SVD를 기반으로 한 고차원 데이터 및 질의 집합의 생성

김 상 욱*

An SVD-Based Approach for Generating High-Dimensional Data and Query Sets

Sang-Wook Kim*

Abstract

Previous research efforts on performance evaluation of multidimensional indexes typically have used synthetic data sets distributed uniformly or normally over multidimensional space. However, recent research result has shown that these kinds of data sets hardly reflect the characteristics of multimedia database applications. In this paper, we discuss issues on generating high dimensional data and query sets for resolving the problem. We first identify the features of the data and query sets that are appropriate for fairly evaluating performances of multidimensional indexes, and then propose HDDQ_Gen(High-Dimensional Data and Query Generator) that satisfies such features. HDDQ_Gen supports the following features : (1) clustered distributions, (2) various object distributions in each cluster, (3) various cluster distributions, (4) various correlations among different dimensions, (5) query distributions depending on data distributions. Using these features, users are able to control the distribution characteristics of data and query sets. Our contribution is fairly important in that HDDQ_Gen provides the benchmark environment evaluating multidimensional indexes correctly.

※ 본 연구는 강원대학교 멀티미디어 연구센터를 통한 정보통신 우수시범학교 지원사업과 한국과학재단의 99 해외 Post-Doc 방문 프로그램 및 2000 목적기초연구 중 지역대학 우수과학자 지원 프로그램(과제 번호 : 2000-1-51200-006-1)의 연구비 지원을 받았습니다.

* 강원대학교 컴퓨터정보통신공학부

1. 서 론

멀티미디어 데이터베이스에서는 질의에서 주어진 객체와 유사한 객체를 찾는 최근접 질의(nearest neighbor query)가 빈번하게 사용된다[Fal1994][Fal1995]. 기존의 연구에서는 각 객체로부터 색상, 질감, 명암 등의 특징 벡터(feature vector)를 추출한 후, 각 객체를 다차원 벡터 공간(vector space)내의 한 점으로 간주한다[Ary1994][Jag1991][Nib1993]. 최근접 질의는 "다차원의 벡터 공간 내에 객체 점들의 집합과 질의 점이 주어질 때, 질의 점으로부터 유클리드 거리(Euclidean distance)가 최소인 객체를 찾는 질의"로 정의된다[Ber1998][Kor1996][Rou1995][Sei1998].

최근접 질의의 효과적인 처리를 위하여 기존의 기법들은 대부분 다차원 색인(multidimensional index)을 이용한다. 다차원 색인은 다차원 벡터 공간내의 점들을 빠르게 검색하기 위한 색인 구조이다[Gae1998]. 그러나 기존의 다차원 색인은 GIS 등 저차원 응용의 경우 매우 좋은 성능을 나타내지만, 멀티미디어 응용에서와 같이 고차원 응용의 경우에는 그 성능이 크게 떨어지는 것으로 알려져 있다[Web1998][Ber1996]. 따라서 이러한 차원의 저주(dimensionality curse)를 극복하기 위하여 새로운 색인 구조에 관한 많은 연구들이 진행되고 있다.

새로운 색인의 성능을 공정하게 평가하기 위해서는 해당 응용에서 실제로 사용되는 데이터 및 질의를 대상으로 올바른 실험이 수행되어야 한다. 그러나 색인을 설계한 시점에서 이러한 실제 데이터 및 질의를 구하는 것이 용이하지 않는 경우에는 표준으로 인정된 인위적 데이터 및 질의를 대상으로 실험이 수행되어야 하며, 이러한 인위적 데이터 및 질의는 실제 데이터 및 질의와 유사한 특성을 가져야 한다[Zob1996].

최근접 질의에 관한 기존의 연구에서는 대부분 인위적 데이터로서 객체들이 다차원 공간에서 균일 분포(uniform distribution) 혹은 정규분포(normal distribution)를 취하는 데이터를 많이 사용한다[Ber1996][Web1998][Ber1998]. 그러나 이러한 분포는 멀티미디어 데이터베이스 응용에서 실제 환경을 전혀 반영하지 않으며, 특히, 고차원 환경에서 최근접 질의가 요구되는 상황에서는 전혀 의미 없는 실험 데이터로 밝혀졌다[Bey1998].

본 논문에서는 이러한 문제점을 해결하기 위한 고차원 데이터 및 질의 집합의 생성에 관하여 논의한다. 본 논문에서는 먼저 고차원 색인 기법 및 최근접 질의 처리 기법의 성능을 공정하게 평가하기 위한 실험에서 사용될 데이터 및 질의들이 갖추어야 할 특성을 지적하고, 이러한 특성을 가지는 데이터 및 질의 집합을 인위적으로 생성할 수 있는 HDDQ_Gen(High-Dimensional Data and Query Generator) 기법을 제시한다. HDDQ_Gen의 주요 특성은 (1) 클러스터 단위의 분포 제공, (2) 클러스터 내에서의 객체들의 다양한 분포 제공, (3) 클러스터들의 다양한 분포 제공, (4) 차원들간의 상관 관계 지원, (5) 객체 분포를 반영하는 질의 분포 지원 등이다. 따라서 HDDQ_Gen 기법은 사용자가 데이터 및 질의의 분포 특성을 자유롭게 제어할 수 있도록 한다. 본 연구는 응용의 특성을 반영하는 벤치마킹 환경을 제공함으로써 고차원 색인 기법 및 최근접 질의 기법의 성능을 올바르게 평가할 수 있는 기반을 마련하였다는 점에서 의미가 있다.

본 논문의 구성은 다음과 같다. 먼저, 제 2장에서는 제안된 HDDQ_Gen의 설계 요건을 기술한다. 제 3장에서는 HDDQ_Gen에서 사용된 수학적 배경인 SVD(singular value decomposition)에 대하여 간략히 설명한다. 제 4장에서는 HDDQ_Gen 알고리즘을 제시하고, 채택된 세부 방안에 관하여 구체적으로 설명한다. 또한, 생

성된 객체 및 질의 집합을 예로 제시함으로써 사용자가 원하는 특성에 적합한 분포를 올바르게 제어할 수 있음을 보인다. 제 5장에서는 논문을 요약하고, 결론을 내린다.

2. 설계 요건

본 장에서는 고차원 색인 기법들의 성능을 공정하게 평가하기 위한 실험에서 사용될 데이터 및 질의들의 생성자 HDDQ_Gen의 설계 요건에 관하여 논의한다.

2.1 객체 클러스터링

참고 문헌 [Bey1998]의 분석 결과에 의하면 주어진 질의 점에서 가장 가까운 객체와 가장 먼 객체간의 거리 차는 차원이 증가함에 따라 점차 작아지는 것으로 나타났다. 특히, 균일 분포를 취하는 차원 수가 20이상인 객체들의 경우, 이 거리 차는 매우 작아지므로 이러한 상황에서는 최근접 객체의 의미 자체가 없어지게 된다. 이러한 결과는 참고 문헌 [Web1998]의 연구 결과와도 일치한다. 전체 공간상에서 객체들이 클러스터(cluster)들의 집합으로 분포되는 응용에 한하여 최근접 질의를 의미 있게 사용할 수 있다 [Bey1998]. 본 연구에서는 이를 반영하기 위하여 HDDQ_Gen이 객체들을 클러스터 단위로 분포시킬 수 있도록 한다.

2.2 클러스터내의 객체 분포

실제 응용에서 각 클러스터 내에서 객체들이 분포하는 형태는 매우 다양하다. 본 연구에서는 HDDQ_Gen이 객체들의 집합으로 구성되는 클러스터의 모양과 크기를 다양하게 제어할 수 있도록 한다.

2.3 클러스터들의 분포

클러스터내의 객체 분포와 마찬가지로 실제 응용에서 클러스터들이 분포하는 형태 또한 매우 다양하다. 본 연구에서는 HDDQ_Gen이 클러스터의 중심점들이 벡터 공간상에서 분포하는 형태를 다양하게 제어할 수 있도록 한다.

2.4 상관 관계

균일 분포 가정(uniform distribution assumption)과 더불어 다차원 색인에 관한 많은 연구들이 분석적인 성능 평가를 위하여 사용하는 기본적인 가정은 "객체들은 서로 다른 차원간에 상관 관계를 가지지 않는다"라는 상호 독립 가정(independence assumption)이다. 그러나 대부분 실제 데이터 집합에서는 차원간의 상관 관계가 존재하며, 특히 고차원 벡터 공간에서는 이러한 상관 관계가 존재할 가능성이 더욱 높아진다[Agg2000]. 본 연구에서는 이러한 실제 상황을 반영하기 위하여 HDDQ_Gen이 각 클러스터내의 객체들이 모든 가능한 서로 다른 두 차원간의 상관 관계를 다양하게 제어할 수 있도록 한다.

2.5 질의 분포

고차원 색인들의 성능을 평가하는 기존의 연구에서 범하는 중요한 오류 중의 하나는 질의 점의 분포를 객체 분포와 독립적으로 만드는 것이다. 그러나 많은 실제 환경에서 질의 점들은 대부분 원하는 객체 주변에서 발생하게 된다 [Pag1993]. 특히, 고차원 벡터 공간에서 최근접 질의가 의미를 갖기 위해서는 질의 점은 반드시 객체 클러스터 내부 혹은 주변에서 발생되어야 한다[Bey1998]. 본 연구에서는 이를 반영하기 위하여 질의 점의 분포가 전체 벡터 공간상에 균일하게 분포하는 기능과 데이터베이스에 저장

된 객체 분포를 따르도록 하는 기능을 모두 HDDQ_ Gen이 지원하도록 한다.

3. Singular Value Decomposition

본 장에서는 HDDQ_Gen의 동작을 이해하기 위한 기본 배경 지식으로서 Principal Component Analysis에서 사용되는 SVD(singular value decomposition)의 정의를 간략히 소개하고, 본 연구에서 채택된 의미에 관하여 논의한다.

3.1 SVD의 정의

M × N 행렬 X(= x_{ij})에 대하여 열(column)의 평균 u(= u_j)를 다음과 같이 정의하자.

$$u_j = (1/M) \sum_{i=1}^{M} x_{i,j}, 1 \le j \le N$$

또한, 1_M을 모든 요소의 값이 1인 M × M의 행렬이라 하자.

SVD는 다음과 같이 행렬 $X - 1_M \cdot u^T$를 M × N의 column-orthonormal[1] 행렬 U, M × M의 대각(diagonal) 행렬 S, 그리고 M × M의 행렬 V의 곱으로 표현한다[Jol1986].

$$X - 1_M \times u^T = U \cdot S \cdot V^T \qquad (1)$$

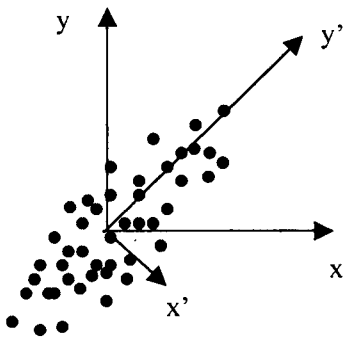
X의 공분산(covariance) 행렬 M × M의 C(= c_{ij})[2]는 이를 이용하여 다음과 같이 표현된다. 여기서 Λ는 M×M의 대각 행렬이다. 또한, Λ와 V는 행렬 C의 아이겐 값(eigen value)와 아이겐 벡터(eigen vector)를 포함한다.

$$C = (1/M) \cdot X^T \cdot X - u^T \cdot u = V \cdot \Lambda \cdot V^T \qquad (2)$$

3.2 SVD의 의미

위의 정의에서 나타난 X는 M개의 차원을 가지는 N개의 객체들을 표현하며, C는 이러한 객체들로 구성되는 객체 집합에서 서로 다른 두 차원간에 존재하는 공분산들을 요소로서 포함한다. SVD는 (식) 2를 통하여 주어진 객체 집합의 경향을 파악할 수 있는 정보를 행렬 V와 Λ를 통하여 제시한다. 행렬 V에는 객체 집합의 분포에서 축들간의 상관 관계를 최소화하는 새로운 축 시스템(axis system)을 나타나며, 행렬 Λ에는 이러한 새로운 축 시스템 내의 각 차원에 대해서 객체들이 얼마나 넓게 분포되고 있는지에 대한 정보가 나타난다.

(그림 3-1)은 두 개의 차원을 가지는 벡터 공간에서 객체들이 분포하는 상황을 나타낸 것이다. 각 점들은 2차원 공간내의 객체 위치를 의미하며, x와 y는 원래의 축 시스템을 의미한다. SVD에 의하여 이 분포에 적합한 축 시스템 x'과 y'을 행렬 V로부터 얻을 수 있다. 또한, 객체들이 새로운 각각의 축 x'과 y'에 대하여 얼마나 넓게 분포하는가에 대한 정보를 Λ로부터 얻을 수 있다.



(그림 3.1) SVD의 의미

1) I를 identity 행렬이라 할 때, $U^T \cdot U = I$임을 의미한다.

2) 여기서 $c_{i,j} = \frac{\sum_{k=1}^{N} x_{k,i} \times x_{k,j}}{N} - \left(\frac{\sum_{k=1}^{N} x_{ki}}{N} \times \frac{\sum_{k=1}^{N} x_{kj}}{N} \right)$로 정의된다.

4. 제안하는 기법

본 절에서는 HDDQ_Gen에 관하여 상세히 설명한다. 제 4.1절에서는 HDDQ_Gen에서 허용하는 제어 인자에 대하여 설명하고, 제 4.2절에서는 HDDQ_Gen 알고리즘을 제시한다. 제 4.3절에서는 HDDQ_Gen에 의하여 생성된 데이터 및 질의 집합의 예를 제시한다.

4.1 제어 인자

- *numDims* : 데이터 집합 및 질의 집합의 차원 수를 의미한다.
- *numObjects* : 데이터 집합에 속하는 총 객체 수를 의미하며, 데이터 집합의 크기를 결정한다.
- *struct_numObjInCluster* : 각 클러스터 내에 속하는 객체 수를 결정한다. 입력 값으로서 (최대값, 최소값)의 쌍을 주게 되며, 수행 시에는 각 클러스터에 대하여 최소값 및 최대값 사이의 임의의 값을 랜덤하게 할당시킨다. 따라서 같은 데이터 집합 내에서 서로 다른 클러스터는 서로 다른 수의 객체들을 포함한다.
- *struct_objDistInCluster* : 각 클러스터 내에 속하는 객체들의 분포를 결정한다. uniform, normal, exponential 중의 하나를 선택할 수 있다. uniform인 경우에는 각 차원에 대하여 (최소값, 최대값)의 쌍을 입력으로 받아 이 범위 내에서 임의의 한 구간을 랜덤하게 선택하고, 이 구간 내에서 균일 분포를 생성한다. normal인 경우에는 각 차원에 대하여 표준 편차의 (최소값, 최대값)의 쌍을 입력으로 받아 이 범위 내에서 한 값 s를 랜덤하게 선택하고, 평균 0, 표준 편차 s인 정규 분포를 생성한다. exponential인 경우에는 각 차원에 대하여 평균의 (최소값, 최대값)의 쌍을 입력으로 받아 이 범위 내에서 한 값 a를 랜덤하게 선택하고, 평균 a인 지수 분포를 생성한다. 세 가지 경우에서 모두 차원 값들은 상호 독립적이다.
- *struct_clusterDist* : 클러스터의 중심점들의 분포를 결정한다. uniform, normal, exponential 중의 하나를 선택할 수 있다. 즉, 클러스터들의 중심점들이 struct_ObjDistInCluster에서와 마찬가지 방식으로 균일 분포, 정규 분포, 지수 분포를 가질 수 있도록 한다.
- *queryRatio* : 전체 객체 수에 대하여 생성될 질의 점의 수를 %로 나타낸다. 예를 들어, 이 값이 10인 경우에는 생성될 객체 수의 10%에 해당되는 질의 점들이 생성된다.
- *queryDist* : 질의 점들의 분포를 결정하며, 입력으로서 'independent'와 'dependent' 중 한 값을 줄 수 있다. 'independent'는 질의 점들이 객체 분포와 독립적으로 벡터 공간에서 균일하게 분포하도록 하며, 'dependent'는 질의 점들이 객체의 분포를 따르도록 한다.

4.2 HDDQ_Gen 알고리즘

본 절에서는 HDDQ_Gen 알고리즘에 관하여 상세히 설명한다. 알고리즘 1은 HDDQ_Gen 알고리즘을 C 스타일로 스케치한 것이다.

HDDQ_Gen은 제 4.1절에서 설명한 제어 인자들을 입력으로 받고, 객체 집합과 질의 집합을 서로 다른 두 파일 dataFile 및 queryFile에 각각 출력한다. numObjInCluster와 numQueriesInCluster는 알고리즘 내부에서만 사용되는 변수를 의미한다.

HDDQ_Gen은 원하는 만큼의 객체들을 모두 생성할 때까지(라인 1), 클러스터 단위로 객체들과 질의 점들을 생성한다. 라인 2~5에서는

```
Algorithm HDDQ_Gen:
Input: numDims, numObjects, struct_numObjInCluster, struct_objDistInCluster, struct_clusterDist,
       queryRatio, struct_queryDist;
Output: dataFile, queryFile;
Local variable: numObjInCluster, numQueriesInCluster;
1  while (numObjects > 0) {
2      determine numObjInCluster, the number of objects in the cluster using struct_numObjects;
3      numQueriesInCluster = numObjInCluster * queryRatio;
4      get centerPoint, the center point of the cluster using struct_clusterDist;
5      determine the axis system for this cluster;
6      while (numObjects > 0 && numObjInCluster > 0) {
7              generate an object belonging to the cluster using struct_objDistInCluster;
8              adjust the object to the axis system of the cluster;
9              shift the object so that all the objects in the cluster are centered around the centerPoint;
10             output the object into dataFile;
11             numObjects--, numObjInCluster--;
       }
12     while (numQueriesInCluster > 0) {
13             generate a query point belonging to the cluster using struct_objDistInCluster;
14             adjust the query point to the axis system of the cluster;
15             shift the query point so that all the objects in the cluster are centered around the
               centerPoint;
16             output the query point into queryFile;
17             numQueriesInCluster--;
       }
   }
```

(알고리즘 1) HDDQ_Gen

먼저 클러스터의 특성을 결정한다. 라인 2에서는 해당 클러스터를 위하여 생성될 객체 수를 결정하며, 라인 3에서는 이 객체 수에 비례하여 생성될 질의 점의 수를 결정한다. 라인 4에서는 이 클러스터의 중심점의 위치를 결정한다. 이러한 결정들은 사용자에 의하여 주어지는 제어 인자들에 의하여 제어된다.

라인 5에서는 해당 클러스터가 서로 다른 차원간에 상관 관계를 가지도록 하기 위하여 사용할 축 시스템을 결정한다. 대상이 되는 벡터 공간의 차원 수가 M일 때, 이 축 시스템은 orthonormal한 M개의 벡터로 구성된다. 이러한 벡터를 랜덤하게 생성하는 것은 용이한 일이 아니므로 본 연구에서는 이의 생성을 위하여 제 3장에서 소개한 SVD를 이용한다. 즉, orthonormal한 M개의 벡터의 생성을 위하여 먼저, M×M의 행렬내의 각 요소 값들을 랜덤하게 생성한다. 이 행렬이 가상의 공분산 행렬의 역할을 하게 된다. 이 행렬에 대하여 SVD를 수행함으로써 (식) 2의 결과를 얻을 수 있으며, 이 결과 orthonormal한 M개의 벡터로 구성되는 축 시스템을 구할 수 있다.

라인 6~11에서는 이러한 클러스터 특성에 맞는 객체들을 생성한다. 먼저, 라인 7에서는 사용자가 struct_objDistInCluster를 통하여 제시한 분포 특성에 맞는 객체를 생성한다. 라인 8에서는 이 객체를 라인 5에서 결정된 새로운 축 시스템에 맞도록 조정한다. 라인 9에서는 조정된 객체를 라인 4에서 결정된 클러스터 중심점에 맞게 이동시킨다. 라인 10에서는 최종적으로

생성된 객체를 dataFile에 출력시킨다. 실제로 라인 8과 9는 다음과 같은 간단한 행렬식으로 표현할 수 있다[Will979].

$$y = A + x \cdot B \quad (3)$$

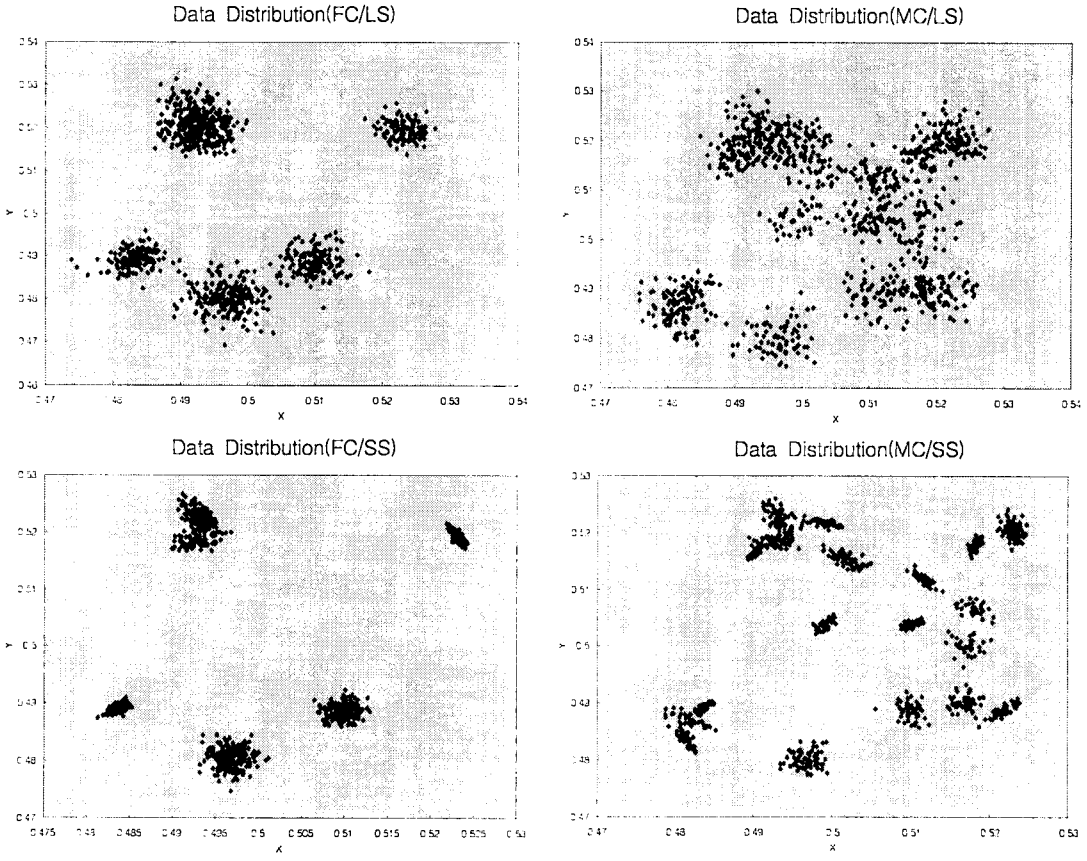
M차원의 벡터 공간을 고려할 때, 여기서 x는 1×M의 행렬이며, 라인 7에서 생성된 객체이다. B는 M×M의 행렬이며, 라인 5에서 생성된 클러스터의 축 시스템이다. A는 1×M의 행렬이며, 라인 4에서 생성된 클러스터의 중심점이다. 끝으로, y는 1×M의 행렬이며, 최종적으로 생성된 객체이다.

라인 12~17에서는 이러한 클러스터 특성에 맞는 질의 점들을 생성한다. 동작 원리는 객체 생성의 경우와 동일하며, 최종적으로 생성된 질의 점들은 queryFile에 출력된다.

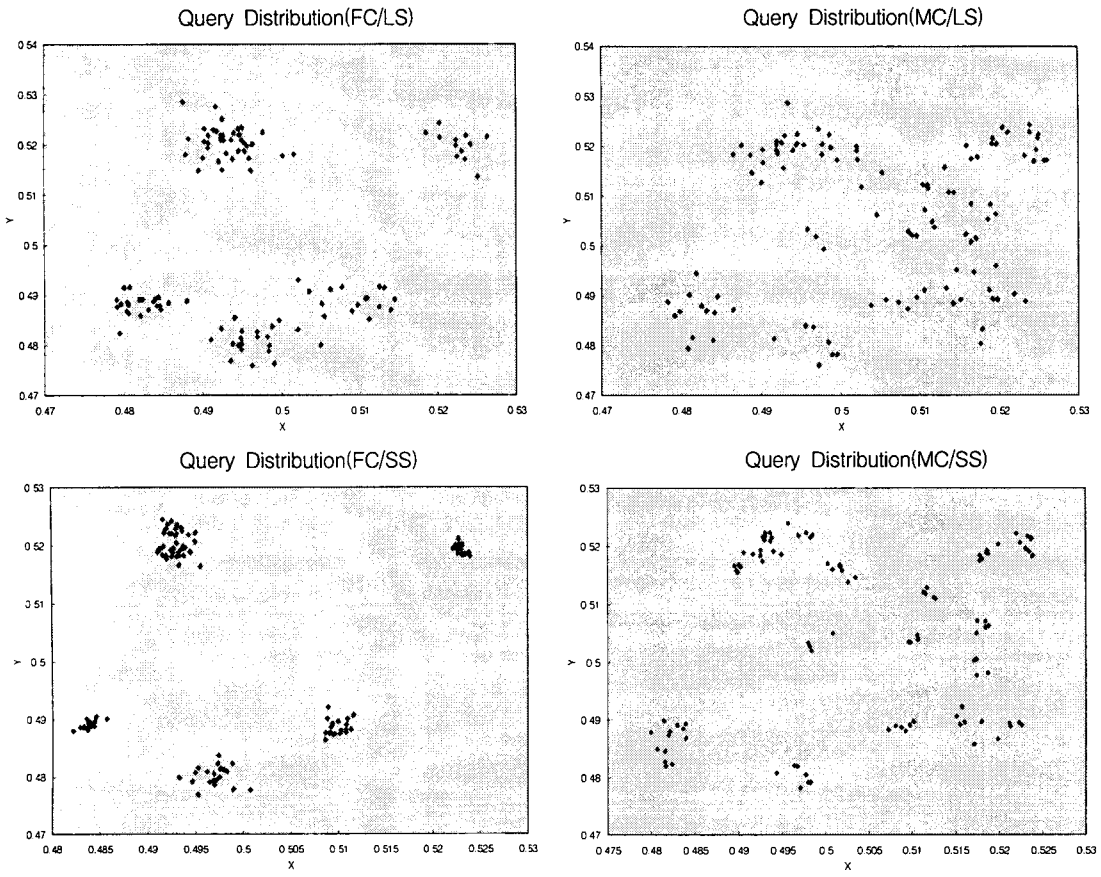
4.3 생성 예

본 절에서는 HDDQ_Gen 알고리즘의 수행으로 생성된 데이터 및 질의 집합의 예를 제시함으로써 사용자가 원하는 데이터 및 질의 집합의 특성을 적절하게 제어할 수 있음을 보인다.

<표 4.1>은 데이터 및 질의 집합의 생성에서 사용된 제어 인자를 나타낸 것이다. 10차원의 객체 1000개를 생성하는 예이다. struct_num-ObjInCluster에 나타난 MC(many clusters)는 각 클러스터에 속하는 객체 수를 [30, 70] 사이



(그림 4.1) 데이터 집합의 생성 예



(그림 4.2) 질의 집합의 생성 예

로 조정하도록 하는 옵션이다. 반면, FC(few clusters)는 각 클러스터에 속하는 객체 수를 [90, 210] 사이로 조정하도록 하는 옵션이다. 또한, struct_objDistInCluster로는 정규 분포를 취하도록 하였다. SS(small standard deviation)는 표준 편차가 전체 영역의 [0.5%, 3.5%] 사이에서 조정되도록 하는 옵션이며, LS(large standard deviation)은 표준 편차가 전체 영역의 [3.5%, 7%] 사이에서 조정되도록 하는 옵션이다. 클러스터들의 중심점들은 벡터 공간에서 균일하게 분포하도록 하였다. 질의 점의 수는 객체 수의 10%로 설정하였고, 질의 분포는 객체 분포를 따르도록 하였다.

〈표 4.1〉 데이터 및 질의 집합 생성을 위한 제어 인자의 설정.

제어 인자 명	할당 값
numDims	10
numObjects	1,000
struct_numObjInCluster	MC, FC
struct_objDistInCluster	normal (SS, LS)
struct_clusterDist	uniform
queryRatio	10
queryDist	dependent

(그림 4.1)과 (그림 4.2)는 MC, FC와 SS, LS에 의하여 가능한 네 가지 조합으로 생성된 질의 집합과 질의 집합을 각각 나타낸 것이다[3]. 먼저, 데이터 집합을 살펴보자. MC의 경우 적

은 수의 객체들을 가지는 많은 수의 클러스터들이 존재하며, 반면, FC의 경우에는 많은 수의 객체들을 가지는 적은 수의 클러스터들이 존재함을 볼 수 있다. LS가 SS에 비하여 좀더 넓은 공간에 분포하며, 클러스터별로 두 차원간의 다양한 상관 관계를 가짐을 볼 수 있다. 이와 같이, 응용 분야의 특성에 따라 이러한 인자들을 적절하게 설정함으로써 실제 데이터 분포 특성을 잘 반영하는 인위적 데이터 집합을 생성할 수 있다. 질의 집합을 보면, 데이터 집합과 유사한 특성을 가지며, 객체 분포 특성을 잘 반영하고 있음을 볼 수 있다.

4. 결 론

최근접 질의(nearest neighbor query)는 멀티미디어 데이터베이스 환경에서 질의에서 주어진 객체와 유사한 객체를 찾는 중요한 연산으로서 사용된다. 최근접 질의의 효과적인 처리를 위하여 다차원 벡터 공간내의 점들을 빠르게 검색하기 위한 다차원 색인과 이를 기반으로 한 질의 처리에 관한 많은 연구가 수행되어 왔다.

본 논문에서는 기존의 연구에서 다차원 색인 및 최근접 질의 처리 기법의 성능 평가에서 사용된 인위적 데이터 및 질의들이 실제 환경을 올바르게 반영하지 않고 있음을 밝히고, 이러한 문제점을 해결하기 위한 방안에 관하여 논의하였다. 본 논문에서는 먼저 고차원 색인 기법 및 최근접 질의 처리 기법의 성능을 공정하게 평가하기 위한 실험에서 사용될 데이터 및 질의들이 갖추어야 할 특성을 제시하고, 이러한 특성을 가지는 데이터 및 질의 집합을 인위적으로 생성할 수 있는 HDDQ_Gen(High-Dimensional Data and Query Generator) 기법을 제시하였다.

HDDQ_Gen은 (1) 클러스터 단위의 분포, (2) 클러스터 내에서의 객체들의 분포, (3) 클러스터들의 분포, (4) 차원들간의 상관 관계, (5) 객체 분포를 반영하는 질의 분포 등을 자유롭게 제어함으로써 데이터 및 질의의 분포 특성을 사용자가 다양하게 선택할 수 있다. 본 연구는 응용의 특성을 반영하는 벤치마킹 환경을 제공함으로써 고차원 색인 기법 및 최근접 질의 기법의 성능을 올바르게 평가할 수 있는 기반을 마련하였다는 점에서 의미가 있다.

향후, 이러한 HDDQ_Gen과 더불어 멀티미디어 응용에서 사용되는 다양한 실제 데이터 및 질의 집합 등을 수집함으로써 WWW 환경에서 벤치마킹에 사용할 수 있는 유용한 데이터를 관련 연구자들에게 제공하는 것을 계획하고 있다.

Acknowledgment

Sang-Wook Kim would like to thank Jung-Hee Seo, Suk-Yeon Hwang, Grace(Ju-Young) Kim, and Ju-Sung Kim for their encouragement and support.

참 고 문 헌

[Agg00] Aggarwal C., S.-W. Kim, P.S. Yu, Nearest Neighbor Search in Multimedia Databases, unpublished manuscript, 2000.

[Ary94] Arya M., et al., "QBISM : Extending a DBMS to Support 3D Medical Images," In Proc. Intl. Conf. on Data Engi-

3) 10차원의 데이터 및 질의를 생성하였으나, 이들 중 두 번째와 세 번째 차원을 선택하여 2차원 공간으로 프로젝션 하였다.

neering, IEEE, pp.314-325, 1994.

[Ber96] Berchtold, S., D.A. Keim, and H.-P. Kriegel, "The X-tree : An Index Structure for High-Dimensional Data," In Proc Intl. Conf. on Very Large Data Bases, VLDB, pp.28-39, 1996.

[Ber98] Berchtold, S. et al., "Fast Nearest Neighbor Search in High-Dimensional Space," In Proc. Intl. Conf. on Data Engineering, IEEE, pp.209-218, 1998.

[Bey98] Beyer, K. et al., "When Is Nearest Neighbor Meaningful?," In Proc. Intl. Conf. on Database Theory, pp.217-235, 1998.

[Fal94] Faloutsos, C. et al., "Efficient and Effective Querying by Image Content," Journal of Intelligent Information Systems, Vol.3, No.3, pp.231-262, 1994.

[Fal95] Faloutsos, C., "Fast Searching by Content in Multimedia Databases," IEEE Data Engineering Bulletin, Vol.18, No. 4, pp.31-40, 1995.

[Gae98] Gaede, V. and O. Guether, "Multidimensional Access Methods," ACM Computing Surveys, Vol.30, No.2, pp.170-231, 1998.

[Jag91] H.V. Jagadish, "A Retrieval Technique for Similar Shapes," In Proc. Intl. Conf. on Management of Data, ACM SIGMOD, pp.208-217, 1991.

[Jol86] Jolliffe, I.T., Principal Component Analysis, Springer-Verlag, 1986.

[Kor96] Korn, F. et al., "Fast Nearest Neighbor Search in Medical Image Databases," In Proc. Intl. Conf. on Very Large Data Bases, VLDB, pp.215-226, 1996.

[Nib93] Niblack, W. et al., "The QBIC Project : Querying Images by Content Using Color, Texture, and Shape," In Proc. Intl. Conf. Storage and Retrieval for Image and Video Databases, pp.173-187, 1993.

[Pag93] Pagel, B.-U., et al., "Towards an Analysis of Range Query Performance in Spatial Data Structures," In Proc. Intl. Symp. on Principles of Database Systems, ACM SIGACT-SIGMOD-SIGART, pp.214-224, 1993.

[Rou95] Roussopoulos, N., S. Kelley, and F. Vincent, "Nearest Neighbor Queries," In Proc. Intl. Conf. on Management of Data, ACM SIGMOD, pp.71-79, 1995.

[Sei98] Seidl, T., and H.-P. Kriegel, "Optimal Multi-Step k-Nearest Neighbor Search," In Proc. Intl. Conf. on Management of Data, ACM SIGMOD, pp.154-165, 1998.

[Web98] Weber, R., H.-J. Schek, and S. Blott, "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces," In Proc. Intl. Conf. on Very Large Data Bases, VLDB, pp.194-205, 1998.

[Whi96] White, D.A. and R. Jain, "Similarity Indexing with the SS-tree," In Proc. Intl. Conf. on Data Engineering, IEEE, pp.516-523, 1996.

[Zob96] J. Zobel, A. Moffat, K. Ramamohanarao, "Guidelines for Presentation and Comparison of Indexing Techniques," ACM SIGMOD Record, Vol.25, No.3, pp.10-15, 1996.

저자소개

김 상 욱

저자는 1989년 서울대학교 컴퓨터 공학과를 졸업하고, 1991년과 1994년에 각각 한국과학기술원에서 석사와 박사 학위를 취득하였다. 또한, 1994년에서 1995년까지 한국과학기술원 정보전자연구원에서 Post-Doc으로서 연구를 수행하였으며, 1999년에서 2000년까지 미국의 IBM T.J. Watson Research Center에서 Visiting Researcher로서 방문 연구를 수행하였다. 현재에는 강원대학교 공과대학 컴퓨터정보통신공학부에 부교수로 재직하고 있으며, 주요 관심분야는 DBMS, 실시간 주기억장치 DBMS, 트랜잭션 관리, 데이타 마이닝, 멀티미디어 정보 검색, 공간 DBMS/GIS 등이다.