고차원 멀티미디어 데이터에 대한 내용기반 검색을 위한 인덱싱 방법들의 성능 평가

문주선 최정훈○ 낭종호
서강대학교 컴퓨터공학과
serenity0605@mlneptune.sogang.ac.kr, drumist@mlneptune.sogang.ac.kr

A Performance Evaluation of Indexing Methods for Content-based Retrieval of High Dimensional Multimedia Data

Joosun Moon Jeonghoon Choi○ Jongho Nang
Department of Computer Science and Engineering, Sogang University

멀티미디어 데이터베이스의 효과적인 내용 기반 검색을 위한 많은 색인 방법들이 연구되어왔지만 정작 동일한 데이터 집합과 동일한 평가 기준으로 서로 다른 검색 방법들의 성능을 분석한 실험은 이뤄지지 않았다. 본 논문에서는 기존의 대표적인 색인 방법들을 구현하고 공통의 데이터 집합에 대한 색인 검색을 여러 성능 측정 기준에 따라 분석함으로써 각 색인 방법들의 특징 및 성능을 객관적으로 평가하였다. 향후 본 논문에서 실험한 결과들을 이용하면 특정 데이터 집합에 효과적인 색인 방법을 선택할 수 있을 것이다.

1. 서론

오늘날 고차원 데이터를 위한 빠르고 효과적인 검색방법들이 많이 소개되고 있지만 동일한 데이터 집합과 동일한 평가 기준으로 서로 다른 검색 방법들에 대한 성능 분석은 이뤄지지 않았다. 본 논문에서는 기존의 멀티미디어 데이터 색인 검색을 위해 제안된 대표적인 색인방법들을 구현하고 공통의 데이터 집합에 대한 색인 검색 성능을 여러 측정 기준에 따라 분석함으로써 각 색인 방법들의 특징 및 성능을 객관적으로 평가한다. 이를 위해 본 논문에서는 색인 방법에 따른 평가 기준을 정하고, 구현상의 이슈 및 실험 결과를 분석, 평가하며 끝으로 각 색인 방법의 성능과 특징에 대한 객관적인 평가를 서술한다.

2. 고차원 데이터 검색방법들의 성능비교 평가 기준 및 데이터 집합

평가 기준은 크게 1)공간적 평가기준, 2)시간적 평가기준, 3)검색 성능 평가기준으로 세분화하였다. 1)공간적 평가기준은 각 색인 방법들을 통해 만들어진 색인의 크기를 비교하는 것이며, 색인된 파일의 크기를 KBytes단위로 N(Total Number of Dimension), M(Total Number of Objects)의 값을 비교하였다. 2)시간적 평가기준은 색인을 만드는 데 걸리는 시간(Indexing time(ms))과 색인을 검색하는 시간(Searching time(ms))으로 나누어 평가하였다. 3)검색 성능 평가기준은 색인을 이용한 최종 결과의 값들과 실제 순차 검색을 통해 얻어진 결과를 비교하여 정확성을 평가한 것으로 Precision/Recall 및 ANMRR(Averaged Normalized Modified Retrieval Rate) 방법을 이용하여 측정하였다.

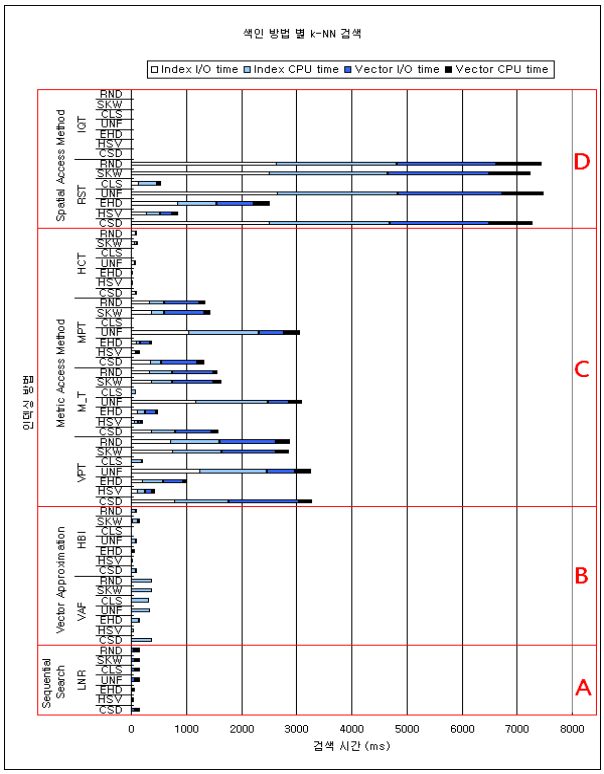
3. 각 색인 방법 별 구현 및 이슈에 대한 해결방안

R^*-tree의 경우 R-tree의 단점인 MBR의 겹침 부분을 최소화시킴으로써 검색 성능의 향상을 도모하는 방법으로써, 성능 향상을 위해선 Directory Rectangle이 서로 겹치지 않으면서 각 Directory Rectangle의 영역 및 둘레를 최소화시켜야 하는데, 이 세 가지 방법이 서로 다를 경우 겹침을 최소화시키면 된다[1]. HBI(Hierarchical Bitmap Imdexing)는 비트맵의 개수에 따라 검색 성능의 차이가 있는데, 각 데이터 집합마다 평균적으로 최적의 검색 시간을 보여주는 적정 비트맵 개수가 존재하므로 몇 번의 실험을 통해서 최적의 검색 시간을 보여주는 적정 비트맵의 개수를 설정할 수 있다. M-tree는 Routing Object를 정할 때 I/O비용과 CPU계산비용을 고려했을 때 Random한 방법을 이용하는 것이 가

장 효과적임을 알 수 있다. HCT(Hierarchical Cellular Tree)는 대용량 멀티미디어 DB의 색인을 위해 고안된 방법이지만 색인화 과정에서 너무 많은 계산비용을 초래하므로, 전체 tree의 중간까지는 preemptive cell-search방식을, 하부는 기존의 M-tree 방식을 사용하면 된다. IQ(Independent Quantization)-tree는 Hybrid방식으로써 Quantization을 위한 bit의 개수를 고정하거나 Flexible하게 결정하느냐에 따라 검색에 필요한 시간의 차이가 커짐을 알 수 있다.

4. 실험 및 결과 분석

고차원에서의 멀티미디어 데이터의 검색 성능을 비교하기 위해 순차검색과 6가지 색인 방법 별 복잡도를 정의하였으며, k-NN 검색을 위해 다양한 색인 방법들을 사용하여 각각의 속도를 측정하였으며 결과는 <그림 1>과 같다. 각 색인 방법들을 Sequential Search, Vector Approximation, MAM, SAM 4가지의 카테고리로 분류하였다. k값은 10으로 하였으며 오브젝트의 수는 35,000개, 차원의 수는 HSV와 EHD를 제외하고 모두 200개로 통일시켰다. Section A는 순차 검색으로써, 데이터 파일의 값을 읽는데 걸리는 I/O time이 CPU time에 비해 결정적이지 않음을 알 수 있었다. Section B는 Vector Approximation 방법으로써 HBI의 벡터를 위한 I/O time과 CPU time이 VA-file에 비해 매우 큼을 알 수 있다[2]. Section C는 MAM으로 HCT의 성능이 가장 좋은 것으로 보아 HCT가 MAM방법의 대표적인 색인방법이라 할 수 있겠다. Section D는 SAM으로 IQ-tree가 SAM방식의 한계를 Quantization이라는 방법을 통해 해결하여 그 성능이 가장 좋음을 알 수 있었다. 또한 차원 증가에 따른 각 색인 방법들의 속도 비교 결과 HBI, HCT, IQT가 상대적으로 좋은 성능을 보였으며 특히 IQ-tree가 차원 증가에 따른 성능이 가장 좋음을 알 수 있었다. 오브젝트 증가에 따른 비교 실험 결과 모든 방법들이 오브젝트 수에 따라 성능이 Linear하게 증가함을 알 수 있었다. 이외에도 차원 증가에 따른 실험 및 동적 색인 구현 시 속도 비교 실험을 진행하였다.



< 그림 1> 색인 방법 별 k-NN검색 시간 비교

5. 결론 및 향후 연구방향

기존의 멀티미디어 데이터 색인 검색을 위해 제안된 다양한 방법들을 구현하고 여러 성능 측정 기준에 따라 분석함으로써 각각의 특징 및 성능을 객관적으로 평가하였다. 향후 새로운 색인 방법이 제시될 경우, 본 논문에서 제안한 방법을 사용하여 객관적인 성능을 검증할 수 있을 뿐 아니라, 해당 알고리즘의 bottle neck을 쉽게 파악할 수 있을 것이다.

6. 참고문헌

[1] 문주선, "고차원 멀티미디어 데이터에 대한 내용기반 검색을 위한 인덱싱 방법들의 성능 비교", 12, 2007

[2] P. N. Yianilos, "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces," in *Proceedings Fourth Annual ACM-SIAM Symposium Discrete Algorithms*, Austin, TX, Jan. 25-27, pp. 311-321, 1993.

[3] S. Berchtold, C. Bohm, H. V. Jagadish, H. -P. Kriegel, and J. Sander, "Independent Quantization: An Index Compression Technique for High-Dimensional Data Spaces," in *Proc. 16th Int. Conf. Data Engineering*, SanDiego, CA, pp. 577-588, Feb, 2000