논문 2011-48CI-2-3

배아 데이터의 효율적 검색을 위한 계층적 구조화 방법

(Hierarchical Organization of Embryo Data for Supporting Efficient Search)

원 정 임*, 오 현 교**, 장 민 희**, 김 상 욱***

(Jung-Im Won, Hyun-Kyo Oh, Min-Hee Jang, and Sang-Wook Kim)

요 약

배아란 동물이나 식물과 같은 다세포 생물의 초기 단계를 의미한다. 배아의 단계에서 다세포 생물의 기초적인 체제가 결정되기 때문에 배아는 개체발생의 기구를 연구하는 중요한 연구대상이 된다. 생물학자들은 배아 연구를 위해 대용량의 배아 이미지 데이터를 소유하고 있으며, 이러한 대용량 데이터 중 원하는 이미지를 효율적으로 검색하기 위해서는 데이터 구조화가 필요하다. 데이터베이스 구조화를 위해 주로 사용되는 방법으로 계층적 클러스터링이 있다. 그러나 기존의 계층적 클러스터링 방법은 데이터베이스를 트리 형태로 구조화 하는 과정에서 클러스터의 크기와 클러스터 내의 객체 수를 동시에 고려하지 못하기 때문에 결과 클러스터링 트리가 경사 트리일 가능성이 매우 높다. 경사 트리인 경우 사용자가 원하는 이미지를 검색하기 위해 트리를 순회할 때 많은 시간이 걸린다. 따라서 본 논문에서는 대용량의 배아 이미지 데이터를 경사 되지 않으며 균형 상태에 가까운 트리 형태로 구조화하기 위한 방안을 제시한다. 제안하는 방안은 데이터베이스 내에 저장된 배아 이미지를 그래프로 변환하고 반복적으로 그래프 분할 알고리즘을 적용하여 클러스터를 생성한다. 이 때 클러스터의 크기와 클러스터 내의 객체 수를 동시에 고려하여 특정 클러스터의 크기가 지나치게 커지거나 객체 수가 많아지는 것을 방지한다. 실험을 통해서 제안하는 방안의 우수성을 규명하고 시각화 툴을 제공하여 사용자가 원하는 배아 이미지를 쉽게 찾을 수 있도록 돕는다.

Abstract

Embryo is a very early stage of the development of multicellular organism such as animals and plants. It is an important research target for studying ontogeny because the fundamental body system of multicellular organism is determined during an embryo state. Researchers in the developmental biology have a large volume of embryo image databases for studying embryos and they frequently search for an embryo image efficiently from those databases. Thus, it is crucial to organize databases for their efficient search. Hierarchical clustering methods have been widely used for database organization. However, most of previous algorithms tend to produce a highly skewed tree as a result of clustering because they do not simultaneously consider both the size of a cluster and the number of objects within the cluster. The skewed tree requires much time to be traversed in users' search process. In this paper, we propose a method that effectively organizes a large volume of embryo image data in a balanced tree structure. We first represent embryo image data as a similarity-based graph. Next, we identify clusters by performing a graph partitioning algorithm repeatedly. We check constantly the size of a cluster and the number of objects, and partition clusters whose size is too large or whose number of objects is too high, which prevents clusters from growing too large or having too many objects. We show the superiority of the proposed method by extensive experiments. Moreover, we implement the visualization tool to help users quickly and easily navigate the embryo image database.

Keywords : 데이터베이스 구조화, 계층적 클러스터링, 대표 객체, 유사도 측정 함수, 배아 데이터

* 정회원, 한양대학교 전기정보통신 기술연구소
(Research Institute of Electrical and Computer Engineering, Hanyang University)
** 학생회원, *** 평생회원(교신저자), 한양대학교 전자컴퓨터통신학과
(Department of Electronics and Computer Engineering, Hanyang University)
접수일자: 2011년2월14일, 수정완료일: 2011년3월7일

Ⅰ. 서 론

배아(embryo)란 동물이나 식물과 같은 다세포 생물이 한번 이상 세포 분열을 하기 시작한 시기부터 하나의 완전한 개체가 되기 전까지의 과정 중 발생 초기 단계를 의미한다. 배아 단계에서 다세포 생물의 기초적인 체제가 결정되기 때문에 배아는 개체발생의 기구를 연구하는 중요한 연구 대상이 된다[1]. 이러한 배아를 연구하는 학문을 발생 생물학(developmental biology)이라고 한다[2]. 발생 생물학에서 배아의 다양한 특징을 분석하기 위해서는 대용량의 배아 이미지 데이터베이스가 요구된다. 생물학자들은 이 데이터베이스를 통해 자신이 원하는 배아 이미지를 찾아 특징을 분석하고 연구를 진행 할 수 있다.

대용량 데이터베이스에서 사용자가 자신이 찾고자 하는 이미지를 정확히 알고 있느냐에 따라 검색 방법을 질의 기반 검색(query-based search)과 브라우징 기반 검색(browsing-based search) 방법으로 나눌 수 있다. 질의 기반 검색 방법은 사용자가 자신이 원하는 이미지의 텍스트 태그 등의 메타 데이터 또는 이미지 자체의 특징을 구체적으로 가지고 있을 때 쓰이는 방법으로 찾고자 하는 이미지의 특징을 질의로 이용하여 유사한 이미지를 데이터베이스로부터 검색 하는 것이다. 반면에 브라우징 기반 검색 방법은 사용자가 자신이 원하는 이미지를 정확히 모르고 있을 때 사용하는 방법으로 사용자가 직접 데이터베이스 내에 존재하는 이미지들을 살펴보면서 자신이 원하는 이미지를 찾는 방법이다.

배아 이미지의 특징을 구체적으로 기술하는 것이 쉽지 않고 생물학자들이 자신이 찾고자 하는 배아 이미지의 특징을 정확히 모르는 경우가 많기 때문에 질의 기반 이미지 검색을 수행하는 것은 어려운 일이다. 따라서 생물학자는 원하는 배아 이미지를 검색하기 위하여 브라우징 기반 검색 방법을 이용하여 데이터베이스 내에 저장되어 있는 모든 배아 이미지 데이터를 직접 살펴봐야 한다. 그러나 데이터베이스 내에는 매우 많은 수의 배아 이미지들이 저장되어 있으므로 이러한 배아 이미지들을 일일이 살펴보는 것은 현실적으로 매우 어렵다. 이러한 문제점을 해결하기 위한 방법으로 데이터베이스 구조화 방법을 사용할 수 있다.

데이터베이스 구조화란 데이터베이스 내에 있는 이미지들을 유사한 혹은 연관되어 있는 이미지들끼리 나누어 놓는 것으로서, 이를 통해 데이터베이스 사용자의 브라우징 기반 검색 범위를 제한하여 사용자가 원하는 것과 유사한 이미지들만 집중적으로 살펴 볼 수 있도록 한다[3]. 본 논문에서는 대용량의 배아 이미지 데이터베이스를 효과적으로 구조화하기 위한 방법에 대하여 논의한다.

데이터베이스 구조화를 위해 주로 사용되는 방법으로 계층적 클러스터링(hierarchical clustering)이 있다. 계층적 클러스터링은 객체간의 유사도를 계산하여 객체들을 유사한 특징을 가진 클러스터들로 구분하고 이를 트리 구조 형태로 표현하는 방법이다[3]. 데이터베이스 구조화 방법으로 계층적 클러스터링을 사용하면 사용자가 전체 데이터베이스 구조를 직관적으로 파악할 수 있으며 브라우징 기반 검색을 사용하여 사용자가 원하는 배아 이미지를 빠르게 검색할 수 있다. 기존의 계층적 클러스터링 알고리즘으로는 CHAMELEON[4], BIRCH[5], CURE[6], ROCK[7] 등이 있다.

이러한 기존의 계층적 클러스터링 방법들은 병합적인 계층적 클러스터링(agglomerative hierarchical clustering)을 사용한다[3]. 병합적인 계층적 클러스터링 방법은 시작 단계에서 데이터베이스 내의 모든 이미지들을 유사도에 따라 원자 클러스터들로 분리한 후, 정해진 기준에 따라 유사한 두 개의 클러스터간의 병합을 수행한다. 이 과정을 모든 이미지들이 하나의 클러스터로 병합되거나 미리 정의된 클러스터의 수와 같은 특정 종료 조건이 만족될 때까지 반복한다. 이러한 방법을 상향식(bottom-up) 전략의 계층적 클러스터링이라고 하는데, 이 방법은 클러스터의 유사도 같은 기준에 따라 단계적으로 합쳐지기 때문에 특정 클러스터로만 계속 병합될 수 있다. 따라서 이 결과, 계층적 클러스터가 경사 트리(skewed tree) 형태로 구축될 가능성이 매우 높다. 경사 트리인 경우, 사용자가 직접 찾아봐야 하는 이미지가 트리 높이에 비례하여 많아지기 때문에 브라우징 검색 시 소모되는 시간이 매우 클 수밖에 없다. 따라서 본 논문에서는 대용량의 배아 이미지 데이터들을 경사(skew) 되지 않으며 균형(balance) 상태에 가까운 트리 형태로 구조화하기 위한 새로운 계층적 클러스터링 방법을 제안한다.

제안하는 계층적 클러스터링 기법은 다음과 같은 순서로 진행된다.

- 1단계: 각 이미지의 색상 특성과 유사도를 고려한 유사도 그래프 생성
- 2단계: 생성된 그래프에 그래프 분할 알고리즘을

적용하여 클러스터들을 구성
- 3단계: 구성된 각 클러스터들을 대표할 수 있는 대표 객체 선정
- 4단계: 2단계와 3단계를 반복하여 트리 구조 형성

본 논문에서는 배아 이미지의 RGB 히스토그램을 추출하여 이를 이미지간의 유사도를 판단하기 위한 특성으로 이용한다. 그러나 이러한 히스토그램은 고차원 데이터이기 때문에 고차원 데이터를 그대로 이용하여 클러스터링을 수행한다면 차원의 저주 문제가 발생한다[8-10]. 이러한 문제를 해결하기 위하여 제안하는 방법은 1단계에서 각 이미지를 노드로 하고 이미지 간의 유사도를 간선으로 하는 유사도 그래프를 생성한다. 유사도 그래프 생성 시 근접한 k개 이미지간의 유사도만을 이용한다. 2단계에서는 생성된 그래프에 그래프 분할 알고리즘을 적용하여 클러스터들을 생성한다. 이때 고려해야 될 것은 클러스터의 크기와 클러스터 내에 포함되는 이미지 수이다. 클러스터의 크기가 너무 크다면 유사하지 않은 이미지들이 한 클러스터에 속해 있을 수 있다. 그리고 한 클러스터의 이미지 수가 너무 많다면 결국 브라우징 기반 검색 방법을 통해 사용자가 확인해야 하는 이미지 수가 많아 질 수밖에 없게 된다. 따라서 본 논문에서는 특정 클러스터의 크기가 지나치게 커지거나 클러스터 내의 이미지의 수가 많아지는 것을 방지하는 방법을 제안한다. 3단계에서는 각 클러스터의 대표 객체를 선정한다. 대표 객체(representative object)란 클러스터 내의 모든 배아 이미지들 중에서 클러스터의 특성을 가장 잘 반영할 수 있는 배아 이미지를 의미한다. 마지막으로 4단계에서는 2단계와 3단계를 반복하여 각 클러스터들을 재분할함으로써 데이터베이스를 트리 구조로 구조화 한다. 기존의 계층적 클러스터링 방법은 병합적인 계층적 클러스터링을 수행하였기 때문에 결과로 얻는 트리 구조가 경사되는 문제가 발생하였으나 본 논문에서 제안하는 방법은 분할 방법을 통해 계층적 클러스터링을 수행하기 때문에 균형 상태에 가까운 트리 구조를 결과로 얻을 수 있다.

본 논문의 공헌을 요약하면 다음과 같다.

- 제안하는 계층적 클러스터링은 클러스터의 크기, 클러스터 내에 포함된 이미지의 수를 고려하여 클러스터링 하기 때문에 이 방법을 통해 대용량의 배아 이미지 데이터를 경사 되지 않으며 균형 트리에 가까운 형태로 구조화 한다.
- 사용자는 제안하는 방안의 결과인 트리 구조를 이용한 브라우징 기반 검색을 통해 손쉽게 원하는 배아 이미지를 얻을 수 있다. 또한 트리 구조를 눈으로 쉽게 볼 수 있도록 돕는 시각화 툴을 구현하였다. 이 툴은 사용자가 클러스터들의 대표 객체들을 눈으로 직접 확인 하면서 자신이 원하는 객체를 쉽고 빠르게 찾을 수 있도록 도움을 준다.
- 다양한 실험을 통해 제안하는 방안을 이용하여 이미지를 구조화 한 결과가 기존의 계층적 클러스터링 방법의 결과보다 훨씬 더 효율적인 검색을 하는 것으로 나타났다. 또한, 제안하는 방안의 클러스터링의 결과가 정확도 측면에서도 우수한 것으로 나타났다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 제안하는 계층적 클러스터링 방안을 다룬다. 제 Ⅲ장에서는 제안하는 방안을 통한 검색이 얼마나 효율적인지를 보이고 이 방법을 통해 도출된 클러스터링 결과에 대한 정확도를 보인다. 끝으로, 제 Ⅳ장에서는 결론을 제시한다.

Ⅱ. 제안하는 방안

본 장에서는 대용량 배아 이미지 데이터의 효율적 검색을 지원하기 위한 데이터베이스 구조화 방법인 계층적 클러스터링을 제안한다. 제 Ⅱ장 1절에서는 데이터베이스 구조화를 위하여 고려되어야 할 사항들을 설명하고, 제 Ⅱ장 2절에서는 제안하는 데이터베이스 구조를 보인다. 제 Ⅱ장 3절에서는 제안된 구조를 구현하기 위한 계층적 클러스터링 방법을 서술한다. 제 Ⅱ장 4절에서는 검색 효율성을 향상시키기 위한 대표 객체 선정 방식을 설명한다.

2.1. 데이터베이스 구조화를 위한 고려 사항

제안하는 계층적 클러스터링 방법을 이용하여 대용량 배아 이미지 데이터베이스를 구조화하기 위해서는 먼저 다음과 같은 사항들이 고려되어져야 한다.

(1) 클러스터의 크기

일반적으로 클러스터의 크기(diameter)란 클러스터

내의 모든 객체 쌍 간의 거리(distance)들 중에서의 최댓값을 의미한다. 본 연구에서는 배아 이미지 데이터의 클러스터링 위한 척도로 유사도(similarity)를 이용하는데 이 경우 클러스터의 크기는 클러스터 내의 모든 객체 쌍 간의 유사도들 중에서의 최솟값이다.

데이터베이스 구조화 과정에서 클러스터의 크기가 너무 크면 해당 클러스터 내에 서로 유사하지 않은 배아 이미지 데이터가 모여 있을 수 있으므로 사용자가 원하는 이미지와 거리가 먼 후보 이미지들까지도 살펴봐야하는 문제점이 있다. 또한, 이 경우 해당 클러스터 내에서 대표 객체를 선정하는 것이 무의미하다. 반면에, 클러스터의 크기가 너무 작으면 서로 유사한 배아 이미지가 여러 클러스터에 분산되어 저장되어 있을 수 있으므로 사용자가 원하는 배아 이미지를 검색하기 위하여 여러 후보 클러스터를 접근하여 살펴봐야 하는 문제점이 있다.

(2) 클러스터 내에 포함된 데이터의 수

클러스터의 크기가 크지 않더라고 클러스터 안에 포함된 이미지의 개수가 너무 많으면 사용자가 원하는 배아 이미지를 검색하기 위하여 살펴봐야 하는 후보 이미지의 개수가 많아지게 된다. 반면에 클러스터 안에 포함된 이미지의 개수가 너무 적으면 사용자가 원하는 배아 이미지가 해당 클러스터에 없을 가능성이 있기 때문에 여러 후보 클러스터를 접근하여 살펴봐야 하는 문제점이 있다.

(3) 계층 구조의 균형 정도

계층적 클러스터링 방법을 통하여 구성된 데이터베이스의 구조가 경사 구조인 경우에는 일정한 검색 성능을 보장할 수 없다는 문제점이 있다. 계층 구조의 깊이(depth)가 깊어지면 사용자는 최상위 루트 노드부터 브라우징 기반 검색을 통해 원하는 이미지를 찾기 위하여 다수의 후보 클러스터들을 접근해야 하는 문제가 발생하게 된다. 반면에 계층 구조의 깊이가 너무 낮으면 특정 클러스터의 크기가 너무 커지거나 혹은 클러스터 내에 포함된 이미지 데이터의 개수가 많아지는 현상이 발생할 수 있어 사용자가 한 클러스터 내의 매우 많은 후보 이미지 데이터를 살펴봐야 하는 문제점이 있다.

기존 대부분의 계층적 클러스터링 알고리즘은 클러스터의 크기와 클러스터 내에 데이터 개수를 동시에 고려하지 못하므로 클러스터링을 수행하여 얻어진 클러스터마다 크기와 데이터의 개수가 매우 큰 차이를 보일 가능성이 크다. 또한, 두 클러스터 내에 존재하는 데이터 간에 유사성이 낮더라도 두 클러스터가 다른 클러스터들 보다 인접해 있으면 이들 클러스터를 병합하기 때문에 클러스터의 크기가 크고 그 안의 데이터의 개수가 매우 많을 가능성이 있다. 그리고 기존의 클러스터링 기법들은 병합적인 클러스터링 방법을 사용하였기 때문에 데이터베이스 구조가 경사 트리의 형태를 띨 수 있다. 이로 인해 브라우징 기반 검색을 수행하기 위해 접근되는 클러스터의 개수와 이미지 데이터의 개수가 많아지는 결과를 초래하게 된다.

(4) 접근되는 클러스터의 개수와 데이터의 개수

대용량 배아 이미지 데이터베이스를 대상으로 브라우징 기반 검색을 수행하는 사용자는 최소한의 이미지 데이터만을 접근하여 자신이 원하는 이미지를 검색하길 원한다. 이를 지원하기 위해서는 사용자가 최종 원하는 배아 이미지를 검색하기까지 접근하는 클러스터의 개수 혹은 이미지 데이터의 개수를 최소화해야 한다.

계층적 클러스터링 방법을 사용하여 데이터베이스를 구조화 하면 사용자가 원하는 이미지와 거리가 먼 대다수의 후보 이미지 집합을 처음부터 검색 대상에서 제외하고 검색을 수행할 수 있어 접근되는 클러스터의 개수 혹은 이미지 데이터의 개수를 줄일 수 있다는 장점이 있다.

본 논문에서는 이러한 사항을 고려하여 사용자가 최소한의 이미지 데이터만을 접근하여 자신이 원하는 이미지와 유사한 이미지를 검색할 수 있도록 클러스터의 크기와 데이터의 개수를 동시에 고려한 계층적 클러스터링 방식을 제안한다.

2.2. 제안하는 데이터베이스 구조

그림 1은 제안하는 데이터베이스 구조를 나타낸 것이다. 각 계층의 비단말 노드는 미리 정해진 n개 이하의 클러스터들 $[C_0, C_1, ..., C_k](0 <= k < n)$로 구성되며, 각 클러스터 C_i는 클러스터의 크기 C_i^{size}, 클러스터 내의 데이터 개수 C_i^{num}, 그리고 대표 객체 C_i^{rep} 에 대한 정보와 하위 노드에 대한 포인터 정보를 엔트리로 저장한다. 각 비단말 노드의 클러스터들은 각 클러스터가 일정 개수까지의 이미지 데이터를 포함할 때까지 계속 분할되며 이 과정을 통해 단말 노드까지의 계층 구조가

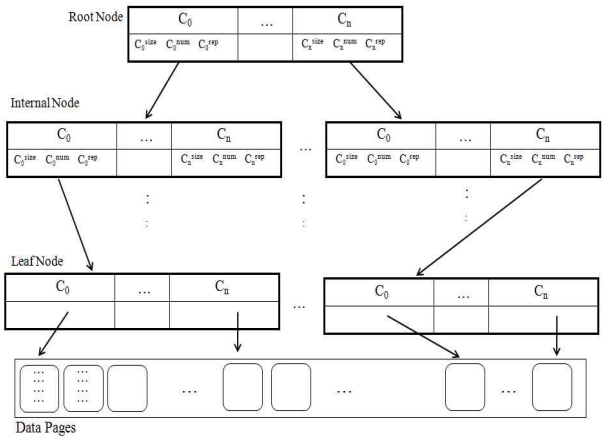

그림 1. 제안하는 데이터베이스 구조
Fig. 1. Overview of proposed database structure.

형성된다. 단말 노드내의 각 클러스터 C_i는 클러스터 내에 포함되는 일정 개수 이하의 유사한 배아 이미지 데이터들이 모여 저장된 데이터 페이지에 대한 포인터 정보를 엔트리로 저장한다.

제안하는 데이터베이스 구조를 이용하면 브라우징 기반 검색 시 사용자는 최상위 노드에서 출발하여 하위 노드로 내려가면서 노드의 각 클러스터들의 대표 객체들을 살펴보면서 자신이 원하는 배아 이미지를 찾을 수 있다. 하위 노드로 내려갈수록 사용자가 원하는 이미지와 유사한 배아 이미지 데이터들을 좀 더 구체적으로 파악 할 수 있고, 최종적으로 단말 노드에서 사용자가 원하는 배아 이미지 데이터와 유사한 이미지 데이터를 찾을 수 있다. 만약 원하는 배아 이미지 데이터가 없다면 다시 상위 노드로 이동하여 원하는 배아 이미지를 검색하게 된다.

2.3. 그래프 기반의 계층적 클러스터링 방법

본 절에서는 Ⅱ장 2절에서 제안한 데이터베이스 구조를 구현하기 위한 그래프 기반의 계층적 클러스터링 방법을 설명한다. 제안된 계층적 클러스터링 방법의 1단계에서는 모든 배아 이미지 데이터 쌍 간의 유사도를 측정한 후 이를 기반으로 유사도 그래프를 생성한다. 2단계에서는 생성된 유사도 그래프를 기반으로 분할 클러스터링 알고리즘을 수행하여 미리 정해진 n개 이하의 클러스터들을 생성한다. 3단계에서는 생성된 각 클러스터에 대하여 대표 객체를 선정한다. 4단계에서는 구성된 n개의 클러스터들 마다 일정 개수 이하의 이미지 데이터가 포함될 때까지 2와 3단계를 반복 수행하여 계층 구조를 형성한다.

가. 유사도 그래프

제안하는 방안은 배아 이미지 데이터에서 RGB 공간상의 컬러 히스토그램을 추출하여 클러스터링에 사용한다. 따라서 기존의 이미지 검색 방법에서 사용하는 유사도 측정 방식인 Euclidean distance, histogram intersection[11], cross-talk distance[12] 등을 사용하여 모든 이미지 쌍 간의 유사도를 측정하고, 이를 이용하여 클러스터링을 수행할 수 있다. 그러나 컬러 히스토그램은 고차원 데이터이기 때문에 이를 이용하여 클러스터링을 수행할 경우 차원의 저주 (dimensionality curse) 문제가 발생한다. 차원의 저주란, 차원의 수가 증가할수록 최 근접 이웃 간 거리와 가장 멀리 떨어진 이웃 간 거리가 상대적으로 가까워지는 문제이다. 따라서 두 이미지간의 유사성을 판단하기가 어렵다. 또한, 차원의 수가 증가함에 따라 두 객체 간 거리에 대한 계산 성능이 지수 함수적으로 악화된다는 문제점을 가지고 있다[13~15].

본 연구에서는 이러한 문제점을 해결하고자 그래프 기반의 클러스터링을 수행한다. 즉, 배아 이미지 데이터를 d차원 공간상의 점이 아니라 그래프 상의 노드로 표현하고, 노드와 노드 사이의 간선에는 이미지 데이터 간의 유사도를 표현하여 그래프를 구성한 후, 이를 기반으로 클러스터링을 수행하는 것이다. 이때, 모든 이미지 쌍 간의 유사도를 측정하여 간선으로 연결하기에는 많은 계산량이 요구되기 때문에 간선으로 연결할 노드의 개수를 제한하는 방식을 사용할 수 있다. 제안하는 방안에서는 유사도 기반의 k-이웃 질의(k-nearest neighbor query)를 수행하여 검색된 k개의 이미지만을 간선으로 연결하여 표현하는 k-NN 방식을 사용한다.

그래프 구성 방식으로는 대칭(symmetric)과 비대칭(asymmetric) 구성 방식이 존재한다[13]. 대칭 방식은 이미지 A의 유사 이미지로 이미지 B가 검색되고, 역으로 이미지 B의 유사 이미지로 이미지 A가 검색된 경우에만 이들 이미지 A와 B를 간선으로 연결하는 방식이다. 반면에 비대칭 방식은 이미지 A의 유사 이미지로 이미지 B가 검색되거나 또는 이미지 B의 유사 이미지로 이미지 A가 검색이 되는 경우 중 한 가지만 만족해도 이들 이미지를 간선으로 연결하는 방식이다.

m개의 이미지를 표현하기 위한 그래프 $G=(V,E)$에서 노드 V는 $\{1,2,...,m\}$개의 이미지로 구성되고, 노드와 노드 사이의 간선 $E=\{(i,j):i,j\in V\}$로 표현되며 간선 (i,j)는 노드 i와 j사이의 유사도 w_{ij}를 가

중치로 갖는다. 제안하는 방안에서 구성된 그래프 G는 유사도 w_{ij}를 ij번째 엔트리로 갖는 matrix W로 표현하여 저장된다.

나. 계층적 클러스터링

본 연구는 최종적으로 원하는 이미지를 빠르고 쉽게 찾는 것을 목표로 한다. 이러한 목표를 만족시키기 위해서 제 Ⅱ장 1절에서 데이터베이스 구조화시에 고려할 사항 4가지 요소인 (1)클러스터의 크기, (2)클러스터 내의 포함된 데이터 개수, (3)계층 구조의 균형 정도, (4) 접근되는 클러스터의 개수와 데이터의 개수를 고려하는 분할 방식 기반의 계층적 클러스터링 방법을 제안한다.

제안하는 방안은 구성된 유사도 그래프를 그래프 분할 알고리즘을 반복 적용하여 클러스터링을 수행함으로써 계층 구조를 형성하는 방식이다. 그래프 분할 알고리즘으로는 spectral[14-15], metis[16-17] 등을 사용할 수 있다. 본 연구에서는 k-NN 그래프를 간선 절단(edge cut)이 최소화되도록 이분할(bi-partitioning)하는 알고리즘인 hMetis[18]를 사용한다. 구성된 유사 그래프의 각 간선은 배아 이미지 데이터 간에 유사도를 표현하므로 hMetis를 이용하여 서로 다른 클러스터 내에 있는 이미지 데이터들 간의 관련성을 최소화할 수 있다.

그림 2는 제안하는 계층적 클러스터링 알고리즘이다. 제안하는 알고리즘은 먼저 hMetis 그래프 분할 알고리즘을 반복 수행하여 n개의 클러스터(n은 사용자에 의해 미리 정의된 클러스터 개수)를 생성한다. 이때 클러스터의 크기를 고려하기 위하여 hMetis로 이분할하기 위한 대상 클러스터 C_b를 클러스터의 크기를 기준으로 선정한다(라인 7-9). 최소 유사도 값이 가장 작은 클러스터를 hMetis 알고리즘을 사용하여 이분할 대상 클러스터 C_b로 선정하여 분할한다. 이처럼 클러스터의 크기로 최소 유사도를 사용할 경우 특정 클러스터가 지나치게 커지는 것을 방지할 수 있다. 다음, 제안하는 알고리즘은 클러스터 내의 데이터 개수를 고려하기 위하여 구성된 n개의 클러스터 중에서 데이터의 개수가 주어진 허용치(threshold) O_{max} 보다 큰 각각의 클러스터들에 대하여 다시 그래프 분할 알고리즘을 반복 수행하여 n개의 클러스터로 재분할한다(라인 10-15).

제안하는 계층 알고리즘은 이와 같이 알고리즘 수행 중에 클러스터의 크기와 데이터의 개수를 지속적으로

```
1: Perform a k-NN query for every object using a similarity matrix
2: Draw a weighted graph using each object as a node and each k-NN relationship as an edge; Put similarity between two objects on the edge for their edge
3: Noise detection and elimination
4: Perform hierarchical clustering HC(C) as follows
5:     #C = 1 //#C is the number of clusters
6:     While (#C is smaller than n)
7:         Choose the biggest cluster Cb
8:         Perform clustering on Cb in binary graph partitioning fashion
9:         #C = #C+1
10:    For each cluster Ci
11:        IF (#Oi is smaller than Omax) //#Oi is the number of data in cluster Ci
12:            Regard Oi as members of Ci
13:        ELSE
14:            Cluster_Set = HC(Ci)
15:            Regard cluster_Set as sub-clusters of Ci
16:    Return all Cis
```

그림 2. 분할기반 계층적 클러스터링 알고리즘
Fig. 2. Hierarchical clustering algorithm based on graph partitioning fashion.

체크하여 클러스터의 크기가 매우 커지거나 클러스터 내에 포함되는 이미지 데이터의 개수가 많아지는 것을 방지함으로써 계층 구조의 균형 정도를 고려한다.

다. 대표 객체 선정 방법

앞서 설명한 바와 같이 계층 구조의 특징을 이용하면 사용자가 원하는 배아 이미지와 거리가 먼 다수의 이미지를 처음부터 배제시킬 수 있으므로 이미지 검색을 위하여 접근해야 하는 이미지의 개수를 크게 줄일 수 있다. 그러나 본 연구에서는 이에 더하여 대표 객체를 선정하는 방식을 이용하여 접근해야 하는 이미지의 개수를 최소화 시킨다.

계층 구조의 루트 노드와 비단말 노드를 구성하는 클러스터들은 대표 객체를 선정하여 저장함으로써 해당 클러스터안의 유사 이미지들을 직접 보지 않고도 대표 객체를 통해 클러스터의 특징을 파악 할 수 있게 한다. 단말 노드를 구성하는 클러스터에는 유사한 배아 이미지 데이터들이 모여 저장된 데이터 페이지에 대한 포인터 정보를 엔트리로 저장한다. 따라서 브라우징 기반 검색을 수행하는 사용자는 상위 노드에서 자신이 원하는 이미지 데이터와 가장 유사한 대표 객체를 선택하여 하위 노드로 점진적으로 내려가면서 검색을 진행하게 된다. 단말 노드에 도달하게 되면 클러스터 내의 모든 유사 이미지들을 사용자가 직접 확인함으로써 자신이 최종 원하는 이미지 데이터를 찾을 수 있다.

그래프 구조에서 대표 객체를 선정하는 가장 간단한 방법으로는 그래프에서 차수(degree)가 가장 많은 노드의 이미지를 대표 이미지로 선정하는 방식이 있다. 그러나 이 방식은 차수만을 고려하기 때문에 간선에 표현된 이미지 데이터간의 유사도 값이 무시된다는 문제점이 있다. 이러한 문제점을 해결하기 위하여 본 연구에서는 클러스터 내의 모든 이미지들과 가장 유사한 이미지를 대표 객체로 선정하는 방식을 사용한다. 즉, 클러스터 내의 각 이미지에 대하여 평균 유사도를 구한 후, 이 값이 가장 큰 이미지를 대표 객체로 선정한다.

matrix W로 표현된 그래프 $G=(V,E)$를 클러스터링 하여 얻어진 클러스터들이 $\{C_1, C_2, ..., C_n\}$ $(C_i \cap C_j = \varnothing\ (i \neq j))$이고 $\bigcup_{i=1}^{n} C_i = V$임)라면, 클러스터 C_i의 대표 객체 노드는 다음과 같이 평균 유사도가 가장 큰 노드가 된다.

$$\max(\arg_{j \in C_i} \sum_{t \in C_i} w_{jt}) \quad (1)$$

Ⅲ. 실 험

3.1. 실험 방법

본 실험에서는 Berkeley drosophila genome project (BDGP)에서 사용된 초파리 배아 이미지 데이터베이스를 실험 데이터로 사용하였다[19]. 초파리 배아 데이터는 한 세대가 짧아 개체발생 결과를 빨리 볼 수 있고, 번식력이 강하며, 인공적인 교배를 통해 원하는 형질을 쉽게 얻을 수 있기 때문에 발생생물학에서 유용하게 사용된다[20].

BDGP에서 사용된 배아 이미지 데이터 중 5개 이상의 태그 정보를 가진 7,489개의 배아 이미지 데이터를 실험에 사용하였다. 태그 정보는 초파리 배아의 특성을 설명해 놓은 것으로서 BDGP에서 사용된 데이터를 통틀면 총 112개 이다. 각 배아 이미지 데이터들은 311차원의 RGB 색상 히스토그램으로 표현되어 있다.

본 논문에서는 이러한 색상 특성을 이용하여 제안하는 계층적 클러스터링을 수행한 결과가 실제로 동일한 태그 정보를 갖는 배아 이미지들이 같은 클러스터에 속하게 되는지에 대해 실험한다. 또한, 제안하는 방법을 통한 브라우징 기반 검색이 기존의 병합적인 계층적 클러스터링을 통한 검색보다 얼마나 효율적인지를 실험을 통해 보인다.

배아 이미지간의 유사도를 계산하기 위해 사용된 거리 함수는 다음과 같다[20]. x_i와 x_j는 이미지 데이터를 벡터로 표현한 것이고, $d_{i,j}$는 백터 x_i와 x_j간의 유사도를 의미한다.

$$d_{i,j} = \left(\frac{(x_i - \overline{x_i})(x_i - \overline{x_i})'}{[(x_i - \overline{x_i})(x_i - \overline{x_i})']^{1/2}[(x_j - \overline{x_j})(x_j - \overline{x_j})']^{1/2}} \right) \quad (2)$$

실험의 비교 대상으로는 기존의 병합적인 계층적 클러스터링 방식 중 하나인 CHAMELEON을 이용한다.

본 논문에서 수행한 첫 번째 실험은, 제안하는 기법의 효율성을 알아보기 위해 사용자가 자신이 원하는 배아 이미지를 찾기 위해 접근하는 클러스터의 평균 접근 횟수와 구축된 계층 구조의 깊이(depth)에 대해 알아본다. 클러스터의 평균 검색 횟수를 측정하기 위하여 사용자가 자신이 원하는 배아 이미지를 찾기 위해 모든

클러스터들을 균등(uniform)하게 접근한다고 가정하고 루트 노드에서 단말 노드까지의 클러스터 평균 접근 횟수를 측정한다. 또한 계층 구조의 깊이로는 구축된 트리 구조의 가장 깊은 깊이를 측정한다.

두 번째 실험으로는 제안하는 기법의 정확도를 알아보기 위하여 단말 노드에 속한 모든 클러스터들의 변형 자카드 계수(varient of Jaccard coefficient)를 측정하여 그 평균을 구하였다[21~22]. 변형 자카드 계수란 한 클러스터 안에 존재하는 두 배아 이미지가 적어도 하나의 같은 태그를 가지고 있으면 올바르게 클러스터링이 됐다고 정의한다. 이에 따라 한 클러스터의 정확도는 올바르게 클러스터링 된 배아 이미지 쌍들의 수를 동일 클러스터 내에 있는 모든 이미지들의 쌍의 수로 나눈 것이다.

3.2. 그래프 구성 방법

제안하는 기법으로 데이터베이스를 구조화하기 위해서는 배아 이미지들의 유사도 그래프를 생성해야 한다. 본 실험에서는 비대칭(asymmetric) 방식을 이용하여 유사도 그래프를 생성하고, 이 때 한 노드에서 다른 노드들과 유사도 간선을 연결하는 k-NN 방식의 k값을 20으로 설정하였다. 즉, 유사도 그래프 생성을 위해 한 이미지와 유사도 계산을 하는 다른 이미지들은 그 이미지와 가장 유사한 20개의 이미지들이다. 그래프 분할 알고리즘 hMetis를 수행할 때마다 생성되는 클러스터의 개수 n=7로 설정하였고, 클러스터 내의 최대 포함 이미지 개수는 50개로 설정하였다. k-NN의 k값, 클러스터의 개수, 그리고 클러스터 내의 최대 포함 이미지 개수를 바꿔가며 사전 실험한 결과, 위와 같은 값을 설정하였을 때 가장 높은 정확도를 보이는 것으로 나타났기 때문이다.

3.3. 실험 결과

표 1은 제안하는 방안으로 구현된 트리 구조내의 클러스터 평균 접근 횟수와 트리의 최대 깊이를 나타낸 것이다. 실험 결과, 제안하는 방안은 CHAMELEON에 비해 88% 정도 평균 접근 횟수가 감소하였고 트리 구조의 최대 깊이도 크게 낮은 것으로 나타났다. CHAMELEON과 같은 기존의 병합적인 계층적 클러스터링 방법은 하나의 클러스터만이 계속 커지는 경향을 보이기 때문에 계층적 클러스터링의 트리 구조가 경사하게 되어 불균형을 이룰 가능성이 매우 높다. 또한 클러스터의 크기나 그 클러스터 안의 이미지의 개수를 고려하지 않기 때문에 하나의 클러스터가 계속 커지는 것을 방지할 수 없다. 결과적으로 트리가 심한 불균형 구조를 이루게 되어 표 1에서 보는 바와 같이 트리 구조의 최대 깊이가 매우 클 수밖에 없고 이에 따라 클러스터 평균 접근 횟수 또한 크게 증가한다. 이에 비해 제안하는 방안은 클러스터의 크기와 그 안의 이미지 개수를 고려하여 분할적인 계층적 클러스터링을 수행하기 때문에 한 클러스터만이 커지는 현상을 방지할 수 있다. 결과적으로 균형에 가까운 계층 구조를 이루게 되어 트리의 최대 깊이가 매우 낮아지게 되고 클러스터 평균 접근 횟수 또한 크게 감소한다.

표 1. 트리의 평균 검색 횟수와 트리구조의 깊이
Table 1. Average number of searches and depth of the tree.

	평균 접근 횟수	트리 구조의 깊이
제안하는 방안	2.6	4
CHAMELEON	23.2	72

표 2. 단말노드에 속한 클러스터의 총 개수와 평균정확도
Table 2. Number of clusters and average accuracy within leaf nodes.

	단말 노드에 속한 클러스터의 총 개수	평균 정확도
제안하는 방안	254	0.841
CHAMELEON	316	0.842

표 2는 변형 자카드 계수로 계산된 모든 클러스터들의 평균 정확도와 단말 노드에 속한 클러스터의 총 개수를 의미한다. 실험 결과, 제안하는 방안과 CHAMELEON이 거의 같은 평균 정확도를 보이지만 단말 노드에 속한 클러스터의 총 개수는 제안하는 방안이 더 낮은 것으로 나타났다. 일반적으로 클러스터의 개수가 많을수록 클러스터의 정확도가 더 높다[21]. 그 이유는 클러스터의 개수가 많을수록 클러스터의 평균 크기가 작아져서 한 클러스터 안에 유사한 이미지가 들어갈 가능성이 높아지기 때문이다. 그러나 클러스터의 크기가 너무 작으면 사용자가 브라우징 기반 검색 시 직접 살펴봐야 할 클러스터의 개수가 많아지기 때문에

검색 시간이 크게 증가할 수밖에 없다. 제안하는 기법이 CHAMELEON에 비해 적은 클러스터의 개수를 유지하면서도 같은 평균 정확도를 보인다는 것은 제안하는 기법으로 구성된 클러스터들이 적당한 크기를 유지하면서도 그 안에 구성되어 있는 이미지들의 특성이 매우 유사하다는 의미이다.

즉, 제안하는 기법을 이용하면 경사되지 않고 균형에 가까운 계층적 클러스터를 구축할 수 있기 때문에 브라우징 기반 검색 시 사용자가 직접 접근해야하는 클러스터와 이미지 개수를 크게 줄일 수 있다. 또한, 각 클러스터안의 이미지들이 유사도가 높기 때문에 사용자가 원하는 이미지를 쉽게 얻을 수 있다.

그림 3은 본 논문에서 사용자의 브라우징 기반 검색을 돕기 위하여 구현한 시각화 툴의 전체적인 모습이다. 그림 3에서 보는 바와 같이 사용자는 비단말 노드에 속하는 7개 클러스터들의 대표 이미지를 직접 확인할 수 있다. 클러스터의 개수는 사용자에 의해 조절가능하다. 그림 3에서 맨 오른쪽 큰 그림은 왼쪽 7개의 대표 이미지 중 사용자가 선택한 대표 이미지를 확대한 그림이다. 사용자가 현재 선택한 클러스터의 하위 노드를 구성하는 다른 클러스터들을 확인하기 위해서는 그림 3에 오른쪽 하단에 있는 down 버튼을 누르면 된다. 이와 마찬가지로 현재 클러스터의 상위 노드를 구성하는 클러스터들을 확인하고 싶으면 오른쪽 상단에 있는 up 버튼을 눌러 확인할 수 있다. 또한 사용자가 현재 선택한 대표 이미지와 같은 클러스터에 있는 유사한 배아 이미지들을 보기 원한다면 각 배아 이미지 하단에

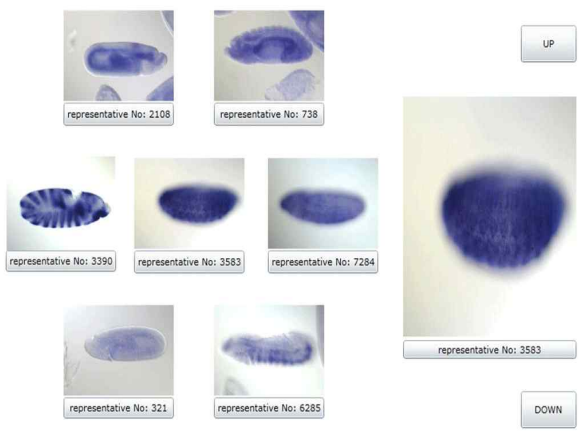

그림 3. 시각화 툴을 통해서 본 비단말 노드에 속하는 클러스터들의 대표객체
Fig. 3. Representative objects of clusters within non-leaf nodes.

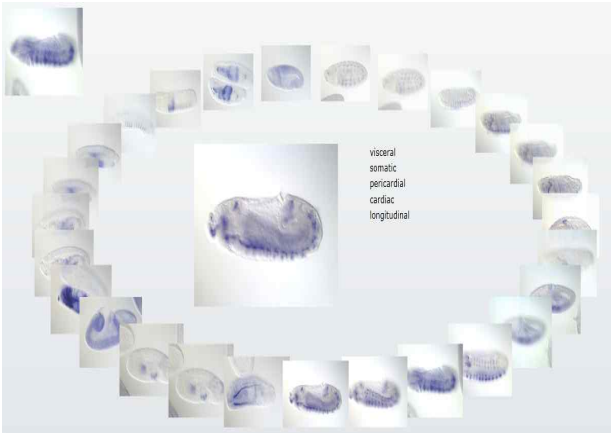

그림 4. 시각화툴을 통해서 본 단말 노드에 속하는 클러스터
Fig. 4. Embryo images of a cluster within leaf node.

있는 이미지 번호를 누르면 된다.

그림 4는 시각화 툴을 통해서 본 단말 노드에 속하는 클러스터이다. 배아 이미지 하단의 이미지 번호를 누르면 이와 같이 그 배아 이미지와 같은 클러스터에 속하는 유사 이미지들을 확인할 수 있다. 사용자는 이 클러스터안의 이미지를 선택함으로써 해당 이미지의 태그 정보를 확인할 수 있다. 그림 4의 가운데에 큰 이미지가 현재 사용자가 선택한 이미지이며, 해당 이미지의 오른쪽에서 태그 정보를 확인할 수 있다.

Ⅳ. 결 론

본 논문에서는 대용량 배아 이미지 데이터베이스의 효과적인 구조화를 위해서 분할기반 계층적 클러스터링 알고리즘을 제안하였다. 기존의 계층적 클러스터링 방법은 데이터베이스를 트리 형태로 구조화하는 과정에서 클러스터의 크기와 클러스터 내의 객체 수를 동시에 고려하지 못하기 때문에 계층적 클러스터링의 결과로 경사 트리 형태의 구조가 될 가능성이 매우 높다. 경사 트리 형태로 데이터베이스가 구조화되면 브라우징 기반 검색 시 사용자가 원하는 이미지를 검색하기 위해 많은 시간이 소요될 수밖에 없다.

제안한 방안은 배아 이미지들을 유사도 그래프로 표현한 후 그래프 분할 알고리즘을 반복적으로 수행하여 계층 구조를 구축하는데 이때, 클러스터의 크기와 클러스터 내에 포함되는 이미지 수를 동시에 고려하여 특정 클러스터의 크기가 지나치게 커지거나 이미지 수가 많아지는 것을 방지하였다. 실험 결과, 제안하는 기법은

경사되지 않고 균형에 가까운 계층 구조를 구축하여 브라우징 기반 검색 시 사용자가 직접 접근해야하는 클러스터와 이미지 개수를 크게 줄이는 것으로 나타났다. 또한, 각 클러스터안의 이미지 유사도 또한 높게 측정되었다. 더불어 본 논문에서는 시각화 툴을 제공함으로써 사용자가 원하는 배아 이미지를 쉽게 찾을 수 있도록 하였다.

감사의 글

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것입니다(2010-0004815와 2008-0061006). 또한 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업(NIPA-2011-C6150-1101-0001)의 부분적인 지원과 정보통신산업진흥원의 IT/SW 창의연구과정의 연구결과로 지식경제부와 삼성전자주식회사에 의해 지원된 과제로 수행되었습니다(NIPA-2010-(C1810-1003-0007)).

또한, 2011년도 두뇌한국21사업에 의하여 지원되었습니다. 그리고 이 논문에서 제안한 데이터베이스 구조를 위한 시각화 툴의 구현을 도와준 한양대학교 컴퓨터학과 김형규 학생과 홍석민 학생에게 감사드립니다.

참 고 문 헌

[1] U. Tepass and V. Hartenstein, “The Development of Cellular Junctions in the Drosophila Embryo,” *Developmental Biology*, Vol. 161, No. 2, pp. 563-596, 1994.

[2] S. Gilberta, J. Opitzc, and R. Raff, “Resynthesizing Evolutionary and Developmental Biology,” *Developmental Biology*, Vol. 173, No. 2, pp. 357-372, 1997.

[3] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[4] G. Karypis, E. Han, and V. Kumar, “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling,” *IEEE COMPUTER: Special Issue on Data Analysis and Mining*, Vol. 32, No. 8, pp. 68-75, 1999.

[5] T. Zhang, R. Ramakrishnan, and M. Linvy, “BIRCH: An Efficient Data Clustering Method for Large Databases,” In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 103-114, 1996.

[6] S. Guha, R. Rastogi, and K. Shim, “CURE: An Efficient Clustering Algorithm for Large Databases,” In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 73-84, 1998.

[7] S. Guha1, R. Rastogi and K. Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes,” *Information Systems*, Vol. 25, No. 5, pp. 345-366, 2000.

[8] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, “Automatic Subspace Clustering of High Dimensional Data,” *Data Mining and Knowledge Discovery*, Vol. 11, No. 1, pp. 5 - 33, 2005.

[9] C. Böhm, K. Kailing, P. Röger, and A. Zimek, “Computing Clusters of Correlation Connected Objects,” In *Proceedings of the ACM International SIGMOD Conference on Management of Data*, pp. 455-466. 2004.

[10] A. Tung, X. Xu, and B. Ooi, “CURLER: Finding and Visualizing Nonlinear Correlation Clusters,” In *Proceedings of the ACM International SIGMOD Conference on Management of Data*, pp. 467-478, 2005.

[11] A. Barla, F. Odone and A. Verri, “Histogram Intersection Kernel for Image Classification,” In *Proceedings of the ICIP International Conference on Image Processing*, pp. 513-516. 2003.

[12] C. Faloutsos, *Searching Multimedia Databases by Content*, Kluwer Academic Publishers, 1996.

[13] 오 현교, 윤 석호, 김 상욱, “이미지 데이터베이스에서 매개변수를 필요로 하지 않는 클러스터링 및 아웃라이어 검출 방법,” *대한전자공학회논문지*, 제 47권 1호, 80-91쪽, 2010년.

[14] R. Kannan, S. Vempala, and A. Vetta. “On Clusterings Good, Bad, and Spectral,” In *Proceedings of the Annual Symposium on Foundations of Computer Science*, 2000.

[15] A. Ng, M. Jordan, and Y. Weiss. “On Spectral Clustering: Analysis and an Algorithm,” In *Proceedings of Neural Information Processing Systems*, 2001.

[16] G. Karypis and V. Kumar, “METIS 4.0: Unstructured Graph Partitioning and Sparse Matrix Ordering System,” Technical Report, Department of Computer Science, University of Minnesota, 1998; http://www.cs.umn.edu/~metis.

[17] G. Karypis and V. Kumar, “Multilevel Algorithms for Multi-constraint Graph Partitioning,” *Journal of Parallel and Distributed Computing*, Vol. 48, No. 1, pp. 96-129, 1998.

[18] G. Karypis and V. Kumar, “hMETIS 1.5: A Hypergraph Partitioning Package,” Technical Report, Department of Computer Science, University of Minnesota, 1998; http://winter.cs.umn.edu/~karypis/metis.
[19] BDGP: Berkeley Drosophila Genome Project (fruitfly.org).
[20] E. Frise, A. Hammonds1 and S. Celniker, “Systematic Image-driven Analysis of the Spatial Drosophila Embryonic Expression Landscape,” *Molecular Systems Biology*, Vol. 6, No. 345, pp. 1 - 15, 2010.
[21] X. Yin, J. Han, and P. S. Yu, “Linkclus: Efficient Clustering via Heterogeneous Semantic Links,” In *Proceedings of the International Conference on Very Large Data Bases*, pages 427-438, 2006.
[22] 송 석순, 김 상욱, 윤 석호, “블로그 공간에서의 링크 기반 클러스터링 방안,” *대한전자공학회논문지*, 제46권 3호, 42-49쪽, 2009년 5월.

저 자 소 개

원 정 임(정회원)

1992년 한림대학교 전자계산학과 학사 졸업.
1997년 한림대학교 컴퓨터공학과 석사 졸업.
2004년 한림대학교 컴퓨터공학과 박사 졸업.
2004년～2006년 연세대학교 컴퓨터공학과 연구교수
2006년～2010년 한양대학교 BK21 연구교수
2010년～현재 한양대학교 전기정보통신기술 연구소 연구교수

<주관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스, 이동 객체 데이터베이스/텔레매틱스, 공간 데이터베이스/GIS>

장 민 희(학생회원)

2003년 홍익대학교 신소재공학과 학사 졸업.
2006년 한양대학교 전자컴퓨터통신학과 석사 졸업.
2006년～현재 한양대학교 전자컴퓨터통신공학과 박사재학.

<주관심 분야: 데이터베이스 시스템, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 이동객체 데이터베이스, 사회 연결망 분석>

오 현 교(학생회원)

2008년 한양대학교 정보통신학과 학사 졸업.
2010년 한양대학교 전자컴퓨터통신공학과 석사 졸업.
2010년～현재 한양대학교 전자컴퓨터통신공학과 박사재학.

<주관심 분야: 사회 연결망 분석, 신뢰 관리, 인터넷 포탈 데이터 분석, e-비즈니스, 데이터 마이닝>

김 상 욱(평생회원)-교신저자

1989년 서울대학교 컴퓨터공학과 학사 졸업.
1991년 한국과학기술원 전산학과 석사 졸업.
1994년 한국과학기술원 전산학사 박사 졸업.
1991년～1991년 미국 Stanford University, Computer Science Department, 방문 연구원.
1994년～1995년 KAIST 정보전자 연구소 전문 연구원.
1999년～2000년 미국 IBM T.J. Watson Research Center, Post-Doc.
1995년～2003년 강원대학교 정보통신공학과 부교수.
2003년～현재 한양대학교 정보통신대학 정보통신학부 교수.
2009년～2010년 미국 Carnegie Mellon University, Visiting Scholar

<주관심분야 : 데이터베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이터베이스, 이동 객체 데이터베이스/텔레매틱스, 사회 연결망 분석, 웹 데이터 분석>