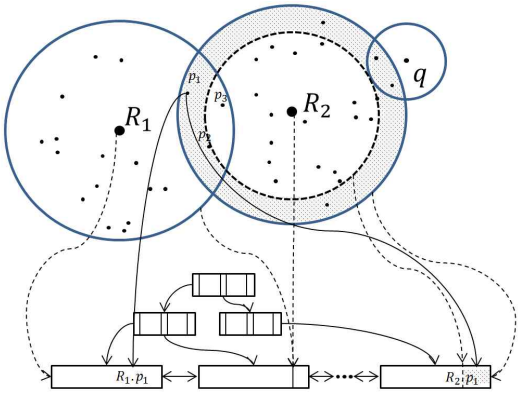
대량의 정적인 데이터를 색인할 때 데이터를 하나씩 삽입하게 되므로 색인 구축 속도도 저하되고 데이터가 삽입되는 순서에 따라 색인구조의 효율성에 영향을 미치게 된다. 따라서 제안하는 기법은 대량의 정적인 고차원 데이터를 빠르게 색인하기 위해 변형된 퀵정렬 기법을 사용한다. 변형된 퀵정렬 기법을 사용하기 위해서는 분할 전략이 있어야 한다. 분할전략은 어떤 차원을 기준으로 어떤 비율로 나눌 것인지를 결정한다. 제안하는 기법에서 분할 차원은 차원의 길이가 최대인 차원을 선택하고 분할 비율은 슬레이브 서버 수로 결정한다. 대량의 정적인 데이터를 분할전략에 따라 변형된 퀵정렬을 사용하여 색인한다.

2.2 마스터 색인

마스터 색인 구조는 질의와 거리 계산량을 감소시키겨 다수의 사용자 질의를 처리하기 위해 그리드 기법을 사용한다. 질의가 요청될 경우 마스터 색인에서는 영역 비교를 통해 해당 슬레이브 서버를 선택하기 때문에 거리를 비교하여 탐색하는 것보다 부하가 적다. 유사도 검색 질의는 구 형태로 구성되기 때문에 실제 질의를 포함하지 않은 슬레이브 서버에도 검색 요청이 될 수 있다. 하지만 슬레이브 색인 구조는 구 형태로 되어 있기 때문에 참조점과 거리 계산을 통해 슬레이브 색인의 탐색 유무를 결정한다.

2.3 슬레이브 색인

유사도 검색 질의는 구 형태로 이루어져 있기 때문에 효율적인 검색을 위해 구 형태의 색인 구조를 사용한다 [3]. 또한, 영역을 중복하여 색인하는 것을 줄이기 위해 데이터의 분포로 참조점의 위치를 결정한다. 그림 2는 슬레이브 색인 구조에서 질의 처리를 나타낸다. 고차원 데이터를 단일차원으로 변환하기 위해서는 기준이 되는 참조점이 존재해야 한다. 초기에 대량의 정적인 데이터의 분포를 기반으로 참조점의 수와 위치를 계산한다. 단일차원 값은 고차원 데이터와 참조점과의 유클리드 거리(euclidean distance)를 이용하여 계산한다. 계산된 단일차원 값들에서 범위 질의와 최근접 질의를 처리하기 위해 임의의 위치에서 순차적으로 값들을 탐색하는 기법이 필요하다. 따라서 계산된 단일차원 값들은 빠르게 검색하고 순차적인 접근이 가능한 B+-tree를 사용하여 색인한다. 이 때, 그림 2의 데이터 p_1과 같이 두 개 이상의 참조점이 겹치는 영역 데이터는 빠른 질의 처리를 위해 영역에 포함되는 모든 참조점에 중복하여 저장한다. 질의 q가 요청되면 모든 참조점과 질의 q의 거리를 계산하여 탐색할 참조점을 결정한다. 탐색할 참조점 R_2가 결정되면 질의가 포함하는 영역을 B+-tree에서 찾고 질의와 실제 데이터의 거리를 계산하여 결과를 마스터 서버에 전달한다.



▶▶ 그림 2. 슬레이브 색인 구조에서 질의 처리

Ⅲ. 결론

본 논문에서는 대용량 고차원 데이터를 효율적으로 관리하고 검색하기 위한 분산 고차원 색인 기법을 제안하였다. 제안하는 기법은 사용자의 질의를 받는 마스터 서버의 부하를 줄이기 위해 그리드 기법을 사용하였다. 또한, 슬레이브 서버에서는 범위 질의와 최근접 질의를 효율적으로 처리하기 위해 고차원 데이터를 저차원으로 변환하여 관리한다. 추후 연구로는 제안하는 기법에서 효율적인 질의 처리 기법을 연구할 것이다.

■ 참 고 문 헌 ■

[1] H. Cheng, W. Yang, R. Tang, J. Mao, Q. Luo, C. Li, and A. Wang, "Distributed indexes design to accelerate similarity based images retrieval in airport video monitoring systems," Proc. FSKD, pp.1908-1912, 2015

[2] 최현화, 이미영, 김영창, 장재우, 이규철, "대용량 데이터의 내용 기반 검색을 위한 분산 고차원 색인 구조," 정보과학회 논문지, 제 37권, 제 5호, pp. 228-237, 2010

[3] C. Yu, B. C. Ooi, K. L. Tan, and H. V. Jagadish, "Indexing the distance: An efficient method to knn processing," Proc. VLDB, pp. 421-430, 2001.