웹상의 이질적 이미지 데이타베이스를 선택하기 위한 복합 추정 방법

(Hybrid Estimation Method for Selecting Heterogeneous Image Databases on the Web)

김 덕 환 † 이 석 룡 †† 정 진 완 †††
(Deok-Hwan Kim) (Seok-Lyong Lee) (Chin-Wan Chung)

요 약 웹상의 이미지 데이타베이스들은 자치성과 이질성이라는 두 가지 다른 특성을 갖고 있다. 즉 독립적으로 만들어지고 유지되며 질의 처리 방법이 서로 다르다. 분산된 이미지 데이타베이스들에 대한 내용기반 검색에서, 메타 서버의 유사성 측정함수에 대하여 서로 다른 지역 유사성 측정 함수를 갖는 데이타베이스들로부터 주어진 질의 객체와 유사한 객체들을 찾는 능력을 갖는 것은 중요하다. 현재까지, 동일한 유사성 측정 함수들을 사용하는 이미지 데이타베이스들을 선택하는 방법에 대하여 많은 연구가 진행되었으나 이미지 데이타베이스들이 다른 유사성 측정함수를 사용하는 경우에 대한 연구는 없었다. 본 논문에서는 웹상의 많은 이질적인 이미지 데이타베이스들 중 질의에 유사한 객체들을 보다 많이 가지고 있는 데이타베이스들을 찾는 문제를 다룬다. 데이타베이스들의 순위는 이미지 데이타베이스들의 압축된 히스토그램 정보와 적은 수의 표본 객체들을 사용하는 복합 추정에 기반을 두고 있다. 구형 영역 질의에 대한 선택률을 추정하기 위하여 히스토그램 정보를 사용하며, 유사성 측정 함수의 차이로 인한 선택률 오차를 보정하기 위하여 표본 객체들을 이용한다. 많은 수의 이미지 데이타에 대한 상세한 실험은 제안된 방법이 이질적인 분산 환경에서 효율적임을 보여준다.

키워드 : 내용기반 검색, 웹 데이타베이스, 유사성 질의, 데이타베이스 선택

Abstract few sample objects and compressed histogram information of image databases. The histogram information is used to estimate the selectivity of spherical range queries and a small number of sample objects is used to compensate the selectivity error due to the difference of the similarity measures between meta server and local image databases. An extensive experiment on a large number of image data demonstrates that our proposed method performs well in the distributed heterogeneous environment.

Key words : Content Based Retrieval, Web Database, Similarity Query, Database Selection

1. 서 론

전자 도서관, 의료 진단시스템, 원격 교육, 분산 출판, 전자 상거래와 같은 새로운 멀티미디어 응용의 등장은 인터넷에 분산되어 있는 이미지 데이타베이스들에 대한 검색의 필요성을 증가시킨다. 현재, 웹상에는 이미지, 비디오 프레임과 같은 객체들을 저장하는 많은 수의 데이타베이스들이 있으며, 점차적으로 시각적인 객체들을 검색하는 것이 중요해지고 있다.

질의에 적합한 튜플들을 검색하는 관계형 데이타베이스들과 같은 전통적인 데이타베이스들과 대조적으로 멀티미디어 데이타베이스들은 '유사성 질의'라고 불리는 내용기반 검색 방법을 사용하여 시각적인 객체들을 검색한다. 유사성 질의는 유사성 측정 함수를 사용하여 질의 객체 q에 유사한 시각 객체들을 검색한다. 예를 들어, $sim(q,x)$가 두 시각 객체 q와 x사이의 유사도를 [0,1] 범위의 실수로 사상하는 함수라고 하자. 값이 클수록 두 객체간의 유사도는 커진다. T가 임계값일 때, 유사성 질의는 $sim(q,x) > T$ 조건을 만족시키는 객체들의 집합을 찾는 것이다.

· 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음
† 비 회 원 : 동양공업전문대학 인터넷정보과 교수
dhkim@dongyang.ac.kr
†† 비 회 원 : 한국외국어대학교 산업정보시스템공학부 교수
sllee@hufs.ac.kr
††† 종신회원 : 한국과학기술원 전자전산학과 교수
chungcw@islab.kaist.ac.kr
논문접수 : 2002년 9월 11일
심사완료 : 2003년 9월 5일

웹과 같은 분산 환경에서의 유사성 질의를 '분산 유사성 질의'라고 한다. 웹상에는 시각적 객체들을 가지고 있는 이질적인 데이타베이스들이 있다. 그와 같은 환경에서 분산 유사성 질의를 효율적으로 처리하기 위하여 메타서버들이 필요하다. 웹상의 분산 유사성 질의의 시나리오는 다음과 같다: 사용자가 질의 객체와 전역 임계값 *GT*(Global threshold)를 메타서버에 준다. 그러면 메타서버는 사용자 질의를 이미지 데이타베이스들에 보낸다. 질의가 각 데이타베이스에서 실행된 후, 메타서버는 질의 결과를 통합하고 정렬하여 사용자에게 보여준다. 사용자 질의에 대하여 웹상의 모든 데이타베이스들을 검색한다면, 질의를 완료하는 데 너무 많은 시간이 걸릴 것이다. 그와 같은 소모적인 과정을 피하기 위하여 적은 수의 후보 데이타베이스들로 검색 범위를 좁히는 방법을 제안해야 한다. 이것이 데이타베이스 선택 문제이다. 지금까지, 텍스트 데이타베이스들에 대한 데이타베이스 선택 문제를 풀기 위해 다양한 방법이 시도되었으나 이미지 검색의 중요성에도 불구하고 이미지 데이타베이스의 선택 문제에 대한 연구는 미미한 실정이다. 본 논문에서는 웹과 같은 분산 환경에서 이미지 데이타베이스 선택 문제를 다룬다.

웹에는 다양한 종류의 이미지 데이타베이스들이 있다. 분산 유사성 질의를 통하여 효율적으로 이미지 데이타를 수집하기 위하여, 데이타베이스들의 이질성(heterogeneity) 및 자치성(autonomy)과 같은 특성과 기능들을 알아야 한다. 특히 분산 유사성 검색의 어려움은 이미지 데이타베이스들이 서로 다른 속성, 특징, 특징표현 방식과 유사성 측정 함수들을 갖는다는 점이다.

두 이미지간의 유사성은 특징 공간에서 특징 벡터들간의 거리로부터 유도된다. 특징 벡터는 색상(color), 질감(texture)과 모양(shape) 등의 속성에 대하여 정의된다. 분산 환경에서 데이타베이스들은 다른 속성들을 가질 수 있다. 색상의 경우 색상 히스토그램, 평균 색상, 대표 색상과 같은 색상 표현에 대한 다양한 시도가 있다[1]. 색상 히스토그램은 가장 널리 사용되는 시각적 특징들중의 하나이다. RGB, HSV, YCbCr과 같은 다양한 색상 공간들이 이미지들의 색상 히스토그램을 나타내기 위하여 사용될 수 있다. RGB, YCbCr색상 공간에서 유클리디안 거리가 거리함수로 사용된다. HSV공간에서는 특징들이 극좌표에서 콘의 형태로 표현되기 때문에 색상의 유사도를 측정하기 위하여 각도 거리(angular distance)가 사용된다[2]. 다른 사이트들에 위치한 이미지 데이타베이스들이 서로 다른 유사성 측정 함수들을 가지는 것은 당연하다. 질의가 이질적인 이미지 데이타베이스들에서 실행되기 전에 유사성 측정 함수들의 차이로 인하여 질의결과의 크기가 달라지는 문제를 해결해야 한다.

대부분의 이미지 데이타베이스들은 특정 메타서버에 의해서 제어되지 않고 자치적으로 데이타를 관리한다. 그들은 이미지 데이타를 저장하기 위한 자신의 기억장치와 색인 구조들을 가지며 유사성 질의를 최적화하기 위한 요약 데이타를 갖는다. 이 요약 정보는 숨겨져서 다른 사이트에게 제공되지 않는다. 그러나, 메타 서버는 분산 유사성 질의를 효율적으로 수행하기 위하여 이미지 데이타베이스들로부터 수집된 메타 데이타를 유지해야 한다. 이는 자치적인 이미지 데이타베이스들로부터 질의 처리를 위해 필요한 정보를 수집하는 아키텍쳐가 필요함을 의미한다. 이를 위한 다양한 방법이 있다. 예를 들어, 메타서버가 정보를 수집할 수 있도록 기존의 이미지 데이타베이스들이 미리 정의된 질의 인터페이스들을 갖도록 한다. 또는, 메타서버가 에이젼트 프로그램을 각 이미지 데이타베이스에게 보내어 분산 유사성 질의에 필요한 정보를 수집하는 작업을 수행하도록 한다. 이와 같은 인터페이스들이나 에이젼트 프로그램이 실행될 수 있는 데이타베이스를 반-자치적인 데이타베이스라고 한다.

본 논문에서 다루는 문제는 반-자치적이고 이질적인 많은 이미지 데이타베이스들로부터 주어진 질의에 가장 적합한 이미지 데이타베이스들을 선택하는 것이다. '적합한 데이타베이스'라는 용어는 다른 데이타베이스들보다 질의 객체에 유사한 객체들을 보다 많이 갖고 있는 데이타베이스를 의미한다. '전역 유사성 측정 함수'라는 용어는 메타서버의 유사성 측정함수를 의미한다. '지역 유사성 측정함수'라는 용어는 이미지 데이타베이스의 유사성 측정함수를 의미한다. 본 논문에서는 메타서버가 이미지 데이타베이스 db_i의 리스트 ($1 \le i \le S$), 각 데이타베이스 db_i로부터 표본 객체 o_{ij} ($1 \le j \le n$), 각 데이타베이스의 선택률 추정을 위한 압축된 히스토그램 정보(이산여현변환(*DCT*)계수들)를 메타 데이타로 저장하고 있다고 가정한다. 이제, 데이타베이스 선택 문제의 형식적인 정의는 다음과 같다.

질의 객체 q, 전역유사성 측정함수, 전역유사성 임계값 GT와 선택될 이미지 데이타베이스의 개수 M ($M \le S$)이 주어질 때, 전역유사성 측정함수 관점에서 q의 질의결과의 크기에 따라서 이미지 데이타베이스들을 선택한다.

메타서버의 데이타베이스 선택 과정은 많은 수의 원격지에 있는 이미지 데이타를 효율적으로 검색하기 위한 필수적인 작업이다. 본 논문에서는 유사성 질의의 결과 이미지의 개수(결과 크기)에 대한 복합 추정자에 기반을 둔 사용자 데이타베이스 선택 방법을 제안한다. 이 방법은 다음과 좋은 성질을 갖고 있다:

1) 메타서버의 유사성 측정함수와 지역 유사성 측정함수가 다를 때 메타서버의 유사성 측정함수 관점에서 이미지 데이타베이스들의 순위를 보다 정확히 매기도록 설계된다. 유사성 측정 함수들이 다르면 질의결과의 예측된 크기도 차이가 난다. *표본 선택률 보상* 기법의 제안을 통해 질의 결과의 예측된 크기 차이로 인한 데이타베이스선택 문제를 해결할 수 있다.
2) 작은 크기의 표본 객체들과 작은 수의 *DCT* 계수들로 질의에 적합한 객체의 수를 추정하는 효율적이고 효과적인 방법을 제안한다. 따라서 대용량 이미지 데이타베이스와 비교하여 메타데이타베이스의 공간을 작게 차지하며 히스토그램 정보의 구축시간도 적게 걸린다.
3) 제안된 방법은 이미지 데이타베이스들의 갱신을 정확히 반영한다. *DCT*를 사용한 압축된 히스토그램 정보는 이미지 데이타베이스들의 최신 정보를 유지하는 데 사용된다. 이 정보는 다차원이더라도 크기가 작고 데이타베이스의 이미지들이 변경될 때 동적으로 갱신되므로 적은 전송비용으로 이미지 데이타베이스로부터 메타서버로 전송될 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 간략히 소개한다. 3장에서는 제안된 방법을 개략적으로 기술하고 4장에서는 다른 유사성 측정 함수들간의 관계에 대한 관찰을 제시한다. 5장에서는 각 이미지 데이타베이스에 대하여 질의에 적합한 객체의 개수를 추정하는 복합적인 방법과 후보 데이타베이스를 결정하는 데이타베이스 선택 알고리즘이 설명된다. 6장에서는 실험에 대한 프레임웍을 설명하고 실험 결과들을 보인다. 7장에서는 결론을 제시한다.

2. 관련연구

텍스트 데이타베이스들에 대한 데이타베이스 선택 문제에 대하여 많은 연구가 이루어졌다[3-8]. Gravano 등[4]은 주어진 질의에 적합한 데이타베이스의 문서 개수를 추정하는 블리언과 벡터-공간 검색 모델에 기초한 키워드 기반 분산 데이타베이스 중개 시스템을 제안하였다. Callan 등[3]은 추론 네트웍에 기반한 정보 검색의 확률 모델을 제안했다. Meng 등[5,6]은 또한 확률 모델에 기초하여 텍스트 데이타베이스들의 유용성을 추정하는 방법을 제안했다. 그러나, 텍스트 데이타베이스의 벡터와 이미지 데이타베이스의 특징 벡터간에 의미적 차이가 존재하기 때문에 텍스트 데이타베이스 선택을 위한 전통적인 방법들이 이미지 데이타베이스에 직접 적용될 수 없다.

QBIC[9], Virage[10], WebSEEk[11], 그리고 VisualSEEk[12] 등과 같이 웹상에는 상이한 이미지 데이타베이스들이 있다. 이미지 데이타베이스 선택에 대한 최근의 연구는 Chang 등[13]에 의하여 이루어졌다. 지역 데이타베이스들의 이미지 군집을 대표하는 템플릿에 대한 질의의 시각적 유사도와 템플릿과 관련된 군집의 통계 데이타를 이용하는 평균-기반과 히스토그램-기반 선택 방법을 제안하였다.

평균-기반 방법은 주어진 시각적 질의에 대한 군집의 우도(likelihood)를 결정하기 위하여 (1) 표본의 개수와 (2) 템플릿에 대한 데이타베이스 이미지들의 유사도 분포의 평균과 분산을 이용한다. 히스토그램-기반 방법은 히스토그램으로 표현되는 데이타베이스 이미지들의 유사도 분포의 통계 값뿐만 아니라 이미지 군집내의 이미지들의 위치에 기반을 둔다. 그러나, 그들은 이미지 데이타베이스들이 메타서버와 같은 특징 추출방법과 거리함수들을 사용한다고 가정한다. 따라서, 웹상의 데이타베이스들이 상이한 유사성 측정함수들을 사용하므로 이 방법은 실제 환경에서 제한적으로 사용될 수 있다.

한편, Benitez 등[14]은 MetaSEEk라는 이미지들을 위한 내용기반 메타검색엔진을 제안하였다. 질의가 주어질 때, MetaSEEk는 사용자들에 의해 만들어진 적합성 피드백에 대한 과거의 데이타를 이용하여 이미지 데이타베이스들의 순위를 결정한다. 그러나, 이 방법은 특정 데이타베이스가 많이 변경되면, 로그 데이타가 동적으로 변경되지 못하므로 과거의 정보가 더 이상 유효하지 못하며 부정확하게 데이타베이스를 선택하게 된다.

3. 제안된 방법의 개략적 기술

제안된 방법은 전처리와 데이타베이스 랭킹의 두 단계로 구성된다. 각 단계에 대해 그림 1에 기술하고 설명한다.

- 전처리 단계 : 이미지 데이타베이스들의 요약 정보는 새로운 데이타베이스가 메타서버에 등록될 때 또는 등록된 데이타베이스의 내용이 크게 변경될 때 백그라운드로 수집된다. 다차원 히스토그램은 주어진 질의에 대하여 각 데이타베이스의 선택률을 추정하기 위하여 이미지 데이타베이스의 이미지 특징 데이타로부터 만들어진다. 이산 여현 변환(*DCT*)을 이용한 압축된 히스토그램 정보는 기억 장소 부담과 네트웍 전송 시간을 줄이는 장점을 갖고 있다. 게다가, 메타서버는 [15, 16]의 표본 추출 방법을 약간 수정한 점진적 질의 기반 표본 추출(progressive query-based sampling) 방법을 사용하여 표본 객체들을 추출하며 표본 객체들의 특징 집합에서 상관계수, 평균, 표준편차와 같은 통계적 메타데이타를 계산한다.
- 데이타베이스 순위결정 단계: 이미지 데이타베이스 선택 문제를 해결하기 위하여 각 이미지 데이타베이스

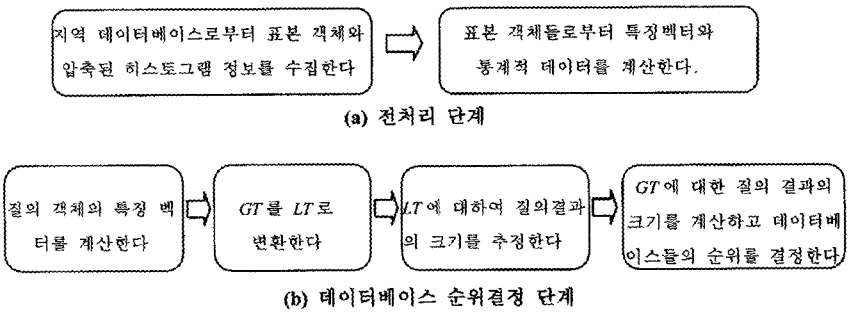

그림 1 분산 이미지 유사성 검색의 개요

의 표본 객체들과 히스토그램 정보를 사용하는 복합 선택률 추정(hybrid selectivity estimation)방법을 제안한다. 메타서버는 회귀분석을 사용하여 표본 객체들의 유사성 분포에 대한 통계 데이타를 수집한다. 이미지 데이타베이스들은 이질적인 환경에서 자신의 유사성 측정 함수를 사용하기 때문에 주어진 사용자의 전역 임계값(GT)은 통계 데이타를 사용하여 각 이미지 데이타베이스의 지역 임계값(LT)으로 변환된다. 이미지 데이타베이스의 지역 유사성 측정 함수에 의해 추정된 질의 결과의 크기는 메타서버의 전역 유사성 측정 함수에 의해 측정된 것과 다를 수 있다. 표본 선택률 보상(sample selectivity compensation) 기법은 표본 객체들을 사용하여 질의 결과의 크기 차이를 보상하기 위하여 개발되었다. GT가 주어질 때 복합 추정자에 의하여 추정된 질의 결과의 크기는 데이타베이스들의 순위를 결정하고 후보 데이타베이스들을 선택하는 기준으로 사용된다.

4. 이질적인 유사성 측정 함수들간의 관계

이장에서는 다양한 통계 데이타를 사용하여 전역 유사성 측정함수와 지역 유사성 측정함수 간의 관계를 기술한다. 전역유사도와 지역 유사도를 다음과 같이 형식적으로 정의한다:

정의 1 전역 유사도, $sim_{global}(q,o)$,는 메타서버의 유사성 측정함수에 의해 계산된 질의 이미지 q와 이미지 o간의 유사도 값에 의해 정의된다. 지역 유사도, $sim_{local}(q,o)$,는 이미지 데이타베이스의 유사성 측정함수에 의해 계산된 질의 이미지 q와 이미지 o간의 유사도값으로 정의된다.

웹상의 데이타베이스들이 이질적이므로, 유사성 질의에 사용된 속성들이 같을지라도 그들의 특징 추출방법과 거리 함수들이 서로 다를 수 있으며 유사성 측정 함수들도 달라진다. 따라서 지역 유사성 측정 함수에 의해 구해진 지역 데이타베이스의 이미지와 질의 이미지사이의 유사도 값은 메타서버의 유사성 측정함수에 의해 구해진 동일한 이미지들 사이의 유사도 값이 다를 수 있다 다음 두가지 예제는 이를 보여준다:

예 1 메타서버와 이미지 데이타베이스는 색상 속성을 사용하여 유사성 질의를 지원한다. 메타서버는 HSV색상 공간의 색상 히스토그램으로부터 평균 색상 특징을 추출하고 이미지 데이타베이스는 RGB 색상공간에서 특징을 추출한다. 메타서버는 이미지 데이타베이스가 하는 것과 같이 질의 이미지에 대해 유사도값을 측정한다. 그림 2는 4,716개의 이미지들로부터 임의로 선택된 4,716 이미지 쌍에 대하여 전역 유사도 값(y축)과 지역 유사도값(x축)들의 산포도를 보여준다. 이 경우에 산포도는 직선 형태로 나타난다.

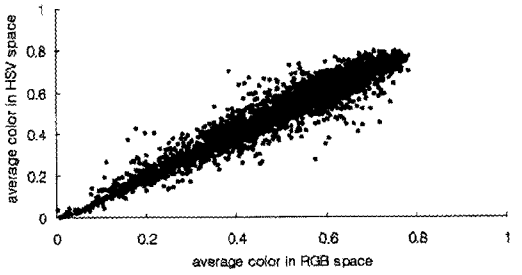

그림 2 RGB공간의 평균 색상과 HSV 공간의 평균색상 간의 산포도

예 2 질의 서버와 이미지 데이타베이스는 질감 속성을 이용한 유사성 검색을 지원한다.메타서버는 HSV 색상 공간의 색상 히스토그램의 2차 모멘트로부터 질감 특징들을 추출하며 이미지 데이타베이스는 YCbCr 색상 공간에서 질감 특징들을 추출한다. 메타서버는 이미지 데이타베이스와 마찬가지로 질의 이미지에 대하여 유사도값을 측정한다. 그림 3은 4,716 이미지들에 대한 전역 유사도값(y축)과 지역 유사도값(x축)의 산포도를 보여준다. 산포도는 직선의 형태로 나타난다.

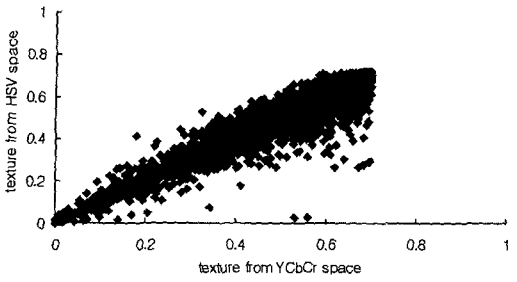

그림 3 YCbCr공간의 질감과 HSV 공간의 질감간의 산포도

메타서버와 이미지 데이타베이스 들간에 유사성 측정 함수가 다르더라도 일부의 유사성 측정 함수들간의 유사도 값들의 산포도는 직선형태로 나타난다. 두 유사성 측정함수로 구해진 유사도 값들이 선형 관계를 만족한다면, 웹상의 분산 유사성 검색을 위해 이 성질을 사용할 수 있다. 같은 속성을 가진 두개의 다른 유사성 측정 함수들이 선형 관계를 보임을 증명할 수 없기 때문에 선형 관계를 갖는 많은 사례들을 보여주는 상세한 실험을 수행했다. 지면관계로 실험 내용은 [17]을 참조하기 바란다.

5. 복합 추정 방법

분산 유사성 질의를 위하여 질의 결과의 크기가 큰 순위로 후보 데이타베이스들을 선택하는 것이 중요하다. 본 논문에서는 이질적인 유사성 측정 함수들을 사용하는 이미지 데이타 베이스들로부터 질의 결과의 크기를 추정하기 위하여 복합 추정 방법을 사용한다. 표 1은 제안된 데이타베이스 선택 방법에서 사용된 기호들을 나타낸다.

표 1 데이타베이스 선택에서 사용된 기호들

기호	의미
DB	모든 데이타베이스들의 집합 = $\{db_1,...,db_s\}$
db_i	i번째 이미지 데이타베이스
q	질의
M	선택될 데이타베이스들의 개수
GT	사용자가 제시한 전역 임계값
LT_i	GT에 대응하는 i번째 이미지 데이타베이스의 지역 임계값
n	확률 표본 객체들의 개수
y	전역 유사도값
x	지역 유사도값

5.1 제안된 방법의 형식적인 정의

복합 추정 방법의 형식적인 정의는 다음과 같다. i번째 이미지 데이타베이스의 이미지 집합을 I_i 라고 하자.

정의 2 질의 q와 GT에 대하여 i번째 이미지 데이타베이스의 전역 질의 결과의 크기 $gnum(db_i, q, GT)$는 다음과 같이 정의된다:

$$gnum(db_i,q,GT) = |\{o \in I_i \mid sim_{global}(q,o) \geq GT\}| \quad (1)$$

정의 3 질의 q와 LT_i에 대하여 i번째 이미지 데이타베이스의 지역 질의 결과의 크기 $lnum(db_i, q, LT_i)$는 다음과 같이 정의된다:

$$lum(db_i,q,LT_i) = |\{o \in I_i \mid sim_{local}(q,o) \geq LT\}| \quad (2)$$

정의 4 질의 q와 GT에 대하여 i번째 이미지 데이타베이스의 전역 질의 선택률 $gsel(db_i, q, GT)$은 다음과 같이 정의된다:

$$gsel(db_i,q,GT) = gnum(db_i,q,GT) / |I_i| \quad (3)$$

여기서 $|I_i|$는 데이타베이스 db_i의 객체의 개수이다.

정의 5 질의 q와 LT_i에 대하여 i번째 이미지 데이타베이스의 지역 질의 선택률 $lsel(db_i, q, LT_i)$는 다음과 같이 정의된다:

$$lsel(db_i,q,LT_i) = lum(db_i,q,LT_i) / |I_i| \quad (4)$$

데이타베이스들의 최적 순위(ideal rank)는 질의 q에 대하여 $gnum(db_i, q, GT)$에 따라서 데이타베이스들을 정렬하여 구할 수 있다. 그러나 이미지 데이타베이스는 전역 특징이 아닌 지역 특징의 히스토그램을 제공하기 때문에 메타서버에서 $gnum(db_i, q, GT)$의 계산이 가능하지 않을 수도 있다. 따라서 $gnum(db_i, q, GT)$과 $lnum(db_i, q, LT_i)$사이의 차이를 보상할 필요가 있다. 이를 위하여, 두 가지 선택률 보상(selectivity compensation) 방법을 다음과 같이 정의한다:

정의 6 i번째 이미지 데이타베이스의 모집단 선택률 보상(PSC_i)은 전역질의 선택률과 지역 질의 선택률의 차이로서 다음과 같이 정의된다:

$$PSC_i(db_i,q,GT) = gsel(db_i,q,GT) - lsel(db_i,q,LT_i) \quad (5)$$

그러나 PSC_i의 정확한 값을 계산할 수 없다. 대신에 PSC_i를 추정하기 위하여 표본 선택률 보상을 정의한다.

$gsel_{sample,n}(db_i,q,GT)$이 크기 n의 임의 표본을 사용하여 추정된 전역 질의 선택률이라 하고 $lsel_{sample,n}(db_i, q, LT_i)$를 크기 n의 임의 표본을 사용하여 추정된 지역 질의 선택률이라고 하자. GT와 LT_i 간에 표본 객체들의 지역 유사도와 전역 유사도에 대해 선형 회귀분석을 적용하여 선형 관계식 $GT = \hat{\alpha}_i + \hat{\beta}_i \times LT_i$을 얻을 수 있다. $\hat{\alpha}_i$와 $\hat{\beta}_i$는 추정된 회귀직선의 계수이다.

정의 7 i번째 이미지 데이타베이스의 표본 선택률 보상($SSC_{i,n}$)은 질의 q의 $gsel_{sample,n}(db_i, q, GT)$와 $lsel_{sample,n}(db_i, q, LT_i)$간의 차이를 보상하기 위한 오프셋 값으로 다음과 같이 정의된다:

$$SSC_{i,n}(db_i,q,GT) = gsel_{sample,n}(db_i,q,GT) - lsel_{sample,n}(db_i,q,LT_i) \quad (6)$$

$lsel'(db_i, q, LT_i)$이 이미지 데이타베이스 db_i의 히스토그램 정보를 사용하여 추정된 지역 질의 선택률이다. 질의 q를 위하여 추정된 db_i의 전역 질의 선택률, $gsel'$, 은 다음과 같은 복합 추정자로부터 구해질 수 있다:

$$gsel'(db_i,q,GT) = lsel'(db_i,q,LT_i) + SSC_n(db_i,q,GT) \quad (7)$$

마찬가지로, 질의 q를 위하여 추정된 db_i의 전역 질의 결과의 크기, $gnum'$,은 다음과 같이 구해진다:

$$gnum'(db_i,q,GT) = gsel'(db_i,q,GT) \times |I_i| \quad (8)$$

$gsel'(db_i, q, GT)$과 $gnum'(db_i, q, GT)$는 각각 $gsel(db_i, q, GT)$과 $gnum(db_i, q, GT)$의 근사값으로 사용된다.

5.2 순수한 표본추출 기반 질의 크기 추정 방법

히스토그램 정보와 함께 작은 수의 표본 객체들을 사용하기 때문에 제안된 데이타베이스 선택 방법은 복합 추정자(hybrid estimator)를 사용하는 방법이라고 볼 수 있다. 전역 유사성 측정 함수에 따라서 질의 결과를 추정하는 대안으로 순수하게 표본만을 이용한 추정자(pure sampling-based estimator)가 이용될 수 있다. $gsel_{sample,n}(db_i, q, GT)$ 만을 사용하여 데이타베이스들의 순위를 정한다. 이 방법은 임의의 데이타 분포에 대해 합리적으로 질의 크기를 추정하며 추출된 표본의 크기에 비례하여 크기 추정의 정확도가 높아진다. 추출된 표본의 크기와 정확도 사이에 상호 보완적 관계(trade off)가 있다. 표본 추출기반 방법은 확률 표본 추출이 질의 처리 시간에 발생하기 때문에 표본 추출을 위한 실행시간 비용이 크며, 추정의 정확성을 높이기 위해서 충분한 크기의 표본들이 필요하므로 메타 데이타베이스의 기억장소를 더 많이 요구하게 된다. 복합추정자를 사용한 방법은 순수한 표본 추출기반 방법보다 작은 크기의 표본들을 사용할 지라도 정확도가 더 높다. 인터넷상의 표본 추출은 높은 비용이 들기 때문에 전처리 단계에서 점진적인 질의 기반 표본 추출 방법을 사용한다.

5.3 데이타베이스 히스토그램 정보

주어진 질의에 적합한 이미지 데이타베이스들을 선택하기 위하여 이미지 데이타베이스들의 질의 결과의 크기를 추정할 수 있는 요약자료가 필요하며, 이는 주로 다차원 히스토그램을 사용하여 구현된다. 이미지 데이타베이스들에 대하여, 이미지 객체들의 특징 벡터들로부터 다차원 히스토그램들이 만들어진다. 이미지 데이타베이스 db_i의 객체 o의 특징값은 실수 공간$[-\infty,\infty]$의 값이다. 그러나 이미지 데이타베이스 db_i의 내용 요약인 히스토그램 정보는 차원 p에서의 정규화된 데이타 공간$[0, 1]^p$에서 생성된다. 따라서, 히스토그램 정보를 구축하기 위하여 각 데이타베이스의 특징 데이타는 다음과 같이 정규화되어야 한다:

객체들의 개수가 충분히 클 때 일반적으로 정규 분포를 가정할 수 있다. 그리고 정규화는 특징값을 평균과 표준편차에 의하여 영역 [0, 1]내로 값을 변환하는 전처리 작업이며 임의의 데이타 분포에 대해서도 적용할 수 있다. 객체 o에 대하여, $F=\{f_1, f_2, \cdots, f_k, \cdots, f_p\}$를 p차원 특징 벡터라고 하자. f_k는 특징 벡터 F의 k번째 특징 값이다. 데이타베이스 db_i에 N객체들이 있다면 $N \times p$특징 행렬 $\Im$을 구성할 수 있다. $\Im$의 각 $\Im_k$열은 길이 N의 k번째 특징 값들의 집합이며 집합 $\Im_k$의 평균 m_k과 표준편차 σ_k을 각각 계산할 수 있다. 차원 p의 특징 벡터에 대하여 특징간 정규화를 수행할 수 있으며 다음과 같이 형식적으로 정의된다:

$$f_k' = \frac{(f_k - m_k)/3\sigma_k + 1}{2} \quad (9)$$

위의 수식에 의하면, 특징값이 영역 [0, 1]내로 들어갈 확률은 약 99퍼센트이다. 집합 매개변수 $m_1, m_2, \cdots, m_k, \cdots, m_p$와 $\sigma_1, \sigma_2, \cdots, \sigma_k, \cdots, \sigma_p$가 메타 데이타로 저장된다. 그들은 실수 데이타 공간의 점들을 정규화된 데이타 공간의 점들로 변환하는 데 사용될 수 있다. HSV색상 특징들의 히스토그램 정보를 구성하기 위하여 극좌표 체계의 특징값을 rectangular좌표 체계의 특징값으로 변환한다. 정규화된 데이타 공간은 여러 개의 사각형 버켓들로 분할되며 버켓과 관련된 빈도수는 DCT 기법을 이용하여 압축된다. DCT 계수를 히스토그램 정보라고 언급한다.

또한 이미지 데이타베이스의 최신 정보를 유지하는 것이 중요하다. 기존의 히스토그램 기반 방법[13]은 데이타 갱신 횟수가 어떤 임계값에 도달하면 정확성이 낮아지기 때문에 히스토그램을 전체적으로 재구성해야 한다. 이에 반해, 제안된 방법은 DCT의 선형적인 성질[18]을 이용하여 변경된 데이타만 적용하여 히스토그램을 재구성할 수 있어 적은 부담으로 동적인 데이타 변경을 지원할 수 있다. 삽입된 객체들의 개수가 어떤 임계치에 도달할 때, 삽입된 객체들을 위한 새로운 DCT 계수값이 계산되어, 메타서버로 전송되며, 데이타베이스 히스토그램의 기존의 DCT 계수들에 더해진다.

5.4 구형 유사성 질의를 위한 선택률 추정

이미지 데이타베이스의 유사성 질의는 실제 데이타 공간에서 구형(hyper-sphere)으로 표현된다. 이전의 연구[18]에서, DCT 기법을 이용하여 압축된 히스토그램 정보를 사용하여 장방형 영역 질의를 위한 다차원 선택률 추정 방법을 제안하였다. 이 방법에 의하면 데이타 공간이 $(0,1)^p$로 정규화 된다고 가정할 때 p차원 장방형 영역질의의 선택률 s이 구해진다. 그러나 이 방법은 장

방형 질의의 선택률만을 추정할 수 있으므로, 구형 질의 q는 실제 데이타 공간에서 장방형으로 근사 되어야 한다. 이 절에서는 구형 유사성 질의의 질의 결과를 추정하기 위하여 DCT를 이용한 히스토그램 정보에 기초한 두 가지 선택률 추정 방법을 제안한다:

- **단일 사각형 근사**(Single Rectangle Approximation; **SRA**): 질의 q의 선택률은 질의 q의 hyper-sphere와 같은 볼륨과 중심을 갖고 있는 hyper-rectangle의 선택률로부터 구해진다. 근사 된 단일 hyper-rectangle은 그림 5에 점선으로 표현된다.
- **다중 사각형 근사**(Multiple Rectangle Approximation: **MRA**): 중심이 $(0, \cdots, 0)$이고 반경이 1인 hyper-sphere안에 v개의 hyper-rectangle R_i i=1, ⋯, v을 미리 생성한다. Hyper-rectangle의 개수를 적게 하고 hyper-sphere와 hyper-rectangle들 간의 공유 부분의 비율을 높이기 위하여, 겹쳐진 영역은 많아야 두개의 hyper-rectangle들로부터 만들어진다는 조건을 준다. 겹쳐진 영역은 다른 hyper-rectangle로 고려한다. s_i는 hyper-rectangle R_i의 선택률이고 λ_i는 (hyper-sphere와 R_i의 공통영역)/(R_i의 볼륨)이며, ρ는 (hyper-sphere의 볼륨)/(R_i의 내부 영역들의 전체 볼륨)이다. 새로운 hyper-rectangle R_i이 hyper-sphere와 겹치는 비율 λ_i이 τ 이상이면 R_i을 취하고 이하이면 R_i을 버린다. 같은 과정을 반복하되 이미 만들어진 hyper-rectangle들과 겹치는 지 검사하여 겹치지 않도록 하는 과정이 추가된다. 차원이 증가함에 따라 hyper-sphere의 볼륨에 대한 hyper-rectangle과의 공유영역의 비율이 감소하므로 3차원일 때는 τ를 0.7, 6차원일 때는 0.5가 되도록 실험적으로 선택하여 MRA 방법이 이 값들을 사용할 때 성능이 좋았다. 이 hyper-rectangle들은 질의 영역(query range)내에 위치하도록 적절한 비율로 확대되고 변환되고 각 사각형의 선택률 s_i, ρ,와 λ_i이 계산된다. 질의 q의 선택률은 $\rho\sum_{i=1}^{v} s_i\lambda_i$ 이다.

두 가지 경우에, hyper-sphere의 볼륨과 hyper-rectangle과 hyper-sphere의 공통부분의 볼륨이 계산되어야 한다. 그러나, 경계 효과[19]가 일어날 수 있으므로 실제 데이타 공간에서 hyper-sphere의 볼륨을 해석적으로 계산하기는 어렵다. 이 경우에, Monte-Carlo 방법[20]을 적용하여 다음과 같이 근사적으로 볼륨을 계산할 수 있다: hyper-sphere의 볼륨 = hypere-sphere를 외접하는 hyper-rectangle의 볼륨 * (hyper-sphere 내의 난수들의 개수 / hyper-rectangle내의 난수들의 전체 개수).

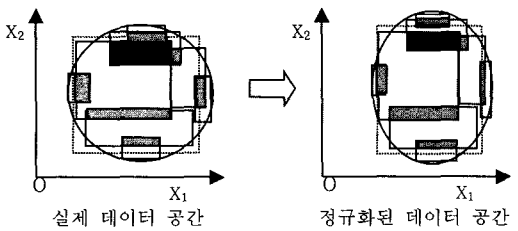

그림 4 실제 데이타 공간과 정규화된 데이타 공간

이미지 데이타베이스들의 히스토그램 정보는 이미지 데이타베이스의 특징 벡터의 차원이 p일 때 정규화된 데이타 공간 $[0,1]^p$에서 생성된다. 그림 4에서 보여준 바와 같이, 실제 데이타 공간에서 hyper-sphere와 hyper-rectangle은 정규화된 데이타 공간에서 식 (9)를 사용하여 각각 hyper-oval과 hyper-rectangle로 변환되어야 한다. 그 다음에 구형 유사성 질의의 선택률은 식 (11)을 사용하여 추정될 수 있다.

5.5 표본 선택률 보상

다른 유사성 측정 함수들은 $gnum(db_i, q, GT)$와 $lnum(db_i, q, LT_i)$사이에 차이 $Diff_i$를 만든다. $Diff_i$를 추정하기 위하여 식 (7)에서 $PSC_i(db_i, q, GT)$가 사용될 수 있다. 그러나 PSC_i는 메타서버에서 계산될 수 없다. 대신에 이미지 데이타베이스에서 메타서버로 표본 객체들을 가져와서 특징들을 추출하고 $SSC_{i,n}$을 계산할 수 있다. 이를 위하여, 전처리 단계에서 점진적인 질의-기반 표본 추출 방법을 사용한다. 이미지 데이타베이스를 위한 상세한 표본 추출 알고리즘은 다음과 같다:

(1) 초기 질의 객체를 임의로 선택한다; $n \leftarrow 0$.
(2) 질의 객체를 i번째 이미지 데이타베이스로 보낸다.
(3) 영역질의를 사용하여 i번째 이미지 데이타베이스로부터 객체들을 검색하고, 검색된 객체들 중 임의로 δ개의 표본 객체들을 선택한다.
(4) 검색된 객체들의 특성에 기준하여 $SSC_{i,n}$을 계산한다.
(5) Stopping criterion이 아직 도달하지 않았으면,
  (a) 새로운 질의 객체를 선택한다; $n \leftarrow n + \delta$;
  (b) 단계 (2)를 실행한다.

점진적인 질의-기반 표본 추출 방법은 작은 크기의 표본들로 시작해서 $SSC_{i,n}$의 정확성이 개선되는 비율이 매우 작아질 때까지 점진적으로 표본의 크기를 증가시킨다. $SSC_{i,n}$을 사용하는 추정 방법의 정확성을 보이기 위하여, $SSC_{i,n}$이 PSC_i로 확률적으로 수렴함을 보여야 한다.

i번째 이미지 데이타베이스 db_i로부터 크기 n의 표본 객체들을 사용하여 질의의 $SSC_{i,n}$을 계산하기 위하여, 표본 회귀직선 계수(α, β)와 표본 결정계수(r^2)와 같은 통계적 메타데이타가 사용된다.

그림 5에 보여준 것처럼 표본 객체들을 사용하여 지

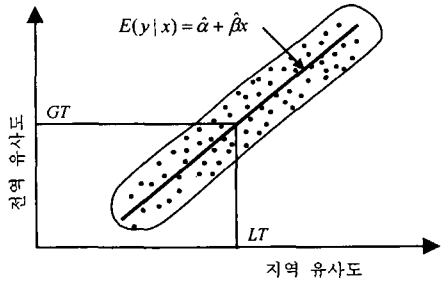

그림 5 회귀 직선

역 유사도값 x와 전역 유사도값 y의 이변량 분포로부터 회귀 직선 $E(y/x)=\hat{\alpha}+\hat{\beta}x$을 추정할 수 있다. 추정된 회귀직선상의 점들$(x_i, E(y|x_i)).i=1, \cdots, n$을 임계점이라고 한다. 다음 lemma와 정리는 표본 크기 n이 증가할수록 $SSC_{i,n}$가 상수값에 확률적으로 수렴함을 보여준다.

Lemma 1 지역 유사도 임계값 x가 주어질 때 크기 n의 확률 표본을 이용하여 추정된 지역 질의 선택률은 표본 크기 n이 커짐에 따라 db_i에 있는 모든 객체들을 사용하여 추정된 지역 질의 선택률에 확률적으로 수렴한다.

Lemma 2 전역 유사도 임계값 $E(y|x)$가 주어질 때 크기 n의 확률 표본을 이용하여 추정된 w전역 질의 선택률은 표본 크기 n이 커짐에 따라 db_i에 있는 모든 객체들을 사용하여 추정된 전역 질의 선택률에 확률적으로 수렴한다.

Theorem 1 질의와 임계점$(x, E(y|x))$이 주어질 때 표본 크기 n이 커짐에 따라 $SSC_{i,n}$은 PSC_i에 확률적으로 수렴한다.

위의 Lemma들과 Theorem에 대한 증명은 [17]을 참조하기 바란다. 이제, 표본 객체들의 크기 n을 어떻게 결정하는 지를 설명한다. $SSC_{i,n}$은 n이 커짐에 따라 PSC_i에 수렴하므로 $SSC_{i,n}$과 $SSC_{i,n+\delta}$는 $SSC_{i,n}$이 충분히 정확해지면 PSC_i에 근접한다. 실제로, $|SSC_{i,n}-SSC_{i,n+\delta}| / |SSC_{i,n}| \le 0.1$을 만족할 때 n을 결정한다.

5.6 데이타베이스 순위결정 알고리즘

제안된 데이타베이스 선택 기준은 다음과 같다:

주어진 질의에 대하여 추정된 전역 질의 결과의 크기 $gnum'(db_i, q, GT)$에 따라 순위가 결정될 때 순위 리스트 $G(q, GT)=(db_{g_1}, db_{g_2}, \ldots, db_{g_r})$안에 있는 상위 M 데이타베이스들을 선택한다.

Algorithm 데이타베이스 순위결정(q, GT, DB)

(1) for each $db_i \in$ DB, $i = 1, \ldots, S$

(2) get $\hat{y}=\hat{\alpha}+\hat{\beta}x$ using the linear regression analysis.

(3) calculate $SSC_{i,n}(db_i, q, GT)$using sample objects of size n.

(4) $LT_i = (GT - \hat{\alpha}) / \hat{\beta}$

(5) compute the estimated local query selectivity, $lsel'(db_i, q, LT)$, using histogram information.

(6) $gsel'(db_i, q, GT) = lsel'(db_i, q, LT) + SSC_{i,n}(db_i, q, GT)$

(7) $gnum'(db_i, q, GT) = gsel'(db_i, q, GT) \times |I_i|$

(8) end for

(9) Rank databases according to $gnum'(db_i, q, GT)$

단계 (2)와 단계 (3)에서 회귀분석과 $SSC_{i,n}$ (db_i, q, GT)의 계산 시간은 각 데이타베이스의 표본객체들의 크기 n에 비례한다. p가 차원이고 α가 싸인 함수의 계산 시간, z가 DCT계수들의 개수이다. MRA를 위하여 v개의 사각형들이 이용된다. 장방형 영역질의의 선택률 계산시간은 식 (11)에서 $2p\ \alpha\ z$임을 알 수 있다. 단계 (5)에서 $lsel'(db_i, q, LT_i)$의 시간 복잡도는 MRA일 때 $O(v\ 2p\ \alpha\ z)$이고 SRA일 때 $O(2p\ \alpha\ z)$이다. n개의 표본 객체들과 MRA가 사용될 때 S데이타베이스들에 대한 알고리즘의 복잡도는 $O(S(n+v\ 2p\ \alpha z))$이다.

6. 실 험

제안된 방법의 효율성을 측정하기 위하여 많은 수의 이미지 데이타와 다양한 질의들에 대한 포괄적인 실험을 수행하였다. 실험은 주어진 질의에 대하여 이미지 데이타베이스들의 순위를 정하는 선택 방법의 정확성을 보여주는 데 초점을 두었다. 시스템은 HP NetServer와 윈도우즈 NT 환경하에서 마이로소프트의 VC++로 구현되었다.

6.1 실험 환경

시험 데이타는 QBIC과 WALRUS[21] 시스템에서 사용된 이미지들을 256-색상 비트맵 형태로 변환하고 4조각으로 분할한 83,476개의 이미지로 구성된다. 시각적인 데이타베이스들은 특정한 형태의 이미지들을 포함하므로 의미적 분류에 기반하여 10개의 이미지 데이타베이스들을 구축하였다. 각 데이타베이스는 다른 특징 추출 방법과 다른 거리 함수를 사용한다. 이미지들의 특성을 묘사하는 시각적 특징들을 구하기 위하여 색상 히스토그램 방법을 이용하여 다양한 색상 공간들로부터 평균 색상과 질감을 추출하였다.

각 이미지에 대하여 평균 색상(μ_1, μ_2, μ_3)은 색상 성분의 평균 명암도를 나타내는 데 사용된다. 질감$(\sigma_1, \sigma_2, \sigma_3)$은 각 색상 성분의 상대적인 매끈한 정도(smoothness)를 나타내며 표준편차로 구해진다. 여기서, 1,2,3은 RGB, HSV 또는 $YCbCr$의 각 성분을 의미한다. 실험에서 메타서버의 특징으로서 HSV 색상공간의 평균 색상과 질감을 사용한다. 표 2는 모든 이미지 데이

표 2 시험 데이타 환경

사이트	추출된 특징들	크기	의미적 분류
1	*RGB* 색상 공간의 평균 색상/질감	8,888	풍경, 사진
2	*RGB* 색상 공간의 평균 색상/질감	8,819	동물, 동물원 모음
3	*HSV* 색상 공간의 평균 색상/질감	8,526	예술 & 설계
4	*RGB* 색상 공간의 평균 색상/질감	8,328	배경, 패턴
5	*YCbCr* 색상 공간의 평균 색상/질감	7,940	꽃, 식물
6	*YCbCr* 색상 공간의 평균 색상/질감	7,835	클립 아트
7	*YCbCr* 색상 공간의 평균 색상/질감	8,328	배경, 패턴
8	*HSV* 색상 공간의 평균 색상/질감	9,908	사람
9	*HSV* 색상 공간의 평균 색상/질감	8,328	배경, 패턴
10	*YCbCr* 색상 공간의 평균 색상/질감	8,819	동물, 동물원 모음

표 3 데이타베이스 선택을 위한 시험 매개변수들

매개변수	의미
차원 D	이미지 특징들의 차원 (D=3 또는 6)
표본 비율 SR	표본객체들의 수를 이미지 데이타베이스의 전체 객체들의 수로 나눈 비율 (SR=0.1% ~ 4.4%)
M	선택될 데이타베이스의 개수 (M=1, 2, ···, 10)
임계값 T	전역 임계값의 범위 (T=0.7, 0.6, 0.5)

타베이스들을 위한 추출된 특징들, 데이타베이스 크기, 의미적 분류를 보여준다. 데이타베이스 선택을 위한 실험은 표 3에 기술한 매개변수들을 변경하여 수행하였다.

6.2 실험 결과

복합 추정자를 위해 사용된 두 사각형 근접 방법들(SRA, MRA)을 비교하는 실험을 수행하였다. 다양한 매개변수들을 사용하여 각 실험마다 30개 질의를 실행하고 그 결과들의 평균을 구했다. 선형 회귀를 실행할 때 전역 유사도 y의 예측 구간을 추정하기 위하여 99.9% 신뢰 수준을 사용한다. *MRA*의 경우 정규 hyper-sphere를 근사하기 위하여 13개의 작은 hyper- rectange들을 미리 생성한다. 먼저 충분히 정확한 전역 선택률 추정을 제공하기 위하여 적당한 표본 객체들의 크기를 관찰한다. 정확성을 측정하기 위하여, 상대 오차(E)를 다음과 같이 정의한다:

$$E = \frac{|\text{query result size - estimated result size}|}{\text{query result size}} \times 100\%$$

그림 6과 7은 D=3과 D=6일 때 각각 *SRA*와 *MRA*의 상대 오차를 보여준다. 같은 데이타베이스(사이트 4, 사이트 7, 사이트 9)를 사용하지만 유사성 측정 함수들(*RGB, YCbCr, HSV* 색상 공간)을 다르게 하였다. *SRA-YCbCr*의 상대 오차는 D=3일 때 12.5~16.2%, D=6일 때 24~26.5%이며 *MRA-YCbCr*의 경우 D=3일 때 2.8~6.6%, D=6일 때 9.5~14.7%임을 관찰할 수 있다. D=3 또는 6일 때 같은 범위의 표본 비율에 대한 *MRA*의 상대 오차는 *SRA*의 상대 오차보다 좋은 결과를 보여준다. *SRA-HSV*의 상대 오차는 전체 범위에서 14.2%와 29.9%로 일정하게 나타난다. 그 이유는 전역유사성 측정함수와 지역 유사성 측정함수(*HSV* 색상 공간)가 동일하며 전체 범위에서 SSC=0이기 때문이다. 실험은 또한 표본 비율이 1.1%이상으로 증가하면 상대 오차의 변화가 적어짐을 보여준다. 따라서, 다음 실험에서는 표본 비율을 1.1%로 고정한다.

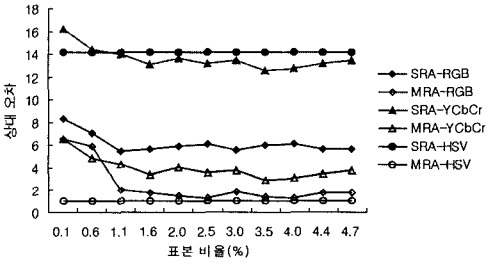

그림 6 D=3일 때 상대 오차

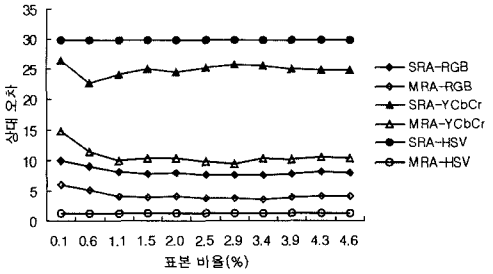

그림 7 D=6일 때 상대 오차

그림 8와 9는 주어진 2개의 질의들에 대하여 표본 선택률 보상($SSC_{i,n}$)의 결과를 보여준다. 표본 비율이 매우 작을 때 $SSC_{i,n}$의 변동이 많으나 표본 비율이 증가하면 $SSC_{i,n}$이 상수값에 확률적으로 수렴함을 관찰할 수 있다. 또한 표본 비율이 1.1% 이상일 때 $SSC_{i,n}$을 이용

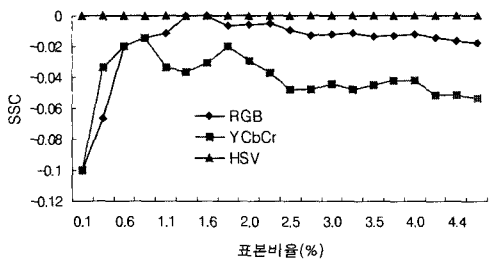

그림 8 질의 1에 대한 *SSC*

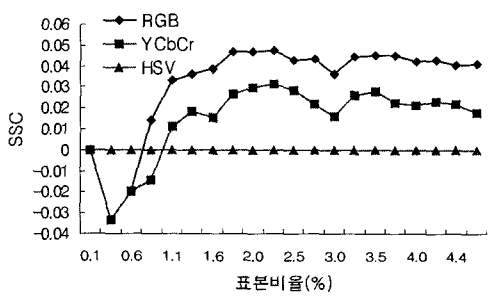

그림 9 질의 2에 대한 *SSC*

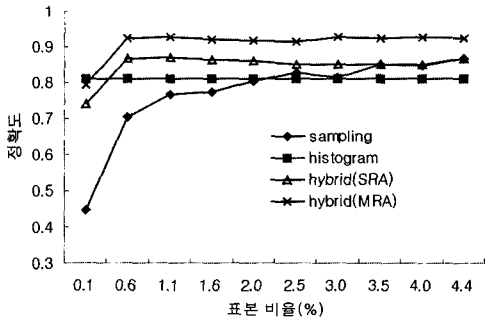

그림 10 데이타베이스 6에서의 정확성

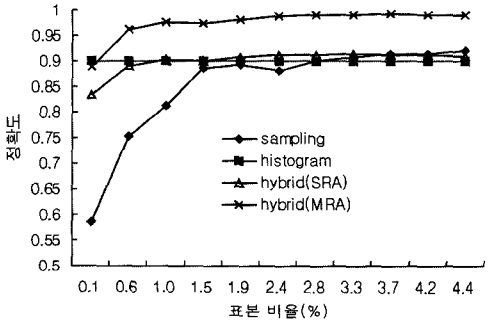

그림 11 데이타베이스 2에서의 정확성

하여 전역 선택률과 지역 선택률의 차이를 상당히 줄일 수 있음을 알 수 있다.

질의 결과의 크기를 추정하기 위하여 제안된 방법의 정확성과 효율성을 측정하기 위하여 (1) 순수한 표본 추출 방법 (2) 히스토그램 기반 방법 (3) 제안된 2가지 복합추정 방법(*SRA+SSC, MRA+SSC*)를 비교하는 상세한 실험을 수행하였다. 순수 표본 추출 기반 방법과 비교할 때 그림 10과 11은 질의 결과의 크기를 추정하는 방법의 정확성이 표본 크기가 작을 때조차도 복합 방법들에 의해 상당히 개선됨을 보여준다. 결과는 또한 복합 방법들이 동일한 정확성을 제공할 때 필요한 표본의 크기가 작아도 됨을 보여준다.

제안된 데이타베이스 선택 방법의 성능을 비교한다. 두 성능 측정 함수들이 사용된다. 첫 번째, 시험 질의들을 위한 순위 결정 오차(E')는 최적 순위값에 대하여 제안한 방법에 의하여 구해진 순위값을 비교하기 위하여 사용된다. 각 데이타베이스의 최적 순위값은 주어진 질의와 각 데이타베이스 이미지사이의 전역 유사도를 측정하여 구할 수 있다. 시험 질의 q_j의 순위 결정 오차는 다음과 같이 정의된다: $E_j' = \frac{1}{k}\sum_{db_i \in ING}(IR_{i,j} - ER_{i,j})^2$, $IR_{i,j}$는 q_j의 실제 결과의 크기에 기반을 둔 데이타베이스 db_i의 최적 순위값이며 $ER_{i,j}$는 q_j에 대한 순위 결정 알고리즘에 의해 결정된 db_i의 예측된 순위값이며, G는 제안된 데이타베이스 선택 방법에 의해 선택된 데이타베이스들의 집합이다. k개의 시험 질의들 $q_1, q_2, \cdots, q_k$가 주어질 때, E'은 수식 $E' = \frac{1}{k}\sum_{j=1}^{k} E_j'$에 의해 계산된다. 둘째로, 상대적 성능($P$)은 데이타베이스 선택 방법에 의해 구해진 전역 질의 결과의 크기와 최적 데이타베이스 선택에 의해 구해진 크기의 비율을 계산하여 데이타베이스 선택 방법의 정확도를 나타낸다. 형식적으로, 시험 질의 q_j ($1 \le j \le k$)에 의해 구해지는 상대 성능 P_j는 다음과 같이 정의된다: $P_j = \frac{\sum_{db_i \in G} gnum(db_i, q_j, GT)}{\sum_{db_i \in B} gnum(db_i, q_j, GT)}$, B는 최적으로 선택된 데이타베이스들의 집합이다. k개의 시험 질의들 $q_1, q_2, \cdots, q_k$이 주어질 때, P는 수식 $P = \frac{1}{k}\sum_{j=1}^{k} P_j$에 의해 계산된다.

그림 12와 13은 D=3과 D=6일 때 선택된 데이타베이스의 개수(M)에 대하여 두 복합 추정 방법들을 사용한 데이타베이스 선택의 성능을 보여준다. S 데이타베이스들 중 임의로 M 데이타베이스들을 선택하는 임의(random) 데이타베이스 선택 방법의 성능을 비교를 위한 참조 기준으로 활용하였다. 제안된 방법들은 임의 데이

표 4 *SRA*와 *MRA*에 대한 hyper-sphere와 hyper-square의 부피

		Sphere Volume	Total Rectangle-Vol	Total In Volume	Total In Volume Ratio(%)
SRA	D=3	4.19	4.19	3.53	84.2
	D=6	5.17	5.17	3.49	67.5
MRA	D=3	4.19	5.45	4.08	97.4
	D=6	5.17	7.72	4.26	82.5

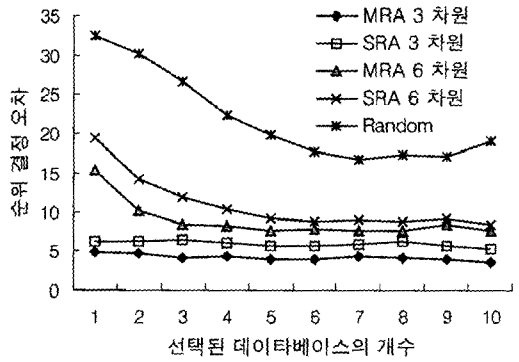

그림 12 순위결정오차

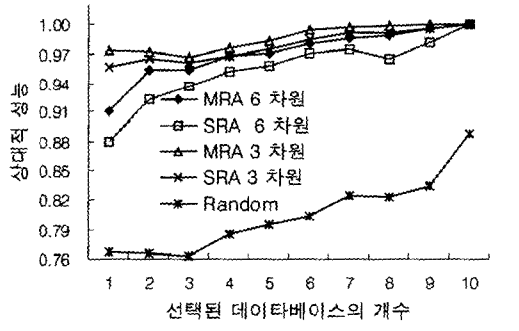

그림 13 상대적 성능

타베이스 선택보다 항상 성능이 좋으며 *MRA*는 *SRA*보다 성능이 좋게 나타났다. 게다가, 두 제안된 방법들과 최적 방법을 비교하여 데이타베이스 선택의 정확성을 검사하였다. 표 4는 D=3, D=6일 때 각각 *SRA*와 *MRA*에 대하여 중심이 (0,…,0)이고 반경이 1인 hyper-sphere의 볼륨과 hyper-rectangle들의 전체 볼륨, 내접 볼륨(=hyper-sphere내에 존재하는 hyper-rectangle들의 볼륨 합)과 내접 볼륨 비율(=내접 볼륨/hyper-sphere의 볼륨)을 보여준다. *MRA*가 *SRA*보다 내접 볼륨 비율이 크기 때문에 *MRA*가 *SRA*보다 성능이 좋음을 알 수 있다. 그러나 이미지 특징들의 차원이 증가함에 따라 *MRA*와 *SRA*의 정확성은 낮아진다. 그 이유는 차원이 증가함에 따라 hyper-rectangle들과 hyper-sphere사이의 내접 볼륨 비율이 감소하기 때문이다.

7. 결 론

본 논문은 웹 상에서의 많은 이미지 데이타베이스에서 적절한 데이타베이스를 선택하는 문제를 탐구하였다. 우리는 이 문제에 대한 해법으로서 새로운 복합 추정 방법을 제안하였다. 이 방법은 비록 이미지 데이타베이스가 다른 유사성 측정 방법을 사용한다고 할지라도 주어진 질의 이미지와 전역적으로 유사한 이미지들의 개수를 비교적 정확히 예측할 수 있다. 이 방법은 구형의 선택을 추정을 위해서 데이타베이스 히스토그램을 사용하며, 메타데이타베이스와 각 이미지 데이타베이스의 선택율의 차이를 보상해주도록 적은 개수의 표본 이미지를 사용한다. 데이타베이스 선택 메카니즘은 메타데이타베이스에서 구현되었으며, 메타데이타베이스는 각 이미지 데이타베이스로부터 필요한 요약 정보를 가져와 저장하는데 그 양은 그리 크지 않다. 이는 복합 추정자에 의해서 사용되는 샘플들의 양과 압축된 히스토그램 정보의 양이 적기 때문이다. 또 다른 잇점으로는 메타데이타베이스가 유지하는 정보가 항상 최신의 정보를 유지할 수 있다는 점이다. 이는 비록 각 이미지 데이타베이스들의 내용이 자주 변한다고 하더라도 *DCT* 계수의 선형성을 사용하면 이미지 데이타 베이스의 가장 최신 요약 정보를 적은 비용으로 갱신할 수 있기 때문이다.

우리는 많은 개수의 실제 이미지 데이타를 사용하여 반복적인 실험을 함으로써 제안된 방법의 효용성(effectiveness)을 입증하였다. 즉 본 논문에서 제시한 방법은 질의에 적합한 데이타베이스를 충분한 정확도를 가지고 선택한다. 앞으로의 연구과제는 이질적인 이미지 데이타베이스의 수집 융합(collection fusion) 문제를 영역 질의(range query)에 대하여 적용하는 것이다.

참 고 문 헌

[1] R. Crane. Simplified approach to Image Processing, Prentice Hall, 1997.

[2] Y. Kanai. Image Segmentation using Intensity and Color Information. *Proceedings of the Visual Communications and Image Processing'98*, Part 2, pages 709-720, January 1998.

[3] J. Callan, Z. Lu, and W. Croft. Searching Distributed Collection with Inference Networks.

Proceedings of the Eighteenth Annual Int'l ACM/SIGIR Conference, pages 21-28, 1995.

[4] L. Gravano, H. Garcia-Molina. Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies. *Proceedings of Int'l Conference on Very Large Data Bases*, August 1995.

[5] W. Meng, K. L. Liu, C. Yu, X. Wang, Y. Chang, N. Rishe. Determining Text Databases to Search in the Internet. *Proceedings of Int'l Conference on Very Large Data Bases*, pages 14-25, August 1998.

[6] W. Meng, K. L. Liu, C. Yu, W. Wu, N. Rishe. Estimating the Usefulness of Search Engines. *Proceedings of Int'l Conference on Data Engineering*, pages 146-153, March 1999.

[7] J. Xu, Y. Cao, E.-P. Lim, W.-K. Ng. Database Selection Techniques for Routing Bibliographic Queries. *Proceedings of ACM Digital Libraries Int'l Conference*, pages 264-273, 1998.

[8] B. Yuwono, D.L. Lee. Server Ranking for Distributed Text Retrieval Systems on the Internet. *Proceedings of Int'l Conference on DB Systems for Advanced Applications*, pages 391-400, 1997.

[9] M. Flickner, H. Sawhney, W. Niblack et al. Query by image and video content: The QBIC system. *IEEE Computer*, Vol.28, No.9, pages 23-32, September 1995.

[10] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B.Horowitz, R.Humphrey, R.Jain and C. Shu. The Virage Image Search Engine: An Open Framework for Image Management. *SPIE Storage and Retrieval for Still Image and Video Databases IV*, pages 76-87, 1996.

[11] J.R. Smith, S.F. Chang. Visually Searching the Web for Content. *IEEE Multimedia*, pages 12-20, Summer 1997.

[12] J.R. Smith, S.F. Chang. VisualSEEk: A Fully Automated Content-Based Image Query System. *Proceedings of the ACM Int'l Multimedia Conference*, pages 87-98, Novemeber 1996.

[13] W. Chang, G. Sheikholeslami, J. Wang, A. Zhang. Data Resource Selection in Distributed Visual Information Systems. *IEEE Transactions on Knowledge and Data Engineering*, Vol.10, No.6, pages 926-946, November 1998.

[14] A.B. Benitez, M. Beigi, S.-F. Chang. A Content-Based Image Meta-Search Engine using Relevance Feedback. *IEEE Internet Computing*, 2 (4), pages 56-69, 1998.

[15] J. Callan, M. Connell, A. Du. Automatic Discovery of Language Models for Text Databases. *Proceedings of ACM SIGMOD Int'l Conference on Management of Data*, pages 479-490, 1999.

[16] F. Provost, D. Jensen, T. Oates. Efficient Progressive Sampling. *Proceedings of ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, pages 23-32, 1999.

[17] D.H. Kim, J.H. Lee, S.L. Lee, C.W. Chung. Heterogeneous Multimedia Database Selection on the Web. *Technical Reports, CS/TR-2000-147*, Korea Advanced Institute of Science and Technology. Available: http://cs.kaist.ac.kr/library/tr.

[18] J. H. Lee, D. H. Kim, C. W. Chung. Multi-dimensional Selectivity Estimation Using Compressed Histogram Information. *Proceedings of ACM SIGMOD Int'l Conference on Management of Data*, pages 205-214, June 1999.

[19] S. Berchtold, C. Bohm, D.A. Keim, H.-P. Kriegel. A Cost Model for Nearest Neighbor Search in High-Dimensional Data Space. *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 78-86, June 1997.

[20] M.H. Kalos, P.A. WhitRock. Monte Carlo Methods. Wiley, NewYork.

[21] A. Natev, R. Rastogi, K. Shim. WALRUS: A Similarity Retrieval Algorithm for Image Databases. Proceedings of ACM SIGMOD Int'l Conference on Management of Data, pages 395-406, June 1999.

김 덕 환

1987년 서울대학교 계산통계학과 학사 1995년 한국과학기술원 정보및통신공학과 석사. 2003년 한국과학기술원 정보및통신공학과 컴퓨터공학전공 박사. 1987년~1997년 2월 LG전자㈜ 통신기기연구소 선임연구원. 1997년 3월~현재 동양공업전문대학 인터넷정보과 조교수. 관심분야는 멀티미디어 데이타베이스, 데이타마이닝, 웹정보검색

이 석 룡

1984년 연세대학교 기계공학과 학사 1993년 연세대학교 산업공학과 전자계산전공 석사. 1984년 1월~1995년 2월 한국IBM 소프트웨어 연구소 선임연구원 1995년~2001년 안산1대학 조교수. 2001년 한국과학기술원(KAIST) 정보및통신공학과 박사. 2002년~현재 한국외국어대학교 산업정보시스템공학부 부교수. 관심분야는 데이타베이스, 데이타 웨어하우스 및 데이타 마이닝, 멀티미디어 검색, 웹 정보 검색

정 진 완

정보과학회논문지 : 데이타베이스
제 30 권 제 1 호 참조