대용량 멀티미디어 데이터의 내용 기반 검색을 위한 고확장 지원 색인 기법

A Scalable Index for Content-based Retrieval of Large Scale Multimedia Data

최현화, 이미영, 이규철*
한국전자통신연구원, 충남대학교*

Choi Hyun-HWa, Lee Mi-Young, Lee Kyu-Chul*
Electronics and Telecommunications Research Institute, Chung Nam Univ.*

요약

카메라 기술의 발전 및 사용자 중심의 인터넷 패러다임인 웹 2.0을 토대로 멀티미디어 데이터가 급증하면서, 멀티미디어 검색은 인터넷 서비스로서 그 중요성이 날로 증가되고 있다. 현재 멀티미디어 검색은 단순한 키워드(keyword) 검색에 의존하고 있는 실정으로, 정보 검색의 정확도 및 사용자의 만족도를 충족시키기 위해서는 내용 기반 검색 지원이 필요하다. 본 논문에서는 대용량의 멀티미디어 데이터의 내용 기반 검색을 지원하기 위하여, 데이터의 분포에 따른 다중 길이의 시그니처를 기반으로 한 새로운 분산 인덱스 구조를 제안한다. 제안하는 인덱스 구조는 고차원 데이터의 클러스터링에 따라 데이터의 분포를 분석하여 서로 다른 요약 파일을 분산 생성하고, 이를 기반으로 유사 검색을 병렬로 수행할 수 있도록 설계되었다. 그리하여, 클러스터 환경 하에서 고차원 데이터의 분산 저장이 용이하고, 각 노드들은 서로 다른 시그니처 파일을 기반으로 검색을 병렬 수행함으로써 효율적인 검색을 지원한다.

Abstract

The proliferation of the web and digital photography has drastically increased multimedia data and has resulted in the need of the high quality internet service based on the moving picture like user generated contents(UGC). The keyword-based search on large scale images and video collections is too expensive and requires much manual intervention. Therefore the web search engine may provide the content-based retrieval on the multimedia data for search accuracy and customer satisfaction. In this paper, we propose a novel distributed index structure based on multiple length signature files according to data distribution. In addition, we describe how our scalable index technique can be used to find the nearest neighbors in the cluster environments.

Ⅰ. 서론

카메라 기술의 발전 및 사용자 중심의 인터넷 패러다임인 웹 2.0을 토대로 멀티미디어 데이터가 급증하면서 멀티미디어 검색은 인터넷 서비스로서 그 중요성이 날로 증가되고 있다. 특히, 멀티미디어 데이터는 웹 페이지와 같은 기존의 텍스트 기반 데이터와 달리, 이미지는 물론 사운드, 애니메이션, 비디오 등 여러 형태의 데이터들이 결합된 동영상 콘텐츠를 포함한다. 그러나, 현재 멀티미디어 검색은 단순한 키워드(keyword) 검색에 의존하고 있는 실정으로, 정보 검색의 정확도 및 사용자의 만족도를 충족시키기 위해서는 내용 기반 검색 지원이 필요하다 하겠다. 멀티미디어 데이터의 내용 기반 검색이란, 색깔, 모양, 질감 등과 같은 객체의 특징들을 각각 벡터 데이터로 표현하여 그와 유사한 벡터 데이터를 지니는 객체들을 검색하는 것을 말한다.

멀티미디어와 같은 고차원 데이터의 유사 검색(similarity search)을 지원하기 위하여 M-tree,

Spill-tree, Hybrid Spill-tree등 다양한 인덱스 구조에 대한 연구가 진행되어 왔다. 그러나, 기존의 인덱스 구조는 고차원 데이터의 유사 검색 서비스를 지원할 만큼 효율적이지 못하다. 뿐만, 아니라 단일 노드에서 효과적으로 수행되도록 고안되었기에, 인터넷 서비스와 같은 클러스터 환경 하에서 대용량 데이터를 관리하기 위하여 갖추어야할 고확장성 및 검색의 병렬 수행에 대한 고려가 부족하다.

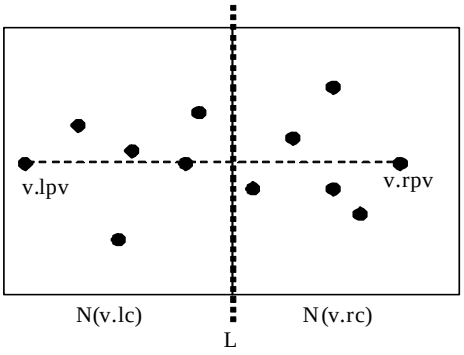
본 논문에서는 대용량 멀티미디어 데이터의 내용 기반 검색을 지원하기 위하여, 데이터의 분포에 따른 다중 길이의 시그니처들을 기반으로 한 새로운 분산 인덱스 구조를 제안한다. 제안하는 인덱스 구조는 고차원 데이터의 클러스터링에 따라 데이터의 분포를 분석하여 서로 다른 요약 파일을 분산 생성하고, 이를 기반으로 유사 검색을 병렬로 수행할 수 있도록 설계되었다. 그리하여, 본 논문에서 제안하는 인덱스 구조는 클러스터 환경하에서 고차원 데이터의 분산 저장이 용이하고, 각 노드들은 서로 다른 길이의 시그니처 파일을 기반으로 검색을 병렬 수행함으로써 효율적인 검색을 지원한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 고차원 데이터의 인덱스 기법에 대한 기존 연구를 살펴보고, 3장에서는 본 연구에서 제안하는 다중 길이 시그니처 기반 파일 기반 분산 인덱스 구조에 대해서 설명한다. 끝으로 4장에서는 결론 및 향후 연구 방향에 대해서 논의한다.

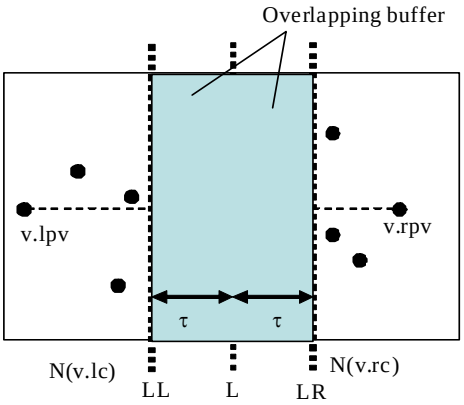
Ⅱ. 관련 연구

멀티미디어 데이터에 대한 내용 기반 검색을 지원하기 위하여, 고차원 데이터의 유사도 검색을 지원하는 인덱스 방법으로는 크게 트리 기반 인덱스 구조와 필터링 기반 인덱스 구조로 분류할 수 있다.

트리 기반 고차원 인덱스 기법에는 M-tree[1], Spill-tree[2] 및 Hybrid Spill-tree[3] 등이 존재한다. M-tree[1]는 그림1의 a)와 같이 모든 데이터 사이의 거리 계산을 통해 가장 먼 거리의 두 점을 선택한다. 선택한 두 점을 잇는 선분에 모든 데이터를 투영한 후, 선분의 이등분점에 가장 가까운 점을 지나는 수직 이등분선을 기준으로 데이터를 분할한다. 이의 반복 수행을 통하여 구축된 M-tree는 MT-DFS(M-tree Depth First Search) 라는 깊이 우선 탐색을 수행함으로써 정확한 k-최근접점 탐색(K-Nearest Neighbor Search)를 지원한다. 한편, Spill-tree[2]는 가장 먼 거리의 두 점을 잇는 선분의 수직이등분선을 기준으로 데이터를 분할하며, 그림1 b)와 같이 "overlapping buffer"라 불리는 데이터 공유 영역을 설정하여 이를 양쪽 노드가 모두 포함한다. 이는 M-tree에서 검색 시간의 대부분을 차지하는 역추적(back-tracking)을 줄이기 위하여 고안되었다. 또한, Hybrid Spill-tree[3]는 M-tree와 Spill-tree를 혼합한 구조로, 구글(Goolge)의 대용량 이미지 데이터의 분산 저장 및 검색 시스템에 활용되기도 하였다. 특히 데이터의 샘플링을 통하여 top-tree로써 M-tree를 구축하고, M-tree의 말단 노드는 특정 분산 시스템 노드를 지정토록 하였다. 분산 시스템 노드는 내부적으로 Hybrid Spill-tree를 구축하여 고차원 데이터에 대한 검색을 수행한다.



a) Partitioning in a metric tree

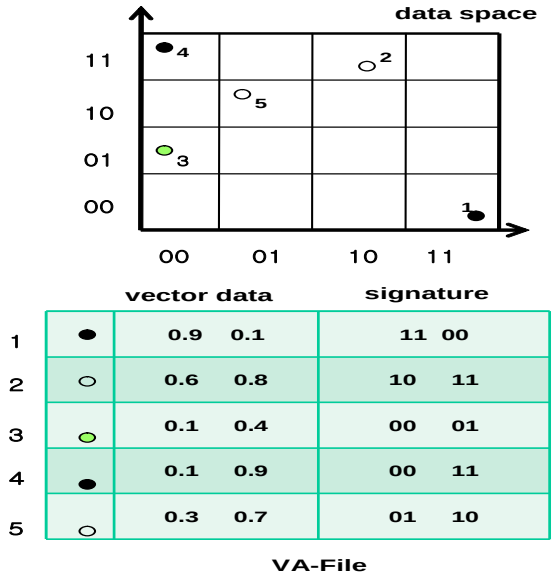


b) Partitioning in a spill tree

▶▶ 그림 1. M-tree와 Spill-tree의 노드 분할

필터링 기반 인덱스 구조는 데이터의 공간을 2^b개의 사각형 셀로 분할하고, 각 셀을 길이 b의 비트로 표현

한다. 그림 2와 같이 셀에 놓이는 벡터를 근사(signature)하고, 근사값들의 집합(VA-file, signature file)을 대상으로 k-최근접점을 탐색한다[4]. 이는 모든 k개의 최근접점을 정확히 찾아내는 방법으로, 트리 기반 인덱스 구조가 가지는 차원의 저주를 극복하기 위하여 제안되었다. VA-file과 같은 일환의 LPC-file[5] 및 CBF(Cell Based Filtering)[6] 연구는 질의 지점과 객체 사이의 거리를 계산하는데 있어, 최소한의 추가 데이터를 통하여 필터링 능력을 높임으로써 검색 성능을 향상시킨 연구이다.



▶▶ 그림 2. 고차원 데이터의 시그니처 구성

기존의 연구들은 한 노드에서 효과적으로 수행될 수 있는 고차원 데이터의 인덱스 구조로, 대용량의 고차원 데이터를 위한 고확장성을 지원하지 못한다. 구글은 Hybrid Spill-tree를 이용한 분산 인덱스 구조를 제안하기도 하였으나, 트리 기반 인덱스가 필터링 기반 인덱스를 통한 검색보다 그 성능이 좋지 않다는 기존 연구 결과를 감안하면, 실제 인터넷 서비스로서 멀티미어 검색을 지원하기에는 최적이라고 할 수 없다.

한편, Peer-to-Peer 환경을 기반으로 한 고차원 데이터의 분산 인덱스 연구가 활발히 이뤄지고 있다. Peer-to-Peer 환경 하에서 고차원 데이터의 대표적인 분산 인덱스 구조로는 Skip Graph를 이용한 SkipIndex[7]와 LSH를 이용한 LSH Forest[8] 등이 있다. 이러한 분산 인덱스는 현재 웹 포털에서 필요로 하는 멀티미디어 데이터의 내용 기반 검색을 지원하기에는 통신 비용에 따른 성능 저하와 검색의 정확도가 낮아 적합하지 않다. 즉, 클러스터 환경 하에서 대용량의 고차원 데이터를 기반으로 유사 검색을 지원하는 색인 구조에 대한 연구가 필요하다.

Ⅲ. 다중 길이 시그니처 기반 분산 인덱스

기존의 필터링 기반 색인은 같은 셀에 여러 개의 점들이 들어가지 않는다는 묵시적인 가정 하에 수행된다. 예를 들어, 단위 하이퍼 큐브(unit hyper-cube) $\Omega=[0,1]^d$ 내에서 한 점이 한 셀에 놓일 확률은 $1/2^{bd}$이다. 여기서 b는 각 차원에 할당된 비트 수이며, d는 차원 수이다. 다른 점이 같은 셀에 들어갈 확률은 $1-(1-1/2^{bd})^{N-1} \approx N/2^{bd}$ 으로, 셀의 수(2^{bd})가 전체 데이터의 집합의 크기 N보다 훨씬 크기 때문에 대부분의 셀은 비어있다고 가정한다.

그러나, 실세계의 데이터는 대부분 클러스터를 이루는 비정규분포 형태로 존재한다. 요즘처럼 인터넷을 통해 멀티미디어의 데이터 수가 급증하면서, 한 셀에 여러 개의 점들이 놓일 가능성 또한 커지고 있다. 한 셀에 여러 개의 점들이 놓일 확률이 증가할수록, 근사의 구분 능력은 현격히 떨어지게 되고, 이는 디스크의 접근 회수를 증가시켜 검색 성능 저하의 요인으로 작용한다. 이런 경우에는 근사를 표현하는 비트의 길이를 늘려, 셀의 크기를 축소함으로써 근사의 구분 능력을 높일 필요가 있다.

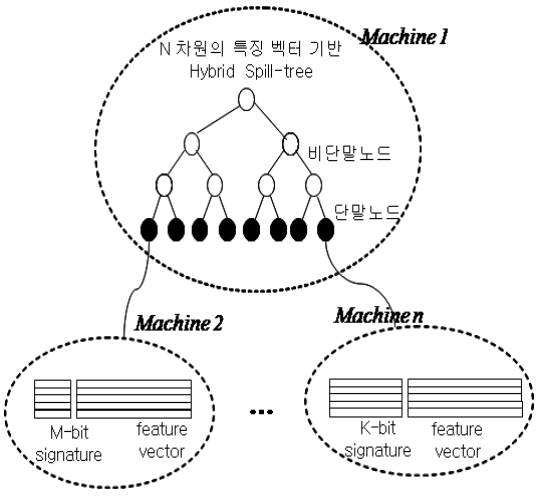
이렇듯 필터링 기반 인덱스 구조에서는 근사를 표현하기 위한 비트의 길이는 읽어야할 시그니처 파일의 크기 및 검색의 정확도를 결정하는 중요한 요소이다. 즉, 시그니처를 위한 비트 길이를 크게 할수록 필터링 대상이 커져 정확도가 증가하는 한편, 검색해야할 대상의 시그니처들의 크기가 커짐으로써 디크스 접근 회수는 증가하게 된다. 그러나 대부분의 필터링 기반 색인 기법들은 시그니처 표현을 위한 비트 길이 결정 시에 데이터의 분포 정보를 고려하지 않거나, 혹은 데이터 분포의 밀도차가 극심한 상황에서도 대체로 4에서 8까지의 길이 중 하나를 선택하여 시그니처로 표현하고 있다. 물론 셀에 할당할 비트의수를 달리 결정하는 연구도 있었느나 한 노드에서 검색을 수행하는 경우에 데이

터마다 다른 길이를 가지는 시그니처를 관리하는데 드는 비용이 존재하여 큰 효과를 거두기는 어렵다.

본 논문에서는 데이터 공간의 지역적 통계 특성에 따라 1차 공간을 분할하고, 최종 분할된 공간의 크기에 따라 사용할 시그니처를 달리 생성한다. 달리 생성된 시그니처 파일들은 서로 다른 분산 노드에 저장하여 검색에 활용하는 다중 시그니처 기반 분산 인덱스를 소개한다.

1. 인덱스 구조

다중 시그니처 기반 분산 인덱스는 기본적으로 상위에 Hybrid Spill-tree로 구성된다. 이는 대용량의 멀티미디어 데이터에서 임의적으로 한 노드 메모리에 적재 가능한 정도의 균등한 표본 추출을 통해 구성된다. Hybrid Spill-tree가 구축되면, 각 말단 노드마다 관리할 데이터 공간의 크기를 계산한다. 이는 공간 내에 존재할 수 있는 가장 긴 거리의 두 점간 거리로 계산될 수 있다. 계산된 데이터의 공간 크기가 특정 임계치(t)를 넘지 못하는 경우, 해당 말단 노드는 더 많은 비트로 구성되는 시그니처를 사용하도록 한다. 즉, 기본적으로 4 비트의 시그니처를 사용한다면, 임계치를 넘기지 못하는 노드들은 6 혹은 8 비트의 시그니처를 사용하는 것이다. 데이터 공간의 크기가 작다는 것은 전체 데이터을 4 비트의 시그니처로 표현하였을 경우, 한 셀에 한 개 이상의 점이 속할 가능성이 더 높다는 것을 의미하기에, 근사값의 구분 능력을 높여주기 위한 것이다.



▶▶ 그림 3. 다중 길이 시그니처 기반 분산 인덱스 구조

특히, 그림 3과 같이 top-tree인 Hybrid Spil-tree의 말단 노드들은 서로 다른 컴퓨팅 노드를 지정하도록 하여, 각 컴퓨팅 노드에서는 특정 하나의 길이로 구성되는 시그니처만을 저장 및 관리하게 한다. 그리하여, 본 노논문에서 제안하는 다중 시그니처 기반 분산 인덱스 구조는 사용자에게 하나의 Hybrid Spill-tree로 보이나, 내부적으로 다중 컴퓨팅 노드에 색인 데이터를 분할 관리함으로써, 한 노드에서 수용할 수 없는 대용량의 데이터에 대한 인덱스를 구축할 수 있다. 그리고, 상위 Hybrid Spill-tree를 관리하는 컴퓨팅 노드를 제외한 컴퓨팅 노드들은 특정 길이의 비트로 구성되는 시그니처 및 각 시그니처에 해당하는 특징 벡터를 함께 관리함으로써 다중 길이 시그니처 관리에 따른 부하를 가지지 않는다.

2. 검색 알고리즘

다중 길이 시그니처 기반 분산 인덱스에 의한 k-최근접점 검색 알고리즘은 다음과 같다. 먼저, 사용자로부터 특정 멀티미디어 데이터를 입력받으면, 그로부터 특징 벡터를 추출한다. 추출한 특징 벡터를 기반으로 top-tree인 Hybrid Spill-tree를 탐색한다. Hybrid Spill-tree 탐색은 MT-DFS 와 defeatist 검색을 통해 이뤄진다. 탐색을 통해 말단 노드가 결정되면, 말단 노드가 가리키는 컴퓨팅 노드에 특징 벡터를 전달한다. 이때, 결정된 말단 노드는 Hybrid Spill-tree 특성 상 하나 혹 그 이상이 될 수 있으며, 이는 가리키는 컴퓨팅 노드가 다르므로 병렬로 검색 수행이 가능하다. 특징 벡터를 전달받은 컴퓨팅 노드는 특징 벡터로부터 해당 노드가 관리하고 있는 특정 길이의 시그니처로 변환한다. 변환된 시그니처를 질의 점으로 하여, 컴퓨팅 노드 내에서 관리하고 있는 시그니처의 집합을 대상으로 거리 연산을 통해 1차 후보 시그니처 결과 집합을 구하고, 구해진 후보 시그니처에 해당하는 특징 벡터와 질의 특징 벡터와의 거리 연산을 통해 2차 후보 특징 벡터 결과 집합을 구성한다. 컴퓨팅 노드는 후보 특징 벡터 결과 집합과 이의 거리값을 top-tree를 관리하는 컴퓨팅 노드에 전달한다. top-tree의 컴퓨팅 노드는 전달받은 특징 벡터와 거리값을 통해 최종 k-최근접점을 결정하여 사용자에게 전달한다.

구글에서 제안한 병렬 Hybrid Spill-tree가 top-tree에서 O(log n), 각 말단 노드에서 관리하는 Hybrid

Spill-tree의 O(log n)에 의해 최종 2O(log n)의 복잡도를 가진다. 한편, 본 논문에서 제안하는 검색 알고리즘은 top-tree에서 O(log n)을 가지나, Hybrid Spill-tree 탐색 결과 얻어지는 말단 노드의 개수는 구글의 M-tree의 역추적(back tracking)에 대비하여 결정되는 말단 노드 수보다는 적다는 이점이 있다. 한편, top-tree의 말단 노드가 가리키는 컴퓨팅 노드에서 구글은 트리 기반 검색을 수행하나, 본 논문에서 제안하는 방법은 전체 데이터의 순차 검색을 수행한다. 순차 검색이 트리 기반 검색보다 데이터의 차원의 수가 클수록 그 성능이 좋다는 것은 Roger Weber의 논문[4]을 통해 이미 알려진 사실이다. 뿐만 아니라, 시그니처 구성 시에 한 셀에 한 개 이상의 점이 들어갈 가능성이 많은 경우 더 많은 비트로 구성되는 시그니처를 구축함으로써, 검색의 정확도는 물론 성능 향상을 얻을 수 있다. 결론적으로 클러스터 환경에서 다중 길이 시그니처 기반 분산 인덱스 구조는 구글에서 제안하는 병렬 Hybirkd Spill-tree보다 좋은 검색 성능을 제공할 것이다.

Ⅳ. 결론

본 논문에서는 웹 서비스와 같은 클러스터 환경 하에서 멀티미디어 데이터에 대한 내용 기반 검색을 지원하기 위한 분산 인덱스를 제안하였다. 본 논문에서 제안하는 인덱스 구조는 top-tree로 Hybrid Spill-tree를 구성하고, 말단 노드가 분산된 컴퓨팅 노드를 가리키도록 하여 검색의 병렬 처리는 물론 대용량 멀티미디어 데이터에 대한 색인이 분산 저장될 수 있다. 또한, 각 컴퓨팅 노드에서 관리하는 특징 벡터의 근사값인 시그니처의 순차 검색을 통해 k-최근접점을 찾는다. 이때, 컴퓨팅 노드가 관리하는 데이터 공간이 작은 경우, 즉 한 셀에 한 개 이상의 점이 들어갈 가능성이 높은 것은 더 많은 비트로 구성되는 시그니처를 관리하도록 설계되었다. 이는 시그니처 기반 검색에서 시그니처의 길이에 따라 검색 성능이 크게 좌우되는 것을 감안한 것이다.

본 논문에서는 검색 복잡도 계산을 통해 구글에서 제공하는 병렬 Hybrid Spill-tree보다 성능을 좋음을 보였고, 데이터의 분포에 따른 다중 길이의 시그니처 관리를 통해 보다 좋은 성능을 예상할 수 있다. 다음 논문에서는 실험을 통하여 다중 길이 시그니처 기반 분산 인덱스가 병렬 Hybrid Spill-tree보다 얼마나 검색 성능을 향상시켰는지를 보일 것이다.

■ 참 고 문 헌 ■

[1] Ciaccia, P., Patella, M. and Zezula, P., “M-tree: An Efficient Access Method for Similarity Search in Metric Spaces”, VLDB, 1997.

[2] Liu, T., Moore, M. Gray, A. and Yang, K., "An Investigation of Practical Approximate Nearest Neighbor Algorithms", NIPS, 2004.

[3] Liu, T., Rosenberg, C. and Rowley, H. A., "Clustering Billions of Images with Large Scale Nearest", WACV, 2007.

[4] Weber, R., Schek H. J. and Blott, S., "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces", VLDB, 1998.

[5] Cha, G. H., Zhu. X., Petkovic, D. and Chung,. C. W., "An Efficient Indexing Method for Nearest Neighbor Searches in High-Dimensional Image Databases", IEEE Transactions and Multimedia, Vol. 4. No. 1, p.76-87, 2002.

[6] Han, S. G. and Chang, J. W., "A New High-Dimensional Index Structure Using a Cell-Based Filtering Technique", LNCS 1884, Springer, p.79-92, 2000

[7] Zhang, C., Krishnamurthy A. and Wang, R., "SkipIndex:Towards a Scalable Peer-to-Peer Index service for High Dimensional Data", Technicla Report TR-703-04, Princeton Univ. CS, 2004. http://www.cs.princeton.edu/~chizhang/skipindex.pdf

[8] Bawa, M., condie T. and ganesan P., "LSH Forset:Self-Tunning Indexes for Similarity Search", WWW, 2005.

본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2007-S-016-1, 저비용 대규모 글로벌 인터넷 서비스 솔루션]