Em-Viz: 배아 데이터의 효율적 검색을 위한 계층적 구조화 기반의 시각화 도구

오현교*, 장민희*, 김형규**, 홍석민**, 원정임***, 김상욱*
*한양대학교 전자컴퓨터통신공학부
**한양대학교 공과대학 컴퓨터공학부
***한양대학교 전기정보통신 기술연구소
e-mail: {rapkyo, zzmini, jiwon, hkkim, smhong, wook}@agape.hanyang.ac.kr

Em-Viz: A Visualization Tool for Efficient Search of Embryo data based on Hierarchical Organization

*Department of Electronics and Computer Engineering, Hanyang University
**Department of Computer Science and Engineering, Hanyang University
***Research Institute of Electrical and Computer Engineering, Hanyang University

요 약

본 논문에서는 배아 데이터의 효율적인 검색을 지원하는 시각화 도구인 Em-Viz의 설계 및 구현에 관하여 논의한다. Em-Viz는 계층적으로 구조화된 대용량 배아 데이터베이스를 기반으로 구현된 시각화 도구로 사용자가 원하는 배아 데이터를 빠르고 정확하게 검색 할 수 있도록 지원한다.

1. 서론

배아(embryo)란 동물이나 식물과 같은 다세포 생물의 발생 초기 단계를 의미한다. 배아의 단계에서 다세포 생물의 기초적인 특성이 결정되기 때문에 배아는 개체발생의 기구(機構)를 연구하는 중요한 연구대상이 된다[1]. 생물학자들은 배아 연구를 위해 대용량의 배아 데이터베이스를 소유하고 있다. 이러한 대용량 데이터베이스에서 원하는 배아 데이터를 효율적으로 선택하기 위해서는 시각화를 통해 전체 데이터 구조를 살펴 볼 수 있어야 한다.

데이터베이스 구조화를 위해 주로 사용되는 방법으로 계층적 클러스터링이 있다. 계층적 클러스터링은 객체간의 유사도를 계산하여 객체들을 유사한 특징을 가진 클러스터들로 구분하고 이를 트리 구조 형태로 표현하는 방법이다[2]. 기존의 계층적 클러스터링 방법인 CHAMELEON, BIRCH, CURE, ROCK 등은 데이터베이스를 트리 형태로 구조화 하는 과정에서 클러스터의 크기와 클러스터 내의 객체 수를 동시에 고려하지 못하기 때문에 결과 클러스터링 트리가 경사 트리(skewed tree)일 가능성이 매우 높다[3]. 데이터 구조가 경사 트리인 경우 시각화를 통해 데이터를 검색 할 때 트리 순회로 인해 많은 시간을 소모하게 된다. 따라서 본 논문의 저자들은 참고문헌[3]에서 대용량의 배아 데이터를 경사 되지 않으며 균형 상태에 가까운 트리 형태로 구조화하기 위한 방안을 제안했다. 제안된 방안은 클러스터의 크기와 클러스터 내의 객체 수를 동시에 고려하여 특정 클러스터의 크기가 지나치게 커지거나 객체 수가 많아지는 것을 방지한다. 이를 통해 대용량의 배아 데이터들을 경사(skew) 되지 않으며 균형(balance) 상태에 가까운 트리 형태로 구조화할 수 있다.

본 논문에서는 이러한 계층적 데이터 구조를 기반으로 시각적인 정보를 제공하는 도구인 Em-Viz (embryo data visualization)를 제안한다. Em-Viz는 배아 데이터 검색 시 각 클러스터를 대표하는 배아 데이터를 눈으로 직접 확인하면서 검색을 수행할 수 있게 한다. 또한, 이를 통해 구조화된 데이터베이스의 트리 구조를 쉽게 파악할 수 있으므로 사용자가 원하는 배아 데이터 검색을 효율적으로 지원한다.

2. Em-Viz

Em-Viz는 효율적인 배아 데이터 검색을 위한 계층적 클러스터링 기반의 시각화 도구이다. Em-Viz는 Silverlight를 이용하여 구현되었으며 Windows 기반에서 동작한다.

그림 1은 효율적인 시각화를 위한 계층적 데이터 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. 대표 객체(representative object)란 클러스터 내의 모든 배아 데이터들 중에서 클러스터의 특성을 가장 잘 반영하는 배아 데이터를 의미한다. Em-Viz를 통해 계층 구조의 루트 노드와 비단말 노드를 구성하는 클러스터들마다 선정된 대표 객체를 사용자들에게 보여줌으로써 해당 클러스터안의 모든 유사 배아 데이터를 직접 확인하지

않고도 클러스터의 특징을 파악할 수 있고, 이를 통해 더욱 효율적인 검색을 가능하게 한다.

단말 노드내의 각 클러스터 C_i는 클러스터 내의 유사한 배아 데이터들이 저장되어 있는 데이터 페이지에 대한 포인터 정보를 엔트리로 저장한다.

그림 1의 계층적 데이터베이스 구조를 기반으로 구현된 Em-Viz를 통해 사용자는 최상위 노드에서 출발하여 하위 노드로 내려가면서 노드내의 모든 클러스터 각각의 대표 객체들을 살펴보면서 원하는 배아 데이터를 찾을 수 있다. 하위 노드로 내려갈수록 사용자는 원하는 배아 데이터와 유사한 배아 데이터들을 좀 더 구체적으로 살펴 볼 수 있고, 최종적으로 단말 노드에서 사용자가 원하는 배아 데이터와 유사한 배아 데이터를 찾을 수 있다. 만약 원하는 배아 데이터가 없다면 다시 상위 노드로 이동하여 원하는 배아 데이터를 검색하게 된다.

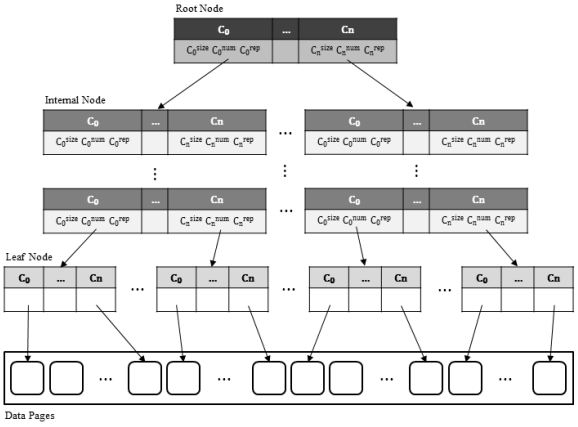

그림 1. 계층적 데이터베이스 구조

데이터 구조화 시 사용자는 클러스터의 개수와 클러스터 내의 최대 포함 데이터 개수를 조절 할 수 있다. 본 논문에서는 각 단계에서 생성되는 클러스터의 개수 n=7로 설정하였고, 클러스터 내의 최대 포함 데이터 개수는 50개로 설정하였다. 그 이유는 클러스터의 개수와 클러스터 내의 최대 포함 데이터 개수를 변화하면서 사전 실험을 수행한 결과 위와 같은 값을 설정하였을 때의 계층적 클러스터링의 정확도가 가장 높았기 때문이다.

본 논문의 저자들은 참고문헌[3]에서 배아 데이터의 검색 효율성을 알아보기 위한 실험과 계층적 클러스터링의 정확도를 측정하기 위한 실험을 하였다. 실험의 비교 대상으로는 기존의 계층적 클러스터링 방식 중 하나인 CHAMELEON을 이용했다.

첫 번째 실험으로 배아 데이터 검색을 효율성을 알아보기 위해 사용자가 자신이 원하는 배아 데이터를 찾기 위해 접근하는 클러스터의 평균 접근 횟수와 구축된 계층 구조의 깊이(depth)을 측정했다.

두 번째 실험은 계층적 클러스터링의 정확도를 알아보기 위하여 단말 노드에 속한 모든 클러스터들의 변형 자카드 계수(varient of Jaccard coefficient)를 측정하여 그 평균을 구했다[4-5].

표 1. 트리의 평균 접근 횟수와 트리구조의 깊이[3]

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

표 1은 참고문헌[3]에서 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 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 배아 데이터의 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 참고문헌[3]에서 제안하는 방안은 클러스터의 크기와 그 안의 배아 데이터 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 2. 단말노드에 속한 클러스터의 총 개수와 평균정확도[3]

	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[4]. 그 이유는 클러스터의 개수가 많을수록 클러스터의 평균 크기가 작아져서 한 클러스터 안에 유사한 배아 데이터가 들어갈 가능성이 높아지기 때문이다. 그러나 클러스터의 크기가 너무 작으면 사용자가 브

라우징 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 배아 데이터들의 특성이 매우 유사하다는 의미이다.

따라서 본 논문에서는 참고문헌[3]에서 제안하는 방법을 이용하여 구축된 경사되지 않고 균형에 가까운 계층적 클러스터를 기반으로 Em-Viz를 구현하였다. Em-Viz에서 제공하는 기능을 이용하면 배아 데이터 검색 시 사용자가 직접 접근해야하는 클러스터와 배아 데이터 개수를 크게 줄일 수 있으며 각 클러스터안의 배아 데이터들의 유사도가 높기 때문에 사용자가 원하는 배아 데이터를 쉽게 얻을 수 있다.

Em-Viz는 사용자에게 크게 2가지의 시각화 기능을 제공한다. 첫 번째로 사용자는 Em-Viz를 통해 각 클러스터들의 대표 배아 데이터들을 살펴 볼 수 있다. 그림 2는 비단말 노드에 속하는 7개의 클러스터들의 대표 배아 데이터다. 그림 2의 오른쪽 큰 배아 데이터는 7개의 대표 배아 데이터 중 사용자가 선택한 배아 데이터를 확대한 것이다. 사용자가 현재 선택한 클러스터의 하위 클러스터들을 확인하기 위해서는 오른쪽 하단의 down 버튼을 누르면 된다. 마찬가지로 현재 선택한 클러스터의 상위 클러스터들을 확인하고 싶으면 오른쪽 상단의 up 버튼은 눌러 확인할 수 있다.

두 번째로 사용자는 단말 노드의 속하는 클러스터 내의 모든 배아 데이터들을 살펴 볼 수 있다. 그림 3은 Em-Viz를 통해서 본 단말 클러스터이다. 사용자는 그림 2의 오른쪽 확대 배아 데이터 하단의 배아 데이터 번호를 클릭함으로써 그 배아 데이터가 속해있는 단말 클러스터 안의 모든 배아 데이터를 확인 할 수 있다. 또한, Em-Viz는 배아 데이터 시각화 정보와 함께 태그 정보를 제공함으로써 보다 구체적으로 해당 배아 데이터의 특징을 파악할 수 있다. 그림 3의 가운데 큰 배아 데이터가 선택된 배아 데이터이며, 배아 데이터의 오른쪽에서 태그 정보를 확인 할 수 있다.

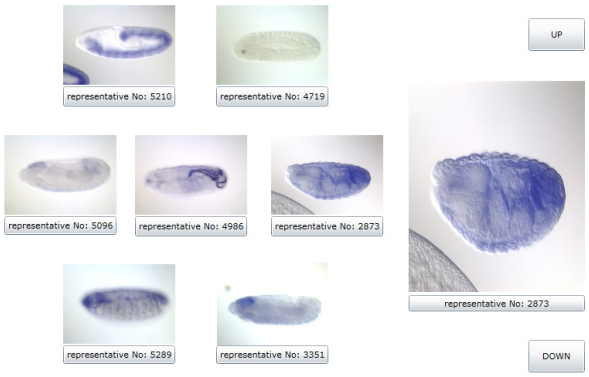

그림 2. Em-Viz를 통해서 본 비단말 노드에 속하는 클러스터들의 대표 배아 데이터

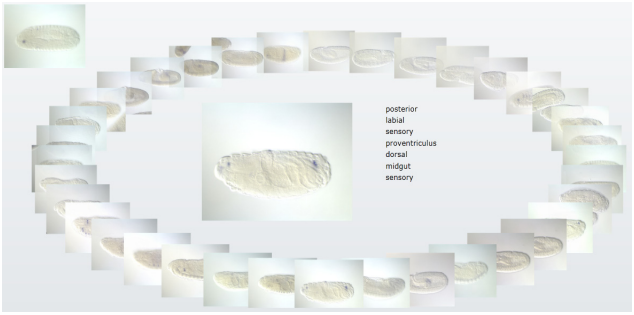

그림 3. Em-Viz를 통해서 본 단말 노드에 속하는 클러스터

3. 결론

본 논문에서는 효율적인 시각화를 위한 계층적 데이터의 구조를 소개했고, 배아 데이터의 효율적 검색을 위해 계층적 구조화 기반의 시각화 도구인 Em-Viz을 구현하였다. Em-Viz는 참고문헌[3]에서 제안된 방법을 이용하여 구축된 경사되지 않고 균형에 가까운 계층적 클러스터를 기반으로 구현되었기 때문에 배아 데이터 검색 시 사용자가 직접 접근해야하는 클러스터의 개수를 크게 줄이며, 각 클러스터의 정확도 또한 높게 측정된다. 따라서 Em-Viz 사용자들에게 빠르고 정확한 배아 데이터 검색 결과를 제공한다.

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)).

참고문헌

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

[3] 원 정임, 오 현교, 장 민희, 김 상욱, “배아 데이터의 효율적 검색을 위한 계층적 구조화 방법,” 대한전자공학회논문지, 2011. 3. (게재 예정)

[4] 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, pp. 427-438, 2006.

[5] 송 석순, 김 상욱, 윤 석호, "블로그 공간에서의 링크 기반 클러스터링 방안," 대한전자공학회논문지, Vol. 46, No. 3, pp. 42-49, 2009. 5.