분산 멀티미디어 데이타베이스에 대한 수집 융합 알고리즘

(Collection Fusion Algorithm in Distributed Multimedia Databases)

김 덕 환 † 이 주 홍 †† 이 석 룡 ††† 정 진 완 ††††

(Deok-Hwan Kim) (Ju-Hong Lee) (Seok-Lyong Lee) (Chin-Wan Chung)

요 약 웹에서의 멀티미디어 데이타베이스가 발달함에 따라 분산 멀티미디어 데이타에 대한 검색 기능의 필요성이 높아지고 있다. 그러나 지금까지는 주로 웹상에 분산된 텍스트 데이타베이스를 선택하고 선택된 텍스트 데이타베이스에 대해서 질의 결과를 결합하는 연구가 이루어졌을 뿐 멀티미디어 데이타베이스에 대해서는 연구가 미진하였다. 웹상의 멀티미디어 데이타베이스는 자율적이고 이질적인 특성을 가지고 있고 주로 내용 기반으로 검색된다. 멀티미디어 데이타베이스에서의 수집 융합 문제는 웹상의 이질적인 멀티미디어 데이타베이스에서 내용 기반 검색으로 검색된 결과를 병합하는 것을 다룬다. 이 문제는 분산 멀티미디어 데이타베이스의 검색에 매우 중요하지만 아직까지 연구된 바가 없다.

본 논문은 웹상에서 이질적인 멀티미디어 데이타베이스의 수집 융합을 처리하는 새로운 알고리즘을 제안한다. 본 논문은 데이타베이스에서 검색할 객체의 개수를 추정하는 휴리스틱 방법과 선형 회귀 분석을 이용한 알고리즘을 사용한다. 그리고 실험에 의해서 이 알고리즘들의 효율성을 보였다. 이 알고리즘들은 향후 웹상의 멀티미디어 데이타베이스들에 대한 분산 내용 기반 검색 알고리즘들의 기본이 될 수 있다.

Abstract With the advances in multimedia databases on the World Wide Web, it becomes more important to provide users with the search capability of distributed multimedia data. While there have been many studies about the database selection and the collection fusion for text databases distributed in the Web, a few researches have been attempted for the case of multimedia databases. The multimedia databases on the Web have autonomous and heterogeneous properties and they use mainly the content based retrieval. The collection fusion problem of multimedia databases is concerned with the merging of results retrieved by content based retrieval from heterogeneous multimedia databases on the Web. This problem is crucial for the search in distributed multimedia databases, however, it has not been studied yet.

This paper provides novel algorithms for processing the collection fusion of heterogeneous multimedia databases on the Web. We propose two heuristic algorithms for estimating the number of objects to be retrieved from local databases and an algorithm using the linear regression. Extensive experiments show the effectiveness and efficiency of these algorithms. These algorithms can provide the basis for the distributed content based retrieval algorithms for multimedia databases on the Web.

· "이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음" (KRF-2000-041-E00262)

† 정 회 원 : 한국과학기술원 정보및통신공학과
dhkim@islab.kaist.ac.kr
†† 종신회원 : 한국과학기술원 정보및통신공학과
jhlee@islab.kaist.ac.kr
††† 비 회 원 : 한국과학기술원 정보및통신공학과
sllee@islab.kaist.ac.kr
†††† 종신회원 : 한국과학기술원 전산학과 교수
chungcw@islab.kaist.ac.kr

논문접수 : 2001년 1월 27일
심사완료 : 2001년 6월 25일

1. 서 론

최근 웹이 비약적으로 발전함에 따라서 인터넷에 분산되어 있는 멀티미디어 데이타베이스들을 검색하는 것이 의료진단, 디지털도서관, 원격 학습, 분산 출판, 전자상거래 등의 다양한 응용 영역에서 중요한 연구 주제가 되고 있다.

웹상에 연결된 멀티미디어 데이타베이스 서버의 개수는 날이 갈수록 증가하고 있으며 사용자가 많은 데이타

베이스에서 데이타를 검색하려면 이들을 통합하여 질의 처리를 하여 주는 메타 서버가 필요하다[1, 2]. 메타 서버는 여러 개의 지역 데이타베이스 들에 각각 질의를 주어 사용자 요구에 맞도록 질의 결과들을 통합하여 주며, 하나의 데이타베이스가 있는 것과 같이 보이도록 뷰를 제공한다.

과거에 멀티미디어 데이타는 주석 방법에 의해 색인되고 검색이 이루어졌으나 이 방법은 주석을 다는 사람의 주관에 따라 쉽게 달라지며 많은 양의 작업을 필요로 하고 융통성을 결여하여 멀티미디어 데이타의 전체 내용에 대한 검색에는 종종 실패하고 있다. 따라서 전체 내용에 대한 검색에는 내용 기반의 검색이 주로 사용된다. 이미지 데이타를 예로 들면 색상(color), 형태(shape), 질감(texture)등과 같은 속성 들에 대해서 다차원 벡터로 표현되는 특징을 추출하여 내용 기반 검색에 사용한다.

일반적으로 웹상에서 운용되는 데이타베이스들은 독립적으로 만들어지거나 운영되고 있으므로 서로 이질적인 특성을 가지고 있다. 예를 들면, QBIC[3], MARS [4], Virage[5] 등과 같은 멀티미디어 데이타베이스들은 내용 기반의 검색을 하기 위해서 각각 다른 특징 추출 방법, 다른 유사성 측정 함수와 상이한 색인 방법들을 지원하고 있다.

메타 서버가 다루어야 할 중요한 문제는 서로 다른 유사성 측정 함수(similarity measure)를 갖는 다수의 멀티미디어 데이타베이스들로부터 질의에 적합한 객체들을 어떻게 구하는 가이다. 이를 위하여 메타 서버는 대체로 다음과 같은 기능을 가진다. 첫째, 모든 데이타베이스로부터 질의에 대하여 유사한 데이타의 개수를 대략적으로 추정할 수 있는 메타 데이타를 수집하여 저장하고 질의가 들어오면 유사한 데이타를 많이 찾을 수 있는 데이타베이스를 선택한다(**데이타 베이스 선택 문제**). 둘째, 선택된 각각의 데이타베이스들로부터 질의 객체와 가장 유사한 데이타를 가져온다(**수집 융합 문제**).

기존의 텍스트 데이타베이스에 대해서는 위의 2 가지 연구 주제들이 많이 연구되었다[6, 7, 8, 9]. 그러나 멀티미디어 데이타베이스에 대해서는 Chang 등[1]이 데이타베이스 선택 문제를 연구했지만 수집 융합 문제에 대해서는 깊이 있게 연구된 바가 없다. 게다가 Chang등은 다중 속성들에 대한 질의를 다루었지만 서로 다른 지역 데이타베이스들이 같은 속성에 대하여 같은 유사성 척도를 갖는 것으로 가정하였으므로 동질적인 환경에서의 데이타베이스 선택 문제로 제한된다. 이질적인 분산 멀티미디어 환경에서 수집 융합 문제를 다룬 논문은 아직 없다.

이질적인 환경에서의 수집 융합 문제는 사용자가 요구하는 유사성 측정 함수와 지역 데이타베이스에서 제공하는 유사성 측정 함수가 서로 일치 하지 않기 때문에 발생한다. 상세한 예는 다음과 같다. 사용자는 메타 서버에서 지원하는 유사성 측정 함수를 사용하여 지역 데이타베이스들로부터 질의 개체와 유사한 데이타를 검색하고자 한다. 그러나 지역 데이타베이스는 메타 서버의 유사성 측정 함수를 지원하지 않으며, 대신에 자신의 유사성 측정 함수를 사용한다. 메타 서버의 유사성 측정 함수가 지역 데이타베이스의 유사성 측정 함수와 완전히 다른 경우, 예를 들면, 메타 서버는 색상에 대한 유사성 측정 함수를 사용하고 지역 데이타베이스는 질감에 대한 유사성 측정 함수를 사용할 때, 사용자는 질의에 적합한 결과를 얻을 수 없다. 따라서, 메타 서버의 유사성 측정 함수는 지역 데이타베이스의 유사성 측정 함수와 상관 관계를 가져야 한다.

본 논문에서는 두 유사성 측정 함수들간에 선형 관계가 성립하는 사례들이 있음을 보이고, 이질적인 멀티미디어 데이타베이스들이 웹에 분산되어 있는 환경에서 그와 같은 경우에 수집 융합 문제를 해결하기 위한 새로운 분산 유사성 검색 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 간략히 소개한다. 3장에서는 수집 융합 문제를 정의하고 그 문제를 풀기 위한 가정과 관찰을 제시한다. 4장에서는 수집 융합 문제에 대한 알고리즘들을 설명한다. 5장에서는 알고리즘들의 실험 결과들을 보인다. 6장에서는 결론과 향후 연구 과제를 제시한다.

2. 관련연구

수집 융합 문제를 풀기 위하여 다양한 방법들이 시도되었다. 그러나 대부분의 연구가 텍스트 데이타베이스에 관한 것이며 멀티미디어 데이타베이스에 대한 연구는 없었다. 따라서 텍스트 데이타베이스에 관한 수집 융합 문제의 연구들을 다루고 웹상의 멀티미디어 데이타베이스에 대한 기존의 연구들을 살펴보겠다.

Voorhees[9] 등은 각 데이타베이스로부터 적합한 객체들의 분포를 추정하기 위하여 연습 질의들(training queries)을 사용하는 방법을 제안하였다. 질의가 주어지면, 시스템은 가장 유사한 연습 질의들을 찾고 이 연습 질의들을 사용하여 얻어진 정보를 이용하여 각 데이타베이스로부터 얼마나 많은 문서들을 가져와야 하는 지 결정한다. 그러나 이 방법은 새로운 데이타베이스들이 자주 추가되는 환경에 적합하지 않으며 너무 많은 연습 질의들을 갖는 것은 시스템에 부담이 될 수 있다.

Gravano와 Garcia-Molina[7]는 각 지역 데이타베이스 들로부터 질의에 적합한 객체들을 모두 구하는 것을 보장하는 해석적 방법을 최초로 제시하였다. 각 지역 데이타베이스 들이 근본적으로 다른 순위(ranking) 알고리즘을 사용하기 때문에 메타 서버에서 지역 데이타베이스들의 점수(score)로 결과를 합칠 수 없으므로, 각 지역 데이타베이스는 질의 결과로서 텍스트 문서의 공통 속성 값을 보낸다. 그런 다음 메타 서버는 공통 속성을 이용하여 최종 점수를 계산한 후 그 결과들을 합친다. 수집 융합을 하기 위하여 사용자에 의하여 주어진 질의 유사도 경계값(global threshold)은 그에 대응하는 지역 유사도 경계값(local threshold)으로 변환되어야 한다. 그러나 [7]에 의해 제안된 알고리즘에 의하여 변환된 지역 유사도 경계값은 최적의 경계값보다 낮게 계산될 수 있다. 결과적으로 많은 수의 부적합한 문서들이 검색될 수 있다.

Meng등[8]은 이러한 단점을 개선하기 위해서 지역 데이타베이스를 위한 최적의 경계값을 얻는 방법을 연구했다. 그들은 선형 프로그래밍(linear programming)과 Lagrange's multipliers를 이용하여 질의 유사도 경계값이 주어질 때 최적의 지역 유사도 경계값을 구하는 방법을 제안하였다.

멀티미디어 자원들이 웹 상에 증가함에 따라서, 멀티미디어 데이타베이스들을 위한 데이타베이스 선택 및 수집 융합 문제에 대한 연구의 필요성이 점차 증가하고 있으나 이 분야에 대한 연구는 아직까지 미미한 실정이다. Chang등[1]은 메타 데이타베이스의 이미지 내용기반 색인에 기반을 둔 시각적 데이타베이스의 선택문제를 연구하고 NetView[10]라는 시스템 아키텍처를 제안하였다. NetView 시스템 아키텍처는 원격지의 시각적 데이타베이스들, 메타 서버, 클라이언트에서의 웹 응용 프로그램 등의 3 가지 주요 부분을 포함한다. 메타 서버는 사용자 질의를 받고 데이타베이스들에 대한 순위를 매기어 데이타베이스 들을 선택하고 선택된 지역 데이타베이스 들에 질의를 보낸다.

그들은 지역 데이타베이스내의 이미지 군집들을 대표하는 이미지인 템플릿(template) 들에 대한 질의의 시각적 유사성과 템플릿들과 관련된 이미지 군집들의 통계적 데이타들을 이용한 평균(mean) 기반 방법과 히스토그램 기반 방법을 제시하고 있다. 평균 기반 방법은 시각 질의에 적합한 이미지 군집들을 결정하기 위하여 템플릿에 대한 데이타베이스 이미지들의 유사성 분포의 평균, 분산 그리고 표본의 개수를 이용한다. 히스토그램 기반 방법은 히스토그램으로 표현되는 유사성 분포의 통계적 특성과 이미지 클러스터 내에 있는 데이타베이스 이미지들의 위치를 이용한다. 이 방법은 지역 데이타베이스와 메타 서버가 같은 특징 추출 방법과 같은 유사성 측정 함수를 사용하는 동질의 분산 환경을 가정하고 있다. 그러나 실제로 웹에 존재하는 대부분의 데이타베이스들은 자치적이고 이질적이므로 실제 상황에서는 제한적으로 사용될 수 밖에 없다.

3. 분산된 멀티미디어 데이타 베이스들에 대한 수집 융합

메타 서버는 분산되어 있는 지역 데이타베이스 들로부터 질의 처리에 필요한 메타 데이타를 수집하여 보관하고 데이타베이스 선택, 질의 변환, 수집 융합 등의 모든 처리를 수행한다. 본 논문에서는 데이타베이스 선택, 질의 변환에 대해서는 다루지 않고 수집 융합만을 다루므로 메타 서버라는 용어 대신에 질의 서버라는 용어를 사용한다.

3.1 수집 융합 문제에 대한 형식적 정의

이질적인 분산 멀티미디어 데이타베이스가 있는 웹 환경에서, k-최근접 유사성 질의의 목적은 질의 서버가 자신의 유사성 측정 함수에 의한 가장 유사한 객체들을 지역 데이타베이스들로부터 검색하는 것이다. 그러나 지역 데이타베이스들은 질의 서버의 유사성 측정 함수에 의해서 검색하지 않고 자신의 유사성 측정 함수에 따라 검색하기 때문에 질의 서버입장에서는 가장 적합한 답을 찾지 못할 수도 있다. 따라서 질의 서버가 질의에 적합한 답을 가져오면서 비적합한 답을 최소한으로 가져오는 방법에 대한 연구가 필요하다. 수집 융합 문제의 형식적인 정의와 목적은 다음과 같다.

정의1. (수집 융합 문제) 멀티미디어 데이타베이스에서의 수집 융합 문제는 질의 서버의 유사성 측정 함수에 의하여 질의에 적합한 객체들을, 즉 k개의 최근접한 객체들을, 찾기 위하여 분산된 이질적인 멀티미디어 데이타베이스들을 검색하고 결과를 가져온다. ◆

수집 융합 문제의 목표. 질의 Q가 k-최근접 유사성 질의라고 할 때, Q에 대해서 i번째 지역 데이타베이스의 적합한 답의 집합을 R_Q^i라고 하고 비적합한 답의 집합을 I_Q^i라고 하자. 그러면 $R_Q^i \cap I_Q^i = \emptyset$ 이고 $R_Q^i \cup I_Q^i =$ {i번째 지역 데이타베이스의 모든 객체}이다. i번째 지역 데이타베이스로부터 가져온 객체들의 집합을 W_Q^i라고 하자. 그러면 지역 데이타베이스들에서 검색한 객체의 전체 개수 ($\sum_{i=1}^{n} |W_Q^i| = ck$, c는 1보다 큰 상수이고 n은 지역 데이타베이스의 개수)는 일정하다는 제약조건 하에

서 수집 융합 문제의 목적은 다음과 같다:

(1) 적합한 객체들에서 가져온 객체의 비율을 최대화한다. 즉 $\sum_{i=1}^{n}|W_Q^i| = ck$라는 조건하에서 $\sum_{i=1}^{n}|R_Q^i \cap W_Q^i| / \sum_{i=1}^{n}|R_Q^i|$를 최대화한다.

(2) 가져온 객체 중에서 비적합한 객체의 비율을 최소화한다. 즉 $\sum_{i=1}^{n}|W_Q^i| = ck$라는 조건하에서 $\sum_{i=1}^{n}|I_Q^i \cap W_Q^i| / \sum_{i=1}^{n}|W_Q^i|$를 최소화한다. ◆

(1)은 리콜을 최대화하는 것이고 (2)는 정확도가 $1 - \sum_{i=1}^{n}|I_Q^i \cap W_Q^i| / \sum_{i=1}^{n}|W_Q^i|$로 주어지므로 정확도를 최대화하는 표현으로 바꿀 수 있다. 실제 상황에서는 (1)과 (2)는 서로 상충되는 목표이다. 따라서 이 둘의 상호 보완적 관계(Tradeoff)를 조절하는 문제가 하나의 이슈가 될 수 있다.

3.2 분산 멀티미디어 데이타베이스 환경에서 중요한 가정들

이 절에서는 질의 서버와 지역 데이타베이스들에 대하여 몇 가지 타당한 가정을 하고 있다. 본 논문에서 제시하는 알고리즘들은 이 가정들을 기반으로 개발되었다.

가정 1. 질의 서버는 자신의 유사성 측정함수와 상관관계가 있는 유사성 측정 함수를 지원하는 지역 데이타베이스들을 선택하여 그들에게 질의를 보낸다.

가정 2. 지역 데이타베이스들은 [11]에 언급된 give-me-more 기능을 사용한 방법처럼 점진적 유사성 순위(incremental similarity ranking)와 영역 질의(range query) 기능을 지원하며 [12]와 같은 방법을 사용하여 고차원 선택률 추정 방법을 제공한다.

가정 3. 지역 데이타베이스에 질의를 주었을 때 지역 유사성 측정 함수에 의해 구해진 질의 결과로서 질의에 유사한 이미지들과 선택률 정보, 지역적 유사도 값 등을 보낸다.

3.3 유사도 측정 함수들 사이의 관계

웹상에서 운영되는 서버들은 독립적으로 만들어지고 운영되기 때문에, 유사성 질의에 사용되는 속성들이 같을 지라도 유사성 측정 함수들이 서로 다를 수 있다. 따라서 지역 유사성 측정 함수에 의해 구해진 지역 데이타베이스의 객체와 질의 객체사이의 유사도 값은 질의 서버의 유사성 측정함수에 의해 구해진 동일한 객체들 사이의 유사도 값이 다를 수 있다. 내용기반 이미지 검색을 위한 많은 유사성 측정 함수들이 있으며, 이들 중 일부의 유사성 측정 함수들간에 상관관계가 있다. 그와 같은 사실을 설명하기 위하여 다음 예제들을 제시한다.

예제 1. 질의 서버는 전체 이미지를 5×5로 분할하여 각 부분 이미지들의 평균 색상을 RGB 색상공간에서 추출하여 특징으로 사용하며, MARS[4]에서 언급한 특징간 정규화 방법을 사용하여 질의 이미지에 대한 유사도 값(y축)을 구한다. 지역 데이타베이스는 전체 이미지를 2×2으로 분할하여 각 부분 이미지들의 평균 색상을 RGB 색상공간에서 추출하여 특징으로 사용하며, 질의 서버와 같은 방법으로 유사도 값(x축)을 구한다. 그림 1은 임의의 질의 이미지에 대하여 3,016개의 이미지들로부터 구한 유사도 값들의 산포도를 보여준다. 산포도는 직선 형태로 나타난다. ◆

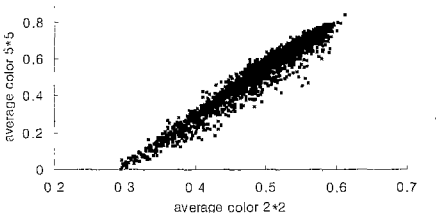

그림 1 평균 색상 2×2와 평균 색상 5×5일 때 산포도

예제 2. 질의 서버는 전체 이미지를 6×6으로 분할하여 각 부분 이미지들의 질감을 특징으로 사용하며 MARS에서 언급한 특징간 정규화 방법을 사용하여 질의 이미지에 대한 유사도 값(y축)을 구한다. 지역 데이타베이스는 전체 이미지를 4×4로 분할하여 각부분 이미지들의 질감을 특징으로 사용하고 질의 서버와 같은 방법으로 유사도 값(x축)을 구한다. 그림 2는 임의의 질의 이미지에 대하여 3,016개의 이미지들로부터 구한 유사도 값들의 산포도를 보여준다. 산포도는 직선의 형태로 나타난다. ◆

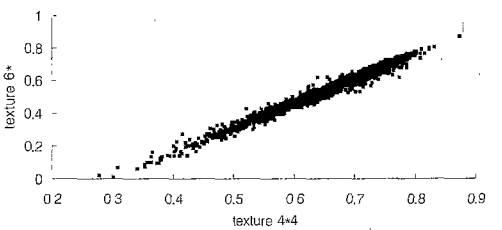

그림 2 질감 4×4와 질감 6×6일 때 산포도

예제 3. 그림 3에서는 y축에는 평균 색상을 속성으로 사용하고 x축에는 질감을 속성으로 사용하는 경우에 유사도 값들의 산포도를 보여준다. y축의 유사도 값들을 구하기 위해 5×5으로 분할 된 영역의 평균 색상을 RGB 색상 공간에서 추출하여 특징으로 사용하고 x축의

유사도 값들을 구하기 위해 6×6으로 분할 된 영역의 질감을 특징으로 사용한다. 앞의 경우와 달리 서로 다른 속성을 갖는 두 유사도 값들간에는 상관관계가 전혀 없음을 관찰할 수 있다. ◆

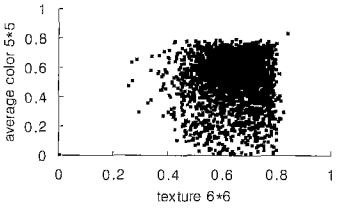

그림 3 평균 색상 5×5와 질감 6×6일 때 산포도

즉, 질의 서버와 지역 데이타베이스 들간에 유사성 측정함수가 다르더라도 일부의 유사성 측정함수들간의 유사도 값들의 산포도는 직선형태로 나타남을 관찰했다.

그러나 같은 속성을 가진 두 개의 다른 유사성 측정함수들이 선형 관계를 보임을 증명할 수 없다. 대신에, 선형 관계를 갖는 많은 사례들을 보여주는 상세한 실험을 수행했다. 표 1은 유사성 측정 함수들을 위하여 사용되어지는 특징들의 3 그룹, 즉 RGB 공간에서의 평균 색상, RGB 공간에서의 질감, RGB 공간에서의 색상 및 질감을 보여준다. 유사도 값을 계산하기 위하여 MARS [4]에서 기술한 특징간 정규화 방법을 사용한다. 통계적 선형 회귀 방법이 직선의 방정식을 얻기 위하여 사용되며, 통계적 가설 검정이 두 유사성 측정 함수간의 선형 관계 여부를 판단하기 위하여 사용된다. 검정 지표로서 산포도, 표본결정계수(r^2), 분산 분석(F_0, $F(\alpha)$)을 사용한다. r^2는 직선 회귀에 의하여 설명되는 변동/총변동(sum of squares due to linear regression/total variation)에 의해 주어지며, F_0는 직선 회귀로 인한 제곱합/오차 제곱합(mean square due to linear regression/mean square of residual)으로 주어지며, $F(\alpha)$는 유의수준이 α일 때 F-분포로부터 얻어진다. 선형 회귀 모델이 두 유사성 측정함수에 대해 유효하다면, $r^2(0< r^2 <1)$는 1에 근접해야 하며 F_0는 $F(\alpha)$보다 커야 한다[13].

표 1 유사도 측정 함수들을 위해 사용되는 특징들의 세 그룹

특징 이름		특징에 대한 설명
색상특징들	feat1	RGB색상공간에서 2×2 부분 이미지들을 위한 평균색상
	feat2	RGB색상공간에서 3×3 부분 이미지들을 위한 평균색상
	feat3	RGB색상공간에서 4×4 부분 이미지들을 위한 평균색상
	feat4	RGB색상공간에서 5×5 부분 이미지들을 위한 평균색상
	feat5	RGB색상공간에서 6×6 부분 이미지들을 위한 평균색상
질감특징들	feat6	RGB색상공간에서 2×2 부분 이미지들을 위한 평균질감
	feat7	RGB색상공간에서 3×3 부분 이미지들을 위한 평균질감
	feat8	RGB색상공간에서 4×4 부분 이미지들을 위한 평균질감
	feat9	RGB색상공간에서 5×5 부분 이미지들을 위한 평균질감
	feat10	RGB색상공간에서 6×6 부분 이미지들을 위한 평균질감
색상/질감특징들	feat11	RGB색상공간에서 2×2 부분 이미지들을 위한 평균색상과 질감
	feat12	RGB색상공간에서 3×3 부분 이미지들을 위한 평균색상과 질감
	feat13	RGB색상공간에서 4×4 부분 이미지들을 위한 평균색상과 질감
	feat14	RGB색상공간에서 5×5 부분 이미지들을 위한 평균색상과 질감

표 2는 이미지 쌍들로부터 유사도 값들을 측정할 때 두 유사성 측정 함수들에 대한 실험 결과를 보여준다. 두 특징들이 같은 그룹에서 선택된 경우와 다른 그룹에서 선택된 경우로 2 가지 부류를 만들었다. 유사성 검색은 주어진 질의 객체에 대하여 실행되므로, 고정된 질의 객체간와의 유사도 값들과 임의의 이미지들과의 유사도 값들을 측정하는 실험을 하였다. 표 2에서 보여진 것과 같이, 같은 그룹에 속한 유사도 측정함수를 사용한 경우 산포도는 직선을 나타내며, r^2는 1에 근접하고 F_0는 $F(\alpha)$보다 매우 크다.

그러나, 유사성 측정 함수들이 서로 다른 그룹에 속한 경우 산포도는 직선의 형태를 보이지 않으며 r^2는 0에 근사한다. F_0가 $F(\alpha)$보다 크게 나타나더라도 선형관계

표 2 두 유사도 측정 함수사이의 직선 회귀에 대한 통계적 가설 검정

특징들	산포도	상관계수 ρ	r^2	F_0	F (0.05)	결과
feat1:feat2	straight line	0.985	0.970	96566	0.000	linear
feat1:feat4	straight line	0.964	0.923	36327	0.000	linear
feat1:feat9	scattered	0.071	0.005	15.16	0.000	non linear
feat6:feat8	straight line	0.960	0.922	35839	0.000	linear
feat3:feat5	straight line	0.997	0.994	490367	0.000	linear
feat6:feat12	scattered	0.132	0.017	53.37	0.000	non linear
feat8:feat10	straight line	0.993	0.985	202061	0.000	linear
feat5:feat10	scattered	0.056	0.003	9.564	0.002	non linear
feat1:feat10	scattered	0.051	0.003	7.945	0.005	non linear
feat11:feat14	straight line	0.968	0.937	44451	0.000	linear
feat7:feat9	straight line	0.984	0.969	93264	0.000	linear
feat12:feat13	straight line	0.996	0.992	377866	0.000	linear

가 만족되는 경우의 F_0보다 훨씬 작은 값을 나타낸다.

두 유사성 측정함수로 구해진 유사도 값들이 선형 관계를 만족한다면, 웹상의 분산 유사성 검색을 위해 이 성질을 사용할 수 있다.

4. 분산 유사성 검색 알고리즘

4.1 분산 k-최근접 검색을 하기 위한 휴리스틱

앞서 정의한 수집 융합 문제의 목적을 달성하려면 리콜과 정확도를 동시에 높여야 한다. 이를 달성하기 위한 방법으로서 질의 Q에 대한 i번째 지역 데이타베이스의 질의 유사도와 검색 효율을 다음과 같이 정의한다.

정의 2. n개의 지역 데이타베이스가 있다고 하자. k-최근접 유사성 질의에 대해서 i번째 지역 데이타베이스 LD_i에서 질의 Q에 대한 질의 유사도(*Query similarity ratio*, S_Q^i)와 검색 효율(*Retrieval efficiency*, RE_Q^i)은 S_Q^i=(LD_i에서 질의 Q에 적합한 객체의 수)/k와 RE_Q^i=(LD_i에서 질의 Q에 대한 적합한 객체의 수)/(LD_i에서 질의 Q에 대한 모든 적합한 객체를 얻기 위하여 검색해야 하는 객체의 수)로 정의된다. ◆

질의 유사도가 높은 지역 데이타베이스에 적합한 객체가 많이 있다. 만약 검색 효율이 높다면 비적절한 객체를 적게 가져오면서 많은 적합한 객체를 가져올 수 있을 것이다. 실제로는 이 값들을 정확히 알 수가 없기 때문에 휴리스틱을 사용하여 그 값을 대략적으로 구한다. 다음 알고리즘은 질의 유사도와 검색 효율을 감안하여 지역 데이타베이스로부터 객체들을 가져오는 휴리스틱 알고리즘이다. 표 3은 휴리스틱 알고리즘과 선형 회귀 분석을 이용한 분산 유사성 검색 알고리즘들에 사용되어진 기호들을 나타낸다.

표 3 분산 유사성 검색 알고리즘에 사용된 기호들

기호	의미
Q	분산 유사성 질의를 위한 질의 객체
k	검색할 데이타의 개수
c	k보다 더 많이 검색하는 비율 (>1)
n	지역 데이타베이스의 개수
a	한 개의 지역 데이타베이스에 대한 평균 검색 수
LD_i	i번째 지역 데이타베이스
P_i	한 스텝에서 LD_i로 부터 검색되는 객체의 개수
r	한 단계에서 검색되는 객체의 개수(미리 정해지는 값임)
y	질의 유사도 축
x	지역 유사도 축
d_y	y축에서 100(1-δ)% 신뢰구간 길이의 반
x'	d_y를 추정하기 위해서 사용되는 지역 유사도 값
GT	영역 검색의 질의 유사도 경계값

4.1.1 휴리스틱 알고리즘

알고리즘 *휴리스틱-검색*(Q, C, k, n, LD_1,..., LD_n)

(1) 모든 지역 데이타베이스 LD에 질의 객체 Q를 보낸다.

(2) While(검색된 객체들의 개수 < ck)

(3) 각각의 LD_i에 대하여 get_more_objects(Q, P_i, LD_i)를 실행한다. $result_i$는 LD_i로부터 검색된 객체들의 집합이다.

(4) merge_results($result_1$, ..., $result_n$)를 실행한다.

(5) 각각의 LD_i에 대하여 질의 유사도와 검색 효율을 추정한다.

(6) 각각의 LD_i에 대하여 질의 유사도와 검색 효율을 사용하여 P_i를 재계산한다.

(10) EndWhile

merge_results($result_1$, ..., $result_n$)는 질의 서버의 유사성 측정 함수를 이용하여 모든 데이타베이스로부터 가져온 검색 결과를 합하고 순위(*rank*)를 정한다. *get_more_objects*(Q, P_i, LD_i)는 [11]에 기술한 바대로 LD_i의 지역 유사성 측정 함수를 이용하여 LD_i로부터 질의 Q와 유사한 객체를 P_i개 더 가져온다.

지역 데이타베이스에서 가져온 객체들에는 비적절한 객체들이 있기 때문에 질의 서버가 k개의 객체만을 가져오면 리콜이 1보다 작을 수 있다. 따라서 리콜이 높아지려면 질의 서버가 k개 보다 더 많이 가져와야 한다. 즉 ck개수의 객체를 가져온다. 여기서 c는 1보다 큰 상수이다. 그러나 c를 너무 큰 값으로 하면 정확도가 떨어진다. 즉 리콜과 정확도는 서로 상호 보완적(tradeoff) 관계가 있으므로 c의 값을 적절히 선택해야 하는 문제가 있다.

만약 모든 지역 데이타베이스의 질의 유사도와 검색 효율이 같다면, 초기 P_i는 $P_i=\left[\frac{ck}{n}\right]$로 주어지며 지역 데이타베이스들에서 각각 한번만 가져오면 된다. 여기서 [ ]는 반올림 함수이다. 그러나 실제로 지역 데이타베이스들의 질의 유사도와 검색 효율은 서로 다르며 그 값을 미리 알 수도 없으므로 반복 실행하면서 그 값을 점차 정확히 한다. 반복 횟수가 a라면 초기의 P_i의 값은 $P_i=\left[\frac{ck}{an}\right]$로 주어지고 알고리즘의 단계(6)에서 지역 데이타베이스의 질의 유사도가 높을수록 리콜을 증가시키기 위해서 P_i에 높은 값을 주고 검색 효율이 높을수록 정확도를 증가시키기 위해서 P_i에 높은 값을 준다. P_i값은 다음 절에서 설명하는 휴리스틱에 의해서 정해진다.

4.1.2 평균 순위 휴리스틱

질의 유사도는 리콜과 관련되어 있으며 검색 효율은 정확도와 관련되어 있다. 지역 데이타베이스에서 가져온 정보에서 정확한 값을 추정하기에는 정보양이 부족하므로 질의 유사도와 검색 효율을 각각 따로 구하는 것이 어렵다. 따라서 이 두 값의 결합된 측정치로서 휴리스틱 α_i를 제안한다. 질의 서버는 α_i값이 큰 지역 데이타 베이스에서는 객체들을 많이 가져오고 α_i값이 작은 지역 데이타베이스에서는 객체들을 적게 가져온다. α_i는 다음과 같이 정의된다: $\alpha_i = M_i \Big/ \sum_{j=1}^{M_i} Rank_{ij}$ α_i는 i번째 지역 데이타베이스에서 가져온 객체들의 통합 순위의 평균의 역이다. $Rank_{ij}$는 i번째 지역 데이타베이스로부터 가져온 j번째 객체의 통합 순위이다. M_i은 i번째 지역 데이타베이스로부터 마지막에 가져온 데이타의 개수이다. 휴리스틱 알고리즘의 P_i는 다음과 같이 주어진다. $P_i = \left[\frac{k}{\alpha} \cdot \frac{\alpha_i}{\alpha_1 + \ldots + \alpha_n}\right]$, 여기서 n은 지역 데이타베이스의 개수이다.

4.1.3 평균 질의 유사성 휴리스틱

이 방법은 평균 순위 휴리스틱과 비슷하다. 순위는 인접한 객체와의 사이의 순위차이가 균일하다. 그러나 유사성의 차이는 균일하지 않을 수 있으므로 휴리스틱 β_i를 다음과 같이 정의한다:

$\beta_i = \frac{\sum_{j=1}^{M_i} Global\_Similarity_{ij}}{M_i}$, β_i는 i번째 지역 데이타베이스에서 검색된 객체들의 평균 질의 유사도이고 M_i는 앞의 경우와 같다. 휴리스틱 알고리즘의 P_i는 다음과 같이 주어진다.

$$P_i = \left[\frac{k}{\alpha} \cdot \frac{\beta_i}{\beta_1 + \ldots + \beta_n}\right].$$

4.2 선형 회귀 분석을 이용한 영역 검색 알고리즘

앞에서 설명한 휴리스틱은 모든 지역 데이타베이스가 비교적 비슷한 비율로 적합한 객체가 있는 경우에 유용한 방법이다. 그러나 지역 데이타베이스간의 적합한 객체의 분포의 편차가 큰 경우에는 좀더 해석적인 방법이 필요할 것이다. 3.3절에 보인 대로 같은 속성을 사용하지만 유사도 측정 함수가 다른 유사도 간에 선형관계가 존재하는 경우 선형 회귀 분석이라는 확률적인 모형에 근거한 해석적인 알고리즘을 고안하였다. 먼저 분산 k-최근접 유사성 질의를 다루기 전에 이 알고리즘의 근간이 되는 선형 회귀 분석에 대해 간략히 설명하고 선형 회귀를 이용한 영역 검색 알고리즘을 다룬다.

4.2.1 선형 회귀 분석

선형 회귀 분석은 두 확률 변수의 관계를 직선에 의해서 추정하는 통계적인 방법이다. 두 변수 간의 관련성을 알려면 먼저 xy평면상에 산포도를 그려 보아야 한다. 산포도가 직선(straight line)의 형태를 보이면 두 변수 간의 함수 관계를 $y = \alpha + \beta x$로서 나타낼 수 있다. 최소 제곱법(least square method)에 의해서 추정된 직선의 식을 $\hat{y} = \hat{\alpha} + \hat{\beta} x$ 이라고 한다면 각각의 값을 다음과 같이 구할 수 있다: $\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$, $\hat{\beta} = \frac{S_{xy}}{S_x^2}$ 이다. 또한 x^*에서의 y의 $100(1-\delta)\%$ 신뢰 구간(confidence interval)은 다음과 같이 주어진다. $\hat{\alpha} + \hat{\beta} x^* \pm t_{\delta/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2}}$, 여기서 $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$이고 $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ 이며, $S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$이다. t는 t분포를 나타내고 s는 $s^2 = \frac{S_y^2 - S_{xy}^2/S_x^2}{n-2}$ 이고, n은 회귀 분석[14]에 사용된 데이타의 개수이다.

4.2.2 영역 검색 알고리즘

영역 검색 알고리즘의 목적은 지역 데이타베이스 LD로부터 질의서버의 유사도(global similarity)가 $GT \in [0,1]$보다 큰 거의 모든 객체들을 검색하는 것이다. 제안된 영역 검색 알고리즘은 선형 회귀 분석 방법을 사용한다. 지역 데이타베이스는 자신의 유사성 측정 함수(지역 유사도)를 사용하여 질의 객체와 유사한 객체들을 검색하므로, 즉 질의 객체에 대해서 주어진 지역 유사도 경계값 보다 큰 지역 유사도(local similarity)를 가진 객체들을 검색하므로, 질의 서버의 유사도의 경계값 GT에 대응되는 지역 유사도의 경계값들을 계산하여야 한다.

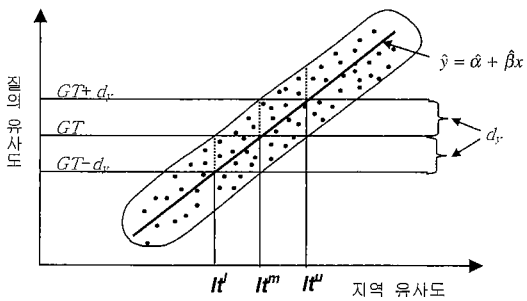

그림 4 질의 유사도 경계값에 대한 세가지 지역 유사도 경계값

그림 4에서 보는 바와 같이 질의 서버의 유사도에 대응되는 지역 유사도에는 세가지(lt^l, lt^m, lt^u)가 있다. 지역 유사도 경계값 lt^m는 $\hat{y} = \hat{\alpha} + \hat{\beta}x$과 $y = GT$의 교점의 x좌표이다. lt^l는 $\hat{y} = \hat{\alpha} + \hat{\beta}x$과 $y = GT - d_y$의 교점의 x좌표이다. lt^u는 $\hat{y} = \hat{\alpha} + \hat{\beta}x$과 $y = GT + d_y$의 교점의 x좌표이다. 지역 유사도의 경계값이 lt^l인 경우에, 영역 검색 알고리즘은 적합한 객체들을 거의 모두 검색하지만 비

적합한 객체들도 함께 많이 포함되는 오버헤드가 있다. 이것은 결과의 리콜은 높으나 정확도가 낮아진다는 것을 의미한다. lt^m의 경우에는 리콜은 lt^l의 경우보다 작으나 정확도는 더 높아진다. lt^u의 경우에는 리콜은 세가지 경우 중에서 가장 낮으나 정확도는 가장 높다. 선형 회귀 분석을 사용한 영역 검색 알고리즘은 다음과 같다:

알고리즘 *영역-유사성-검색* (*Q, r, , GT, LD, T*)

(1) $p \leftarrow r$, $total\# \leftarrow 0$

(2) while ($p > 0$)

(3) get_more_objects(Q, p, LD)를 실행한다.

(4) 선형 회귀 분석을 사용하여 $\hat{y} = \hat{\alpha} + \hat{\beta}x$를 구한다.

(5) $d_y = t_{\delta/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2}}$

(6) if $T=lower$ then $lt = lt^l = \frac{GT - d_y - \hat{\alpha}}{\hat{\beta}}$

(7) if $T=middle$ then $lt = lt^m = \frac{GT - \hat{\alpha}}{\hat{\beta}}$

(8) if $T=upper$ then $lt = lt^u = \frac{GT + d_y - \hat{\alpha}}{\hat{\beta}}$

if (Expected_num(lt,LD) $-$ $total\#$) $\geq r$

(10) then $p = r$

(10) else p=Expected_num(lt,LD) $-$ $total\#$

(11) $total\# \leftarrow total\# + p$

(12) end while

Expected_num(lt,LD)은 지역 유사도 값이 lt보다 큰 객체들의 개수의 기대값으로서 지역 데이타베이스 LD의 선택율 추정 방법을 사용하여 추정된다. T는 지역 유사도 경계값의 타입으로서 *upper*, *middle*, *lower*는 각각 lt^u, lt^m, lt^l를 나타낸다.

이 알고리즘은 초기에 각 지역 데이타베이스로부터 r개의 객체들을 가져와서 선형 회귀 분석 방법으로 그것들을 분석함으로써 직선의 식을 얻는다. 또한 알고리즘은 y의 $100(1-\delta)\%$ 신뢰구간의 길이의 반인 d_y를 추정하고 그림 4에서 보인 대로 지역 유사도 경계값 lt를 구한다. 지역 데이타베이스 안에서 지역 유사도 값이 lt보다 큰 객체의 전체 개수, 즉 Expected_num(lt,LD), 를 추정한다. Expected_num(lt,LD)에서 검색한 객체의 전체 개수 $total\#$을 빼면 더 검색해야 할 객체의 개수가 나온다. 만약 더 검색해야 할 개수가 r보다 크다면 r개만을 더 검색하지만 r보다 작은 경우에는 더 가져와야 할 개수 만큼만 가져온다. 추정된 직선과 lt가 불충분한 정보로부터 추정되기 때문에 정확하지 않을 수 있다. 따라서 알고리즘은 미리 정해진 개수인 r개 만큼만 반복적으로 더 가져오고 직선의 식과 LT의 값을 반복하여 계산하면 점차적으로 정확한 값에 근접하게 될 것이다. r은 미리 정해진 값인데, r이 작으면 알고리즘의 반복 회수가 커지므로 오버헤드가 커지지만 결과는 약간 더 정확할 수도 있다. r이 크다면 오버헤드는 감소하지만 결과의 정확도가 약간 작아질 수 있다.

위의 영역 검색 알고리즘은 4.3절에 제시된 분산 k-최근접 알고리즘의 근간을 이루고 있으며 선형 회귀 분석을 유사성 검색에 적용하는 과정을 나타내기 위해 기술되었다. 성능 실험을 하여 이를 검증하였으나 지면 관계로 실험 부분을 생략하였다.

4.3 선형 회귀 분석을 사용한 분산 k-최근접 검색 알고리즘

분산 k-최근접 검색 알고리즘은 n개의 지역 데이타베이스로 부터 질의 Q에 대해서 질의 서버의 유사성 측정 함수를 사용하여 유사도가 가장 높은 k개의 객체들을 가져오는 알고리즘이다. 이 알고리즘은 검색된 객체들의 지역 유사도 값들의 최소값에 대응되는 세 가지 종류의 질의 서버의 유사도 경계값, gt^u, gt^m, gt^l들을 사용한다. 최소 지역 유사도 값은 지역 유사도 경계값(LT)을 사용하며 gt^m은 $\hat{y} = \hat{\alpha} + \hat{\beta}x$과 $x=LT$의 교점의 y값이다. gt^u는 $\hat{y} = \hat{\alpha} + \hat{\beta}x + d_y$과 $x=LT$의 교점의 y좌표이다. gt^l는 $\hat{y} = \hat{\alpha} + \hat{\beta}x - d_y$과 $x=LT$의 교점의 y좌표이다. T는 질의 서버의 유사도 경계값의 타입으로 사용되며 *upper*, *middle*, *lower*는 각각 gt^u, gt^m, gt^l를 나타낸다.

알고리즘 *K-최근접-검색*(*p, c, k, n, Q, T, LD*$_1$,..,*LD*$_n$)

(1) 각각의 LD_i에 대하여 get_more_objects(Q, p, LD_i)를 실행한다.

(2) 각각의 LD_i에 대하여 선형 회귀 분석을 이용하여 LD_i로부터 구한 객체들을 분석하고 직선 $\hat{y}_i = \hat{\alpha}_i + \hat{\beta}x_i$을 추정하며, T에 따라서 gt_i^u, gt_i^m, gt_i^l중의 하나인 gt_i를 구한다.

(3) LD_l을 모든 지역 데이타베이스들 중에 가장 큰 gt를 갖는 데이타베이스라고 하고 가장 큰 gt를 gt_l이라 한다.

(4) if (검색된 객체들 중 유사도가 gt_l보다 큰 객체들의 개수) $\geq k$ 또는 (검색된 객체들의 전체 개수) $\geq ck$ then 수행을 종료한다.

(5) 모든 데이타베이스들 중에서 가장 큰 gt를 갖는 LD_l을 선택한다.

(6) get_more_objects(Q, p, LD_l)를 수행한다.

(7) 선형 회귀를 이용하여 LD_l로부터 구한 객체들을 분석한다.

(8) 단계 (3)으로 간다.

위 알고리즘의 gt는 세 가지의 질의 서버의 유사도 경계값 gt^m, gt^u, gt^l들 중의 하나이다. 세 가지 질의 서버의 유사도 경계값에 대한 잇점과 불리한 점은 영역 질의 알고리즘의 경우와 비슷하다. gt^u의 경우에는 결과의 리콜이 높으나 정확도는 낮다. gt^m의 경우에는 리콜은 gt^u의 경우 보다 작으나 정확도는 더 높다. gt^l의 경우에는 리콜은 세 가지 경우 중에서 제일 작으나 정확도는 제일 높다.

5. 실 험

제안된 분산 유사성 검색 알고리즘의 효율성과 성능을 측정하기 위하여 많은 수의 이미지 데이타들과 다양한 질의들을 포함하는 환경에서 포괄적인 실험을 하였다. 시험 데이터는 3,016개의 256 *RGB* 색상 비트맵 이미지들로 구성된다. 표 4는 시험에 사용된 이미지들의 내용을 보여준다. 평균 순위 휴리스틱 알고리즘, 평균 질의 유사성 휴리스틱 알고리즘 그리고 선형 회귀를 이용한 분산 k-최근접 검색 알고리즘들은 n개의 지역 데이타베이스 들로부터 질의 서버의 유사도가 가장 높은 k개의 객체들을 구한다. 부분적으로 검색된 결과에 대한 선형 회귀의 정확성을 보이기 위하여, 표 5에 실험 결과들을 제시한다. 부분 결과들이 최종 결과에 점진적으로 근접하며 검색된 객체들의 수가 증가할수록 r^2, α, 그리고 β가 최종 값에 근접함을 알 수 있다.

표 4 시험 이미지들의 내용

범주	이미지의 개수	영역
Plants	720	flower, leaves, grass
Pattern	680	glass, brick, woods
Architecture	820	house, building
Scene	796	water, sky, cloud

표 5 부분적으로 검색된 객체들을 이용한 선형회귀의 정확성

[1]MSE(mean square error)=(residual sum of squares) /(number of retrieved objects).

검색된 객체의 개수	MSE[1]	r^2	α	β
70	$7.71*10^{-5}$	0.7069	-0.071	1.086
110	$8.82*10^{-5}$	0.7060	0.002	0.992
190	$7.16*10^{-5}$	0.8053	0.045	0.936
230	$6.78*10^{-5}$	0.8485	0.025	0.971
299	$6.62*10^{-5}$	0.8832	0.019	0.962
전체 객체들	$5.12*10^{-5}$	0.9836	0.013	0.959

정확도와 리콜 그리고 이들의 곱으로 표현되는 복합 측정 기준을 이용하여 알고리즘의 효율성을 평가한다. 복합 측정 기준으로 전체적인 효율성을 측정할 수 있다. 다양한 매개변수를 사용하여 각 시험마다 10개의 질의를 하고 그 결과들을 평균하였다.

한 개의 질의 서버와 네 개의 지역 데이타베이스를 두어 이미지들이 지역 데이타베이스에 분포되어 있다고 가정한다. 이미지들을 각 지역 데이타베이스에 할당하기 위하여 (1) 균일 할당, (2) 군집 할당의 2가지 방식을 사용한다. 균일 할당에서는 모든 이미지들이 4개의 데이타베이스들에 균일하게 분배되어진다. 군집 할당에서는 유사한 이미지들이 같은 지역 데이타베이스에 할당될 확률이 높으며 비균일하게 분배된다. 이와 같은 두 가지 경우에 대하여 각각 실험을 하였다. 시험에 사용된 다른 매개 변수 값들로서 y의 신뢰 구간을 예측하기 위해 사용된 신뢰수준을 99.9%로 하였으며, 각 단계에서 가져오는 이미지의 개수를 10으로 정하고, 질의 서버의 유사도 경계값으로 gt^u를 사용하였다. 분산 k-최근접 질의에 대한 실험은 제안된 3알고리즘(평균 순위 휴리스틱 알고리즘, 평균 질의 유사성 휴리스틱 알고리즘과 선형 회귀를 이용한 알고리즘)의 비교를 포함한다.

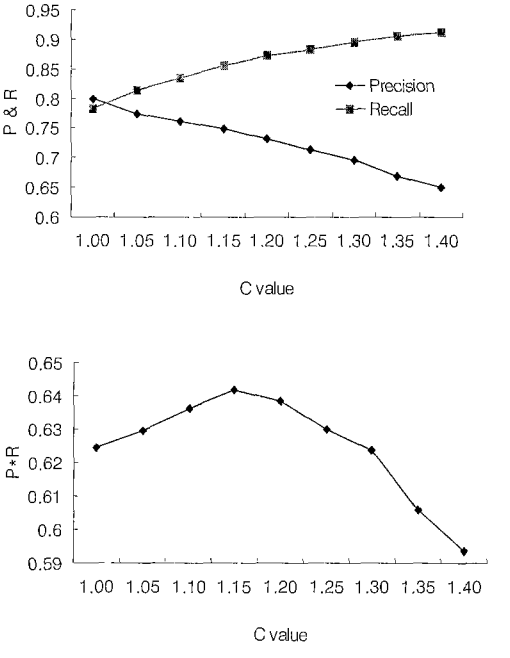

그림 5 균일 분포를 갖는 지역 데이타베이스들에서의 c 값에 따른 정확도의 변화

먼저 지역 데이타베이스들로부터 구해지는 객체의 개수에 관하여 선형 회귀를 이용한 k-최근접 알고리즘을 사용하여 c값의 영향을 관찰하였다. c=1.0은 지역 데이타베이스들로부터 단지 k개의 객체들만을 구하는 것을 의미하며, c=1.40은 1.40*k개의 객체들이 구해지는 것을 의미한다. 이 값을 1.0에서 1.40으로 변경할 때, 그림 5와 그림 6과 같이 각각 정확도와 리콜이 변화됨을 알 수 있다. 균일 할당과 군집 할당의 두 가지 경우에 대하여 c=1.15 근처에서 $P\times R$ 값이 가장 높게 나타난다. 정확도는 C값이 커짐에 따라 감소하고 리콜은 그 반대 방향으로 움직임을 관찰할 수 있다.

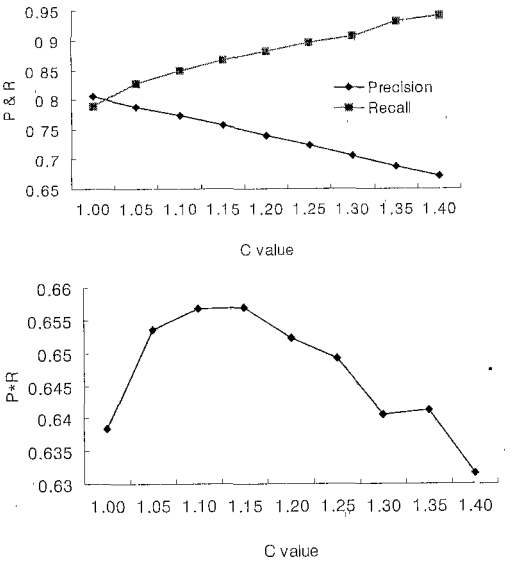

그림 6 군집 분포를 갖는 지역 데이타베이스들에서의 c 값에 따른 정확도의 변화

세가지 k-최근접 질의 알고리즘들을 위한 정확도와 리콜 그리고 복합 측정 기준인 $P\times R$에 대한 그래프가 그림 7에서 그림 9에 걸쳐 보여진다. 군집 분포의 경우 선형 회귀를 이용한 알고리즘(linear)은 데이타 분포의 군집 효과를 잘 반영하기 때문에 이 알고리즘이 평균 순위 휴리스틱 알고리즘(alpha)과 평균 질의 유사성 휴리스틱 알고리즘(beta)보다 좋은 성능을 보인다. 균일 분포의 경우, 선형 회귀를 이용한 알고리즘이 다른 알고리즘들보다 약간 좋게 나타난다. 그러나 실제 상황에서는, 웹에 있는 데이타베이스들의 데이타 분포가 일반적으로 군집되어 있다. 따라서, 선형 회귀를 이용한 알고리즘이 더욱 실제적으로 사용될 것이다.

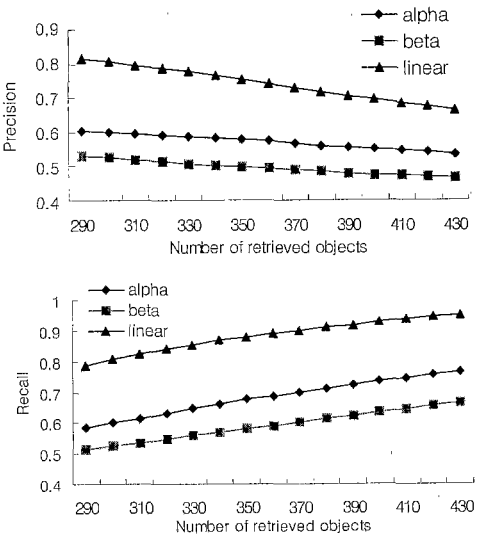

그림 7 군집 분포에서 각 알고리즘의 정확도와 리콜

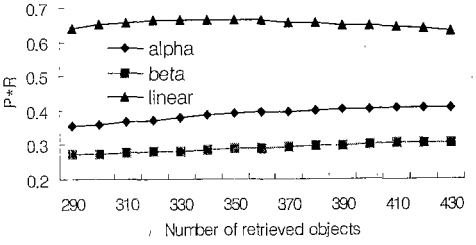

그림 8 군집 분포에서 각 알고리즘의 $P\times R$

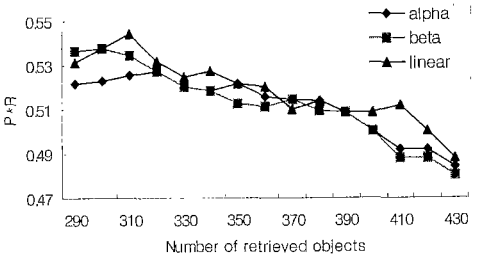

그림 9 균일 분포에서 각 알고리즘의 $P\times R$

6. 결 론

이 논문은 웹상의 멀티미디어 데이타베이스를 위한 수집 융합문제를 해결하는 새로운 알고리즘을 제안한다. 이 알고리즘들이 작동하는 전제 조건으로 웹상의 멀티미디어 데이타베이스는 자치적이며 이질적이라는 실질

적인 가정을 하였다. 이러한 환경과 가정하에서 일부 유사성 측정 함수들에 대하여 서로 다른 유사성 측정 함수를 사용하는 유사도간에 선형 관계가 있다는 중요한 관찰로부터 두 유사도에 대한 직선의 관계식을 선형 회귀 분석 기법으로 구할 수 있음을 보였다. 본 논문에서는 질의 서버의 k-최근접 유사성 질의 처리를 위해 두 가지의 휴리스틱과 선형 회귀분석법을 사용하는 확률적 알고리즘을 제안한다. 질의 서버의 유사성 측정 함수와 지역 데이타베이스의 유사성 측정 함수가 상관 관계를 가지는 경우에, 질의 서버의 k최근접 유사성 질의 처리는 본 논문에서 제시된 알고리즘을 사용하여 지역 데이타베이스들로부터 검색한 결과들을 통합하여 순위를 정할 수 있다. 실험에 의하면 선형 회귀 분석법을 사용한 알고리즘이 가장 성능이 좋은 것으로 나타났다.

이 논문은 웹상의 이질적인 멀티미디어 데이터베이스들에 대한 k-최근접 질의로서의 수집 융합 문제를 최초로 연구하였으며 이 분야의 알고리즘을 제시하였다. 앞으로 웹상에서의 멀티미디어 데이타베이스 검색은 점점 더 중요성이 커지고 있다. 따라서 이 논문에서 제시된 알고리즘은 앞으로의 연구에 중요한 기반을 제공할 것이다.

향후 연구 과제로서 고려되어야 할 이슈는 다음과 같다. 본 논문에서는 객체들이 지역 데이타베이스들로부터 질의 서버에 직접 전달되어야 한다고 가정하였다. 그러나 만약 멀티미디어 객체의 크기가 비디오 클립과 같이 매우 크다면 이 방법은 네트웍 부담이 크게 될 것이다. 따라서 이러한 경우에는 질의 서버가 지역 데이타베이스에게 에이젼트 프로그램을 보내고 에이젼트 프로그램이 지역 데이타베이스들로부터 검색에 필요한 유용한 정보를 수집하여 질의 서버에 전달하고 이 정보를 사용하여 효율적으로 수집 융합을 하는 알고리즘을 고려할 수 있다. 이러한 경우의 알고리즘은 이 논문에서 제시한 알고리즘과 상당한 차이가 있을 것이다.

참 고 문 헌

[1] 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.

[2] L. Gravano, Y. Papakonstantinou. Mediating and Metaserarching on the Internet. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol.21 No.2, pages 28-36, June 1998.

[3] 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.

[4] M. Ortega, K. Chakrababarti, K. Porkaew, S. Mehrotra. Supporting Ranked Boolean Similarity Queries in MARS. *IEEE Transactions on Knowledge and Data Engineering*, Vol.10, No.6, pp.905 925, November/December 1998.

[5] 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.

[6] 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.

[7] L. Gravano, H. Garcia-Molina. Merging Ranks from Heterogeneous Internet Sources. *Proceedings of Twenty-third Int'l Conference on Very Large Data Bases*, pages 14-25, August 1997.

[8] 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.

[9] E. Voorhees, N. Gupta, and B. Johnson-Laird. The Collection Fusion Problem. *Proceedings of Third Text Retrieval Conference(TREC-3)*, pages 95-104, 1994.

[10] A. Zhang, W. Chang, G. Sheikholeslami, and T. Syeda-Mahmood. NetView: Integrating Large-Scale Distributed Visual Databases. *IEEE Multimedia*, pages 47-59, September 1998.

[11] T. Seidl, H. Kriegel. Optimal Multi-Step k-Nearest Neighbor Search. *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data*, pages 154-165, June 1998.

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

[13] R. V. Hogg, E. A. Tanis. Probability & Statistical Inference. MacMillan Publishing Co., 1977.

[14] F. Hillier, G. Lieberman. Introduction to Operations Research. McGraw-Hill, pp.755-762, 1977.

김 덕 환

1987년 서울대학교 계산통계학과(학사). 1995년 한국과학기술원 정보및통신공학과(석사). 1995년 ~ 현재 한국과학기술원 박사과정. 1987년 ~ 1997년 2월 LG 전자(주) 통신기기연구소 선임연구원. 1997년 3월 ~ 현재 동양공업전문대학 인터넷정보과 조교수. 관심분야는 멀티미디어 데이타베이스, 데이타마이닝, 웹정보검색

이 주 홍

1983년 서울대학교 컴퓨터공학과 졸업(학사). 1985년 서울대학교 컴퓨터공학과 졸업(석사). 1985년 ~ 1989년 한국통신공사 사업지원단 전임연구원. 1989년 ~ 1996년 한국아이비엠 소프트웨어연구소 선임프로그래머. 2001년 한국과학기술원 정보및통신공학과 박사. 관심분야는 멀티미디어 데이타베이스, 데이타 웨어하우스, 데이타마이닝, 질의 최적화

이 석 룡

정보과학회논문지 : 데이타베이스
제 28 권 제 2 호 참조

정 진 완

정보과학회논문지 : 데이타베이스
제 28 권 제 2 호 참조