Earth Mover's Distance 기반 M-트리의 성능 분석

이원조, 하성대, 정교성, 장민희, 김상욱
한양대학교 전자컴퓨터통신공학과

A Performance Analysis on the M-tree with the Earth Mover's Distance

Won-Jo Lee, Sung-Dae Ha, Kyo-Sung Jung, Min-Hee Jang, Sang-Wook Kim
Dept. of Electronics and Computer Engineering, Hanyang University

요 약

Earth mover's distance(EMD)는 내용 기반 데이터 검색을 위한 거리 함수로서 정확도가 매우 높은 검색 결과를 가져오지만, 계산 복잡도가 높아 대용량 데이터베이스에서 사용하기 어렵다. 이러한 문제를 해결하기 위한 방법으로 다차원 인덱스인 M-트리를 사용하여 데이터 비교 횟수를 크게 줄일 수 있다. 그러나 고차원의 데이터에 인덱스를 사용하면 차원의 저주 문제로 인해 검색 성능이 크게 저하될 수 있다. 본 논문에서는 이러한 문제를 정량적으로 검증하기 위하여 고차원 데이터를 대상으로 EMD 기반 M-트리를 구축한 후 다양한 실험을 수행한다. 실험 결과, 고차원 데이터에서도 차원의 저주는 일어나지 않는 것으로 나타났다. 이러한 공헌은 EMD의 검색 성능 개선을 위한 정책을 고안하는 데, 중요한 실마리를 제공할 수 있을 것이다.

1. 서론

최근 들어, 인터넷 기술의 발달과 모바일 기기의 보급 확산으로 동영상, 이미지와 같은 멀티미디어 데이터가 폭발적으로 증가하고 있다. 대용량 멀티미디어 데이터베이스에서 사용자가 원하는 데이터를 찾아주기 위해서는 멀티미디어 데이터를 질의로 던져 그 데이터의 색 분포나 형태, 그리고 질감 등을 분석하여 유사한 데이터를 검색하는 기술인 내용 기반 데이터 검색(content-based data retrieval)이 요구된다[1].

내용 기반 데이터 검색을 수행하기 위해서는 멀티미디어 데이터베이스 간에 거리를 측정하기 위한 거리 함수가 필요하다. 멀티미디어 데이터는 주로 고차원 히스토그램으로 표현되는데 이러한 히스토그램 데이터 간에 거리를 측정하기 위한 대표적인 거리 함수로서 EMD가 존재한다. EMD는 하나의 히스토그램을 비교하고자 하는 다른 히스토그램으로 옮기는데 들어가는 최소 비용을 계산하는 거리 함수이다[1]. EMD는 뛰어난 검색 정확도를 보이지만 계산 시간이 매우 크다는 단점을 가지고 있다. 각 데이터 비교 시 마다 거리 계산 시간이 매우 크다면 그 거리 함수는 대용량 데이터에서 사용하기 어렵다.

이러한 문제를 해결하기 위한 방법으로, 다차원 인덱스를 사용하여 데이터 비교 횟수를 줄일 수 있다. 그러나 멀티미디어 데이터는 고차원 히스토그램이기 때문에 차원의 저주 문제가 발생하여 데이터베이스의 모든 데이터와 거리를 계산하는 선형 검색 (linear scan)보다 느릴 가능성이 있다[2]. 이러한 이유로 EMD를 기반으로 한 다양한 연구들이 진행되었지만, 저자들이 아는 한 다차원 인덱스를 적용하여 EMD의 계산 횟수를 줄이고자 한 연구는 없었다. 본 논문에서는 다차원 인덱스인 M-트리를 기반으로 이러한 문제를 검증하고자 한다. M-트리는 각 데이터 간의 거리를 기반으로 다차원 인덱스를 구성하는 트리로서 고차원 데이터의 인덱싱에 효과적인 것으로 알려져 있다[3]. 본 논문에서는 실제 고차원 데이터를 이용하여 다양한 실험을 수행함으로써 EMD를 기반으로 한 M-트리의 검색 성능을 검증한다.

2. Earth Mover's Distance

본 장에서는 EMD를 설명하기 위해 두 이미지의 색상 간 거리 측정을 예로 든다. 두 이미지 P={(p_1, w_{p1}), ... ,(p_m, w_{pm})}, Q={(q_1, w_{q1}), ... ,(q_l, w_{qn})}가 있을 때 p_i와 q_j는 이미지의 색을, w_{pi}와 w_{qj}는 각 색의 비율을 의미한다. P와 Q의 비율 총합은 동일하다고 가정한다. 이때, EMD는 P와 Q의 거리를 측정하기 위해 minimum work를 계산한다. minimum work는 P의 색-비율 분포를 Q의 분포로 옮기는데 들어가는 최소 work의 양을 의미한다. work는 한 이미지에서 다른 이미지로 옮겨진 색들의 양 f와 ground distance d의 곱으로 구할 수 있다. ground distance란 색 간의 거리를 측정하기 위해 사용된 Euclidean distance나 L_1 distance와 같은 기본 거리를 의미한다[1]. work를 공식으로 표현하면 다음과 같다.

$$WORK(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}$$

F=[f_{ij}]는 p_i에서 q_j로 옮겨진 색의 양을 의미하고 D=[d_{ij}]는 옮겨진 색의 거리를 의미한다. EMD는 뛰어난 검색 정확도를 자랑하지만 계산 복잡도가 n 차원의 히스토그램이 존재할 때 O(n^3log$n$)이다. 각 데이터 간 거리 계산마다 이와 같은 계산복잡도가 필요하다면 그 거리 함수는 대용량 데이터베이스에서 사용하기 어렵다.

3. 연구 동기

대용량 데이터베이스에서 EMD를 기반으로 내용 기반 데이터 검색을 수행하기 위해서는 다차원 인덱스를 이용하여 데이터 비교 횟수를 줄여야 한다. 그러나 멀티미디어 데이터는 주로 고차원 히스토그램으로 표현된다. 고차원 데이터에서 인덱스를 사용하면 차원의 저주 문제가 발생할 수 있다. 차원의 저주란 차원의 수가 증가할수록 최근접 이웃 데이터 간 거리와 가장 멀리 떨어진 데이터 간 거리가 상대적으로 가까워지는 문제이다. 이에 따라 인덱스 구축 시 다차원 인덱스의 많은 노드들이 서로 겹치게 되어 선형 검색보다 느린 검색 성능을 보일 수 있다.

다차원 인덱스의 종류는 데이터의 절대 좌표를 이용하여 인덱싱하는 공간 기반 방법 (spatial-based method)과 데이터 간의 거리를 기반으로 인덱싱하는 거리 기반 방법(distance-based method)으로 나눌 수 있다[3]. 공간 기반 방법의 대표적인 방법으로는 R-트리를 들 수 있고, 거리 기반 방법의 대표적인 방법으로는 M-트리를 들 수 있다. R-트리는 데이터의 위치를 기반으로 최소 포함 사각형(minimum bounding rectangle) 형태의 노드를 이용하여 인덱스를 구축한다. 이에 비해, M-트리는 데이터들의 거리를 기반으로 구(sphere) 형태의 노드를 이용하여 인덱스를 구축한다. 기존 논문의 실험 결과, R-트리의 검색 성능은 고차원 데이터일수록 크게 떨어지는 것으로 나타났지만 M-트리는 50차원 이상의 고차원에서도 좋은 성능을 보이는 것으로 나타났다[3]. 즉, M-트리를 이용하면 EMD의 계산 횟수를 줄여 선형 검색보다 더 좋은 성능을 발휘할 가능성이 있다는 것이다.

본 논문에서는 M-트리의 검색 성능을 검증하기 위하여 EMD를 기반으로 M-트리를 구축한 후 다양한 실험을 수행한다. 이에 더해, 선형 검색과 M-트리의 검색 성능을 비교함으로써 고차원 데이터 인덱싱 시 차원의 저주 문제가 발생하는지 검증한다.

4. 실험

4.1. 실험 환경

실험에 사용한 데이터 집합은 3,932개로 구성된 RETINA 데이터와 49,973개로 구성된 DBLP 데이터이다[4]. RETINA 데이터는 96차원의 히스토그램 데이터이고 각 히스토그램의 빈(bin)은 2차원 특성 벡터이다. DBLP 데이터는 8차원의 히스토그램 데이터이고, 각 히스토그램의 빈은 3차원 특성 벡터이다.

각 데이터 집합을 대상으로 EMD 기반 M-트리를 구축한 후 데이터베이스 안의 모든 데이터와 거리를 측정하는 선형 검색 방법과 k-최근접 검색 성능을 비교하였다. k-최근접 검색의 k 값은 10, 30, 그리고 50으로 설정하였다. 검색의 성능을 평가하기 위한 척도로는 파일 접근 시간을 제외한 검색 시간, M-트리의 I/O 횟수, 그리고 거리 계산 수를 사용하였다. 정확한 성능 평가를 위하여 동일한 유형의 임의 질의 100개를 던져 그 결과들의 평균을 측정하였다. 모든 실험은 윈도우즈7에서 4GH의 메모리 크기와 Intel Core 3.10GHZ PC를 이용하여 수행되었다.

4.2 실험결과

표 1은 RETINA 데이터에 대해 k-최근접 검색을 수행한 결과이다. 표에서 보는 바와 같이 선형 검색보다 M-트리의 검색 시간이 약 1/14 정도 감소한 것을 볼 수 있다. M-트리에서는 검색 시 추가적인 I/O 횟수가 들어가지만 EMD 계산 수가 크게 줄어들기 때문에 선형 검색보다 검색 시간이 감소하는 것이다.

표 2는 DBLP 데이터를 대상으로 k-최근접 검색을 수행한 결과이다. 선형 검색보다 약 1/2 정도 검색 시간이 감소한 것을 볼 수 있다. DBLP 데이터는 8차원 데이터이기 때문에 각 데이터 비교 시 EMD의 계산 속도가 매우 빠를 수밖에 없다. 이에 더해 M-트리에서는 검색에 따른 추가적인 I/O 시간이 필요하므로 RETINA 데이터에 비해 검색 성능 향상 효과가 그리 크지 않다. 그러나 거리 계산 수가 선형 검색에 비해 크게 줄어들기 때문에 M-트리를 사용함에 따른 기본적인 성능 향상 효과를 얻을 수 있다.

		시간(s)	I/O reads	거리 계산 수
k=10	Linear+EMD	68.3096	-	3932
	Mtree+EMD	4.13188	63.43	489.09
k=30	Linear+EMD	68.3058	-	3932
	Mtree+EMD	5.04131	70.88	640.79
k=50	Linear+EMD	68.3308	-	3932
	Mtree+EMD	5.88435	72.66	708.1

<표 1> RETINA 데이터의 k-최근접 검색 결과

		시간(s)	I/O reads	거리 계산 수
k=10	Linear+EMD	0.86147	-	49973
	Mtree+EMD	0.4037	97.9	3024.62
k=30	Linear+EMD	0.86606	-	49973
	Mtree+EMD	0.45108	107.54	3735.74
k=50	Linear+EMD	0.87003	-	49973
	Mtree+EMD	0.41486	100.59	3819.94

<표 2> DBLP 데이터의 k-최근접 검색 결과

5. 결론

본 논문에서는 고차원 데이터에서 EMD를 기반으로 한 M-트리의 검색 성능을 검증하기 위해 다양한 실험을 수행하였다. 실험 결과, 고차원 데이터에서도 M-트리를 사용하는 것이 선형 검색보다 빠른 검색 시간을 보이는 것으로 나타났다. 향후 연구로서 이러한 결과를 기반으로 EMD의 계산 속도를 줄이기 위한 다양한 연구를 수행할 예정이다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업의 연구결과로 수행되었습니다. (NIPA-2011-C6150-1101-0001)

참고문헌

[1] Y. Rubner, et al., “The Earth Mover’s Distance as a Metric for Image Retrieval,” *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99-121, 2000.

[2] R. Weber, et al., "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces," In *Proc. Int'l Conf. on VLDB*, pp. 194-205, 1998.

[3] P. Ciaccia, et al., "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces," In *Proc. Int'l Conf. on VLDB*, pp. 426-435, 1997.

[4] J. Xu, et al., "Efficient and Effective Similarity Search over Probabilistic Data based on Earth Mover's Distance," In *Proc. Int'l. Conf. on VLDB*, pp. 758 - 769, 2010.