



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0051300
(43) 공개일자 2020년05월13일

(51) 국제특허분류(Int. Cl.)

G06F 16/00 (2019.01)

(52) CPC특허분류

G06F 16/35 (2019.01)

(21) 출원번호 10-2018-0134483

(22) 출원일자 2018년11월05일

심사청구일자 2018년11월05일

(71) 출원인

서강대학교산학협력단

서울특별시 마포구 백범로 35 (신수동, 서강대학교)

(72) 발명자

정성원

서울특별시 양천구 오목로 300, 604호(목1동, 현대하이페리온2 204동)

유형철

서울특별시 마포구 대흥로 48, 405호(대흥동, 프라임원룸텔)

(74) 대리인

정부연

전체 청구항 수 : 총 10 항

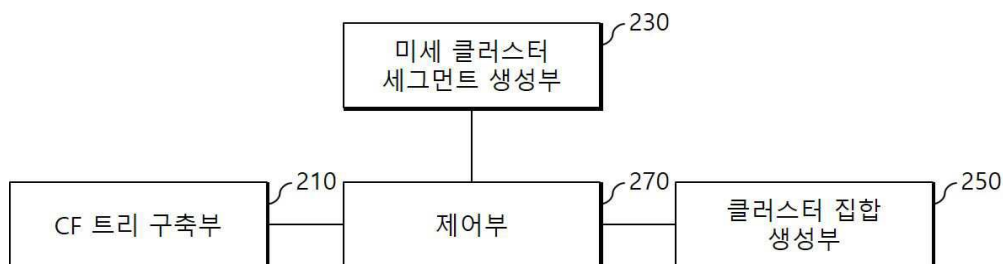
(54) 발명의 명칭 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법

(57) 요약

본 발명은 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법에 관한 것으로, 데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축하는 CF 트리 구축부, 상기 리프노드를 오른쪽 방향으로 순차탐색하면서 상기 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성하는 미세 클러스터 세그먼트 생성부 및 상기 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 클러스터 집합 생성부를 포함한다. 따라서, 본 발명은 CF 트리의 미세 클러스터 간의 유사성 정보를 활용하여 효과적인 클러스터링을 수행할 수 있다.

대표도 - 도2

130



이 발명을 지원한 국가연구개발사업

과제고유번호 201721041

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 일반연구자지원사업

연구과제명 빅데이터 마이닝을 위한 디스크 기반 인덱스 트리를 활용한 군집화 기술 개발

기 여 율 1/1

주관기관 서강대학교

연구기간 2017.11.01 ~ 2018.10.31

명세서

청구범위

청구항 1

데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축하는 CF 트리 구축부;

상기 리프노드를 오른쪽 방향으로 순차탐색하면서 상기 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성하는 미세 클러스터 세그먼트 생성부; 및

상기 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 클러스터 집합 생성부를 포함하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 2

제1항에 있어서, 상기 CF 트리 구축부는

상기 구축이 완료된 경우 상기 보조 클러스터에 대해 반지름을 산출하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 3

제2항에 있어서, 상기 CF 트리 구축부는

상기 반지름 산출 과정에서 상기 보조 클러스터의 자식이 미세 클러스터인 경우 해당 미세 클러스터의 평균 반지름을 사용하고 상기 보조 클러스터의 자식이 보조 클러스터인 경우 해당 보조 클러스터의 반지름을 사용하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 4

제1항에 있어서, 상기 미세 클러스터 세그먼트 생성부는

다음의 수학적식을 통해 상기 미세 클러스터 간의 거리(Inter-Microcluster Distance, IMD)를 산출하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

[수학적식]

$$IMD(MC_x, MC_y) = ED(c(MC_x), c(MC_y)) - (ar(MC_x) + ar(MC_y))$$

(여기에서, MC_x 및 MC_y 는 미세 클러스터에 해당하고, $c(MC)$ 및 $ar(MC)$ 는 해당 미세 클러스터의 중심 및 평균 반지름에 해당하며, $ED(c(MC_x), c(MC_y))$ 는 두 미세 클러스터의 중심 간의 유클리디언 거리(Euclidean Distance, ED)에 해당한다.)

청구항 5

제1항에 있어서, 상기 미세 클러스터 세그먼트 생성부는

제1 및 제2 미세 클러스터들 간의 거리가 임계 거리보다 작은 경우 상기 제1 및 제2 미세 클러스터들을 동일한 미세 클러스터 세그먼트에 추가하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링

장치.

청구항 6

제1항에 있어서, 상기 미세 클러스터 세그먼트 생성부는

상기 미세 클러스터 세그먼트의 중심은 상기 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터들의 중심의 평균값을 통해 산출하고 상기 미세 클러스터 세그먼트의 반지름은 상기 미세 클러스터 세그먼트의 중심에서 각 미세 클러스터의 중심까지의 거리와 각 미세 클러스터의 평균 반지름의 합 중에서 가장 큰 값을 통해 산출하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 7

제1항에 있어서, 상기 클러스터 집합 생성부는

상기 미세 클러스터 세그먼트의 중심을 기준으로 상기 미세 클러스터 세그먼트의 반지름과 임계 거리의 합을 반경으로 하는 범위 영역과 중첩되고 외부 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터 후보들을 결정하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 8

제7항에 있어서, 상기 클러스터 집합 생성부는

상기 복수의 미세 클러스터 후보들 각각에 대해 상기 미세 클러스터 세그먼트에 포함된 모든 미세 클러스터들 간의 거리가 상기 임계 거리 이내인 경우 해당 미세 클러스터 후보가 포함된 외부 미세 클러스터 세그먼트를 상기 미세 클러스터 세그먼트와 연관된 부분 클러스터 집합에 추가하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 9

제8항에 있어서, 상기 클러스터 집합 생성부는

동일한 미세 클러스터 세그먼트를 포함하고 있는 복수의 부분 클러스터 집합들을 하나로 통합하여 상기 클러스터 집합을 생성하는 것을 특징으로 하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치.

청구항 10

데이터 클러스터링 장치에서 수행되는 데이터 클러스터링 방법에 있어서,

데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축하는 단계;

상기 리프노드를 오른쪽 방향으로 순차탐색하면서 상기 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성하는 단계; 및

상기 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 단계를 포함하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 방법.

발명의 설명

기술 분야

[0001] 본 발명은 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 기술에 관한 것으로, 보다 상세하게는 CF 트리의 미세 클러스터 간의 유사성 정보를 활용하여 효과적인 클러스터링을 수행하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법에 관한 것이다.

배경 기술

[0003] 데이터 클러스터 분석 기술, 즉 클러스터링 기법은 거대한 용량의 데이터를 빠른 시간 안에 분석할 수 있는 기법으로 발전하고 있으며, 빅 데이터 시대의 도래와 함께 이 수요는 크게 증가하고 있다. 클러스터링 기법은 전체 데이터의 집합에서 일부의 표본만을 사용하는 표본화 기반 클러스터링 기법, 데이터를 구성하는 여러 속성 값들 중 중요한 속성의 값만 선별해서 사용하는 차원 감소 기반 클러스터링 기법 및 전체 데이터의 집합을 요약한 후 요약된 데이터의 집합으로 클러스터링을 수행하는 요약 기반 클러스터링 기법(혹은 pre-clustering) 등이 연구되었다. 하지만, 표본화 기법은 임의로 선택된 표본이 전체 데이터의 특성을 정확히 나타낼 수 있는가에 따라 성능의 차이가 발생하는 단점이 있고, 차원 감소 기법은 중요한 값이 누락되어 정확도가 떨어지는 단점이 있다.

[0004] 한국등록특허 제10-0987026(2010.10.05)호는 매크로 클러스터링 방법과 그 장치, 및 상기 방법을 구현하는 프로그램이 기록된 기록매체에 관한 것으로, 고가의 물리적인 마이크로어레이 실험 없이 기존의 GEO, ArrayExpress 등과 같이 공개된 마이크로어레이 데이터를 활용하여 생물학적 기능 연관성을 갖는 바이클러스터를 도출할 수 있다.

[0005] 한국등록특허 제10-0803769(2008.02.05)호는 미디어 아이템들을 클러스터링 및 조회하기 위한 방법에 관한 것으로, 모바일 단말들 내에서 계층적 다차원 클러스터들 내에 편성된 미디어 아이템들을 정렬, 보이기, 및 질의하는 기술로서 미디어 아이템들은 다차원 클러스터링을 통해 정렬되고, 한 차원에 대해 최초의 엔트리를 규정하여 조회되며, 다음 엔트리는 그 최초의 엔트리를 만족하는 미디어 아이템들로부터 다른 차원에 기반한다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허 제10-0987026(2010.10.05)호
(특허문헌 0002) 한국등록특허 제10-0803769(2008.02.05)호

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 실시예는 CF 트리의 미세 클러스터 간의 유사성 정보를 활용하여 효과적인 클러스터링을 수행하는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법을 제공하고자 한다.

[0009] 본 발명의 일 실시예는 CF 트리를 구성하는 미세 클러스터에 대한 분할과 정제 과정을 통해 대용량 데이터 셋에 적합하고 높은 클러스터링 성능을 제공할 수 있는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법을 제공하고자 한다.

[0010] 본 발명의 일 실시예는 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트를 생성하고 범위 질의를 통해 직접 또는 간접 연결된 미세 클러스터 세그먼트를 검출함으로써 효과적인 데이터 클러스터링을 수행할 수 있는 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0012] 실시예들 중에서, CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법은 데이터 집합에 포함된

복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축하는 CF 트리 구축부, 상기 리프노드를 오른쪽 방향으로 순차탐색하면서 상기 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성하는 미세 클러스터 세그먼트 생성부 및 상기 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 클러스터 집합 생성부를 포함한다.

- [0013] 상기 CF 트리 구축부는 상기 구축이 완료된 경우 상기 보조 클러스터에 대해 반지름을 산출할 수 있다.
- [0014] 상기 CF 트리 구축부는 상기 반지름 산출 과정에서 상기 보조 클러스터의 자식이 미세 클러스터인 경우 해당 미세 클러스터의 평균 반지름을 사용하고 상기 보조 클러스터의 자식이 보조 클러스터인 경우 해당 보조 클러스터의 반지름을 사용할 수 있다.
- [0015] 상기 미세 클러스터 세그먼트 생성부는 다음의 수학적식을 통해 상기 미세 클러스터 간의 거리(Inter-Microcluster Distance, IMD)를 산출할 수 있다.
- [0016] [수학적식]
- [0017] $IMD(MC_x, MC_y) = ED(c(MC_x), c(MC_y)) - (ar(MC_x) + ar(MC_y))$
- [0018] (여기에서, MC_x 및 MC_y 는 미세 클러스터에 해당하고, $c(MC)$ 및 $ar(MC)$ 는 해당 미세 클러스터의 중심 및 평균 반지름에 해당하며, $ED(c(MC_x), c(MC_y))$ 는 두 미세 클러스터의 중심 간의 유클리디언 거리(Euclidean Distance, ED)에 해당한다.)
- [0019] 상기 미세 클러스터 세그먼트 생성부는 제1 및 제2 미세 클러스터들 간의 거리가 임계 거리보다 작은 경우 상기 제1 및 제2 미세 클러스터들을 동일한 미세 클러스터 세그먼트에 추가할 수 있다.
- [0020] 상기 미세 클러스터 세그먼트 생성부는 상기 미세 클러스터 세그먼트의 중심은 상기 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터들의 중심의 평균값을 통해 산출하고 상기 미세 클러스터 세그먼트의 반지름은 상기 미세 클러스터 세그먼트의 중심에서 각 미세 클러스터의 중심까지의 거리와 각 미세 클러스터의 평균 반지름의 합 중에서 가장 큰 값을 통해 산출할 수 있다.
- [0021] 상기 클러스터 집합 생성부는 상기 미세 클러스터 세그먼트의 중심을 기준으로 상기 미세 클러스터 세그먼트의 반지름과 임계 거리의 합을 반경으로 하는 범위 영역과 중첩되고 외부 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터 후보들을 결정할 수 있다.
- [0022] 상기 클러스터 집합 생성부는 상기 복수의 미세 클러스터 후보들 각각에 대해 상기 미세 클러스터 세그먼트에 포함된 모든 미세 클러스터들 간의 거리가 상기 임계 거리 이내인 경우 해당 미세 클러스터 후보가 포함된 외부 미세 클러스터 세그먼트를 상기 미세 클러스터 세그먼트와 연관된 부분 클러스터 집합에 추가할 수 있다.
- [0023] 상기 클러스터 집합 생성부는 동일한 미세 클러스터 세그먼트를 포함하고 있는 복수의 부분 클러스터 집합들을 하나로 통합하여 상기 클러스터 집합을 생성할 수 있다.
- [0024] 실시예들 중에서, CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 방법은 데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축하는 단계, 상기 리프노드를 오른쪽 방향으로 순차탐색하면서 상기 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성하는 단계 및 상기 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 단계를 포함한다.

발명의 효과

- [0026] 개시된 기술은 다음의 효과를 가질 수 있다. 다만, 특정 실시예가 다음의 효과를 전부 포함하여야 한다거나 다음의 효과만을 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.
- [0027] 본 발명의 일 실시예에 따른 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법은 CF 트리를 구성하는 미세 클러스터에 대한 분할과 정제 과정을 통해 대용량 데이터 셋에 적합하고 높은 클러스터링 성능을

제공할 수 있다.

- [0028] 본 발명의 일 실시예에 따른 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 장치 및 방법은 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트를 생성하고 범위 질의를 통해 직접 또는 간접 연결된 미세 클러스터 세그먼트를 검출함으로써 효과적인 데이터 클러스터링을 수행할 수 있다.

도면의 간단한 설명

- [0030] 도 1은 본 발명의 일 실시예에 따른 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 시스템을 설명하는 도면이다.
- 도 2는 도 1에 있는 데이터 클러스터링 장치를 설명하는 블록도이다.
- 도 3은 도 1에 있는 데이터 클러스터링 장치에서 수행되는 데이터 클러스터링 과정을 설명하는 순서도이다.
- 도 4는 도 1에 있는 데이터 클러스터링 장치에서 활용되는 CF 트리의 일 실시예를 설명하는 도면이다.
- 도 5는 CF 트리 구축 과정에서 사용되는 미세 클러스터 간의 거리를 설명하는 예시도이다.
- 도 6은 CF 트리 구축 과정에서 사용되는 미세 클러스터 세그먼트를 설명하는 예시도이다.
- 도 7은 CF 트리 구축 과정에서 사용되는 미세 클러스터 세그먼트 간의 연결 관계를 설명하는 예시도이다.
- 도 8은 도 1에 있는 데이터 클러스터링 장치에서 미세 클러스터 세그먼트에 대한 범위 질의를 수행하는 과정을 설명하는 예시도이다.
- 도 9는 도 1에 있는 데이터 클러스터링 장치에서 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 과정의 알고리즘을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0031] 본 발명에 관한 설명은 구조적 내지 기능적 설명을 위한 실시예에 불과하므로, 본 발명의 권리범위는 본문에 설명된 실시예에 의하여 제한되는 것으로 해석되어서는 아니 된다. 즉, 실시예는 다양한 변경이 가능하고 여러 가지 형태를 가질 수 있으므로 본 발명의 권리범위는 기술적 사상을 실현할 수 있는 균등물들을 포함하는 것으로 이해되어야 한다. 또한, 본 발명에서 제시된 목적 또는 효과는 특정 실시예가 이를 전부 포함하여야 한다거나 그러한 효과만을 포함하여야 한다는 의미는 아니므로, 본 발명의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.
- [0032] 한편, 본 출원에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.
- [0033] "제1", "제2" 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.
- [0034] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결될 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다고 언급된 때에는 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 한편, 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.
- [0035] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함하다" 또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0036] 각 단계들에 있어 식별부호(예를 들어, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

- [0037] 본 발명은 컴퓨터가 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현될 수 있고, 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.
- [0038] 여기서 사용되는 모든 용어들은 다르게 정의되지 않는 한, 본 발명이 속하는 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한 이상적이거나 과도하게 형식적인 의미를 지니는 것으로 해석될 수 없다.
- [0040] 도 1은 본 발명의 일 실시예에 따른 CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 시스템을 설명하는 도면이다.
- [0041] 도 1을 참조하면, CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 시스템(100)은 사용자 단말(110), 데이터 클러스터링 장치(130) 및 데이터베이스(150)를 포함할 수 있다.
- [0042] 사용자 단말(110)은 대용량 데이터를 제공하여 클러스터링 결과를 확인할 수 있는 컴퓨팅 장치에 해당하고 스마트폰, 노트북 또는 컴퓨터로 구현될 수 있으며, 반드시 이에 한정되지 않고, 태블릿 PC 등 다양한 디바이스로도 구현될 수 있다. 사용자 단말(110)은 데이터 클러스터링 장치(130)와 네트워크를 통해 연결될 수 있고, 복수의 사용자 단말(110)들은 데이터 클러스터링 장치(130)와 동시에 연결될 수 있다.
- [0043] 데이터 클러스터링 장치(130)는 대용량 데이터 집합에 대하여 CF 트리를 활용하여 데이터 클러스터링을 효과적으로 수행할 수 있는 컴퓨터 또는 프로그램에 해당하는 서버로 구현될 수 있다. 데이터 클러스터링 장치(130)는 사용자 단말(110)과 유선 네트워크 또는 블루투스, WiFi 등과 같은 무선 네트워크로 연결될 수 있고, 유선 또는 무선 네트워크를 통해 사용자 단말(110)과 통신을 수행할 수 있다.
- [0044] 일 실시예에서, 데이터 클러스터링 장치(130)는 데이터베이스(150)와 연동하여 대용량 데이터에 관한 클러스터링을 수행하는데 필요한 정보를 저장할 수 있다. 한편, 데이터 클러스터링 장치(130)는 도 1과 달리, 데이터베이스(150)를 내부에 포함하여 구현될 수 있다. 또한, 데이터 클러스터링 장치(130)는 사용자 단말(110)로부터의 데이터 제공이 없더라도 데이터베이스(150)에 저장된 대용량 데이터에 관한 클러스터링을 수행할 수 있으며, 프로세서, 메모리, 사용자 입출력부 및 네트워크 입출력부를 포함하여 구현될 수 있다.
- [0045] 데이터베이스(150)는 데이터 클러스터링 장치(130)가 대용량 데이터 집합에 관한 클러스터링을 수행하는 과정에서 필요한 다양한 정보들을 저장할 수 있다. 예를 들어, 데이터베이스(150)는 클러스터링의 대상이 되는 대용량 데이터 집합을 저장할 수 있고, 대용량 데이터를 기초로 구축된 CF 트리에 관한 정보를 저장할 수 있으며, 반드시 이에 한정되지 않고, 구축된 CF 트리를 활용하여 범위 질의 기반의 데이터 클러스터링을 수행하는 과정에서 다양한 형태로 수집 또는 가공된 정보들을 저장할 수 있다.
- [0047] 도 2는 도 1에 있는 데이터 클러스터링 장치를 설명하는 블록도이다.
- [0048] 도 2를 참조하면, 데이터 클러스터링 장치(130)는 CF 트리 구축부(210), 미세 클러스터 세그먼트 생성부(230), 클러스터 집합 생성부(250) 및 제어부(270)를 포함할 수 있다.
- [0049] CF 트리 구축부(210)는 데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축할 수 있다. CF(Clustering Feature) 트리는 CF 벡터들로 구성된 트리에 해당하고 CF 벡터는 클러스터 내 데이터들로 만들어진 값으로서 각 클러스터의 중심(c:center) 및 평균 반지름(ar:average radius)을 구하는데 사용될 수 있다.
- [0050] CF 트리 구축부(210)에 의해 구축된 CF 트리의 리프노드는 데이터들의 가장 작은 클러스터인 미세 클러스터를 엔트리로 포함할 수 있고 다른 엔트리와 마찬가지로 CF 벡터를 통해 표현될 수 있다. 또한, 인덱스노드는 리프노드 또는 인덱스노드를 자식으로 갖는 내부노드에 해당할 수 있고 보조 클러스터를 엔트리로 포함할 수 있다. CF 트리는 B^+ 트리와 동일하게 내부 노드에 해당하는 인덱스 노드와 리프노드로 구성될 수 있으나 B^+ 트리와 다

르게 데이터가 아닌 미세 클러스터를 관리할 수 있다.

[0051] 일 실시예에서, CF 트리 구축부(210)는 CF 트리의 구축이 완료된 경우 CF 트리를 구성하는 모든 보조 클러스터에 대해 반지름을 산출할 수 있다. 보조 클러스터의 반지름은 CF 트리를 이용한 범위 질의에 사용될 수 있고 범위 질의를 수행할 경우 모든 미세 클러스터 간의 거리 계산을 피하기 위한 방법으로 사용될 수 있다. 이 경우, CF 트리를 구성하는 각 노드는 보조 클러스터에 대한 반지름 정보를 추가로 포함하는 형태로 선언되어 CF 트리 구축에 활용될 수 있다.

[0052] 일 실시예에서, CF 트리 구축부(210)는 반지름 산출 과정에서 특정 보조 클러스터의 자식이 미세 클러스터인 경우 해당 미세 클러스터의 평균 반지름을 사용하고 특정 보조 클러스터의 자식이 보조 클러스터인 경우 해당 보조 클러스터의 반지름을 사용할 수 있다. 예를 들어, 인덱스노드에서 i 번째 보조 클러스터인 SC_i 가 C 클러스터들, 즉 $C_1, \dots, C_c$ 로 구성되고, $ED(c(C_i), c(C_j))$ 가 $c(C_i)$ 와 $c(C_j)$ 간의 유클리디언 거리라고 하면 보조 클러스터 SC_i 의 반지름 $SC_i.r$ 은 다음과 같이 정의될 수 있다.

$$SC_i.r = \begin{cases} \max_{1 \leq j \leq c} \{ED(c(SC_i), c(C_j)) + ar(C_j)\} & \text{if } C_j \text{ is a microcluster} \\ \max_{1 \leq j \leq c} \{ED(c(SC_i), c(C_j)) + C_j.r\} & \text{otherwise.} \end{cases}$$

[0054]

[0056] 특히, CF 트리 구축부(210)는 자식이 보조 클러스터인 경우 다시 자식의 자식을 찾아가 반지름을 구하고, 이러한 과정을 자식이 미세 클러스터일 때까지 재귀적으로 반복 수행할 수 있다. CF 트리 구축부(210)는 CF 트리를 구축한 후 전체 트리를 한번 탐색하여 트리 내 전체 보조 클러스터들에 대한 반지름을 구할 수 있고, 각 보조 클러스터에 대한 반지름은 해당 보조 클러스터의 모든 자손에 해당하는 미세 클러스터들을 포함하는 원의 형태로 표현될 수 있다.

[0057] 미세 클러스터 세그먼트 생성부(230)는 리프노드를 오른쪽 방향으로 순차탐색하면서 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성할 수 있다. CF 트리는 B^+ 트리와 같이 모든 리프노드는 연결리스트 형태로 구현될 수 있다. 따라서, 리프노드에 대한 순차탐색이 가능하고 CF 트리의 가장 왼쪽 리프노드부터 시작하여 링크를 따라 오른쪽 방향으로 진행될 수 있으며 각 리프노드 내에서도 왼쪽부터 오른쪽으로 진행하면서 각 미세 클러스터를 탐색할 수 있다.

[0058] 또한, 미세 클러스터 세그먼트 생성부(230)는 각 미세 클러스터에 대해 오른쪽 형제에 해당하는 미세 클러스터와의 거리를 기초로 미세 클러스터 세그먼트를 생성할 수 있다. 여기에서, 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)는 미세 클러스터 간의 거리가 임계 거리 이내인 미세 클러스터들에 관한 집합에 해당할 수 있다. 임계 거리는 데이터 클러스터링 장치(130)에 의해 사전에 설정될 수 있다.

[0059] 일 실시예에서, 미세 클러스터 세그먼트 생성부(230)는 다음의 수학적식을 통해 미세 클러스터 간의 거리(Inter-Microcluster Distance, IMD)를 산출할 수 있다.

[0061] [수학적식]

$$IMD(MC_x, MC_y) = ED(c(MC_x), c(MC_y)) - (ar(MC_x) + ar(MC_y))$$

[0062]

[0064] 여기에서, MC_x 및 MC_y 는 미세 클러스터에 해당하고, $c(MC)$ 및 $ar(MC)$ 는 해당 미세 클러스터의 중심 및 평균 반지름에 해당하며, $ED(c(MC_x), c(MC_y))$ 는 두 미세 클러스터의 중심 간의 유클리디언 거리(Euclidean Distance, ED)에 해당한다. 결과적으로, 미세 클러스터 간의 거리 IMD는 두 미세 클러스터의 중심 간의 유클리디언 거리에서 각 미세 클러스터의 평균 반지름을 뺀 값을 통해 산출할 수 있다.

[0065] 일 실시예에서, 미세 클러스터 세그먼트 생성부(230)는 제1 및 제2 미세 클러스터들 간의 거리가 임계 거리보다 작은 경우 제1 및 제2 미세 클러스터들을 동일한 미세 클러스터 세그먼트에 추가할 수 있다. 미세 클러스터 세

그먼트 생성부(230)는 인접한 두 미세 클러스터들 간의 거리와 임계 거리를 비교하여 미세 클러스터에 대한 분류를 수행할 수 있다. 결과적으로, 미세 클러스터 세그먼트 생성부(230)는 미세 클러스터 간의 거리가 임계 거리보다 가까우면 동일한 미세 클러스터 세그먼트로 분류하고, 그렇지 않은 경우 서로 다른 미세 클러스터 세그먼트로 분류할 수 있다. 미세 클러스터 세그먼트 생성부(230)는 리프노드에 포함된 모든 미세 클러스터들에 대해 분류를 수행함으로써 복수의 미세 클러스터 세그먼트들을 생성할 수 있다.

[0066] 일 실시예에서, 미세 클러스터 세그먼트 생성부(230)는 미세 클러스터 세그먼트의 중심은 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터들의 중심의 평균값을 통해 산출하고 미세 클러스터 세그먼트의 반지름은 미세 클러스터 세그먼트의 중심에서 각 미세 클러스터의 중심까지의 거리와 각 미세 클러스터의 평균 반지름의 합 중에서 가장 큰 값을 통해 산출할 수 있다. 미세 클러스터 세그먼트는 복수의 미세 클러스터들을 포함하는 집합에 해당하며 미세 클러스터와 동일하게 중심과 반지름이라는 대표값으로서 표현될 수 있다. 미세 클러스터 세그먼트 생성부(230)에 의해 복수의 미세 클러스터 세그먼트들이 생성되면 각 미세 클러스터 세그먼트들은 중심과 반지름의 값으로 표현될 수 있다.

[0067] 예를 들어, 미세 클러스터 세그먼트 MCS_i 가 k 개의 미세 클러스터들 즉, $MC_1, \dots, MC_k$ 로 구성되었다고 가정하면, 각 미세 클러스터의 중심 및 평균 반지름이 $c(MC_1), \dots, c(MC_k)$ 와 $ar(MC_1), \dots, ar(MC_k)$ 인 경우 MCS_i 의 중심은 $MCS_i.c$ 이고 반지름은 $MCS_i.r$ 로 표현될 수 있으며, 각각 다음의 수학적 식 (1)과 (2)를 통해 산출될 수 있다.

[0069] (1)
$$MCS_i.c = \frac{\sum_{y=1}^k c(MC_y)}{k}$$

[0070] (2)
$$MCS_i.r = \max_{1 \leq y \leq k} \{ED(c(MC_y), MCS_i.c) + ar(MC_y)\}$$

[0071] 여기에서, $ED(c(MC_y), MCS_i.c)$ 는 $c(MC_y)$ 와 $MCS_i.c$ 간의 유클리디언 거리에 해당할 수 있다.

[0073] 클러스터 집합 생성부(250)는 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성할 수 있다. 특히, 미세 클러스터 세그먼트 간의 모든 쌍에 대한 IMD 값이 임계 거리보다 작을 때 두 미세 클러스터 세그먼트가 연결되었다고 정의할 수 있다. 이 경우 미세 클러스터 세그먼트들은 서로 직접 또는 간접적으로 연결될 수 있으며, 클러스터 집합 생성부(250)는 직접 또는 간접적으로 연결된 미세 클러스터 세그먼트들을 하나로 통합하여 클러스터 집합을 생성할 수 있다. 클러스터 집합 생성부(250)는 모든 미세 클러스터 간의 거리 계산을 피하기 위하여 범위 질의를 수행할 수 있다.

[0074] 일 실시예에서, 클러스터 집합 생성부(250)는 특정 미세 클러스터 세그먼트의 중심을 기준으로 미세 클러스터 세그먼트의 반지름과 임계 거리의 합을 반경으로 하는 범위 영역과 중첩되고 외부 미세 클러스터 세그먼트에 포함된 복수의 미세 클러스터 후보들을 결정할 수 있다. 외부 미세 클러스터 세그먼트는 특정 미세 클러스터 세그먼트와 서로 다른 미세 클러스터 세그먼트에 해당할 수 있고, 특정 미세 클러스터 세그먼트와 외부 미세 클러스터 세그먼트는 서로 중첩되지 않는다. 예를 들어, 클러스터 집합 생성부(250)는 미세 클러스터 세그먼트 MCS 의 중심인 $MCS.c$ 를 중심으로 하고, MCS 의 반지름 $MCS.r$ 과 임계 거리 T 의 합을 반경으로 하는 가상의 원과 겹쳐지는 모든 미세 클러스터들을 미세 클러스터 후보로서 검출할 수 있다.

[0075] 일 실시예에서, 클러스터 집합 생성부(250)는 복수의 미세 클러스터 후보들 각각에 대해 특정 미세 클러스터 세그먼트에 포함된 모든 미세 클러스터들 간의 거리가 임계 거리 이내인 경우 해당 미세 클러스터 후보가 포함된 외부 미세 클러스터 세그먼트를 특정 미세 클러스터 세그먼트와 연관된 부분 클러스터 집합에 추가할 수 있다. 클러스터 집합 생성부(250)는 특정 미세 클러스터 세그먼트와 직접적으로 연결된 미세 클러스터 세그먼트들을 검출하여 부분 클러스터 집합을 생성할 수 있다.

[0076] 여기에서, 부분 클러스터 집합은 특정 미세 클러스터 세그먼트와 직접적으로 연결된 미세 클러스터 세그먼트들을 포함하는 집합으로서 최종적인 클러스터 집합을 생성하기 위한 중간 집합에 해당할 수 있다. 예를 들어, 특정 미세 클러스터 세그먼트 MCS_a 와 MCS_b 가 직접적으로 연결되고 MCS_a 와 MCS_d 가 직접적으로 연결된 경우 범위 질의 키에 해당하는 MCS_a 에 관한 부분 클러스터 집합에 MCS_b 및 MCS_d 가 포함될 수 있다.

- [0077] 일 실시예에서, 클러스터 집합 생성부(250)는 동일한 미세 클러스터 세그먼트를 포함하고 있는 복수의 부분 클러스터 집합들을 하나로 통합하여 클러스터 집합을 생성할 수 있다. 클러스터 집합 생성부(250)는 모든 미세 클러스터 세그먼트에 대해 부분 클러스터 집합을 생성할 수 있고, 각 부분 클러스터 집합은 범위 질의의 키에 해당하는 미세 클러스터 세그먼트와 직접적으로 연결된 미세 클러스터 세그먼트만을 포함하고 있기 때문에 동일한 미세 클러스터 세그먼트를 포함하고 있는 복수의 부분 클러스터 집합들을 하나로 합병하여 최종적인 클러스터 집합을 생성할 수 있다. 따라서, 클러스터 집합 생성부(250)에 의해 생성되는 클러스터 집합은 직접 또는 간접적으로 연결된 모든 미세 클러스터 세그먼트를 포함하여 생성되고 서로 다른 클러스터 집합 간에는 동일한 미세 클러스터를 포함하지 않게된다.
- [0078] 제어부(270)는 데이터 클러스터링 장치(130)의 전체적인 동작을 제어하고, CF 트리 구축부(210), 미세 클러스터 세그먼트 생성부(230) 및 클러스터 집합 생성부(250) 간의 제어 흐름 또는 데이터 흐름을 관리할 수 있다.
- [0080] 도 3은 도 1에 있는 데이터 클러스터링 장치에서 수행되는 데이터 클러스터링 과정을 설명하는 순서도이다.
- [0081] 도 3을 참조하면, 데이터 클러스터링 장치(130)는 CF 트리 구축부(210)를 통해 데이터 집합에 포함된 복수의 데이터들을 기초로 미세 클러스터(Micro Cluster, MC)로 구성된 리프노드와 보조 클러스터(Sub Cluster, SC)로 구성된 인덱스노드를 포함하는 CF(Clustering Feature) 트리를 구축할 수 있다(단계 S310). 데이터 클러스터링 장치(130)는 미세 클러스터 세그먼트 생성부(230)를 통해 리프노드를 오른쪽 방향으로 순차탐색하면서 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트(Micro Cluster Segment, MCS)를 생성할 수 있다(단계 S330). 데이터 클러스터링 장치(130)는 클러스터 집합 생성부(250)를 통해 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성할 수 있다(단계 S350).
- [0083] 도 4는 도 1에 있는 데이터 클러스터링 장치에서 활용되는 CF 트리의 일 실시예를 설명하는 도면이다.
- [0084] 도 4를 참조하면, CF(Clustering Feature) 트리는 대표적인 요약 기반 클러스터링 기법에서 클러스터의 특징을 요약하기 위해 사용하는 CF 벡터들로 구성된 트리에 해당한다. CF 벡터는 클러스터 내 데이터들로 만들어진 값으로, 각 클러스터의 중심(c:center) 및 평균 반지름(ar:average radius) 등의 대표값으로 표현될 수 있다.
- [0085] 데이터 클러스터링 장치(130)에 의해 구축된 CF 트리의 리프 엔트리들은 데이터들의 가장 작은 클러스터인 미세 클러스터(microcluster)가 되며, 다른 엔트리와 마찬가지로, CF 벡터를 통해 표현될 수 있다. 각 미세 클러스터는 자신의 중심에서 임계값 T보다 가까운 데이터들을 묶은 소형 클러스터에 해당할 수 있다. 데이터 클러스터링 장치(130)는 클러스터링을 위해 입력된 데이터에 대해 CF 트리를 통해 가장 가까운 미세 클러스터를 찾을 수 있고, 해당 입력된 데이터와 미세 클러스터의 중심과의 거리가 T보다 작을 경우 해당 입력된 데이터를 해당 미세 클러스터에 반영할 수 있다.
- [0086] 또한, 반영된 데이터는 해당 미세 클러스터의 CF 벡터에 반영될 수 있고, 해당 미세 클러스터의 중심 및 평균 반지름 역시 입력 데이터에 따라 변경될 수 있다. 데이터 클러스터링 장치(130)는 해당 입력된 데이터와 미세 클러스터의 중심과의 거리가 T보다 클 경우 해당 입력된 데이터를 기준으로 새로운 미세 클러스터를 생성할 수 있다. 데이터 클러스터링 장치(130)에 의해 새롭게 생성된 미세 클러스터들은 CF 트리의 리프노드에 포함될 수 있고, 리프노드들은 보조 클러스터에 의해 표현될 수 있다. 결과적으로, CF 트리의 지역성에 따라 가까운 거리의 미세 클러스터들은 같이 모일 확률이 높을 수 있다.
- [0087] 도 4에서, CF 트리는 $node_1$ 을 루트노드로 하는 트리에 해당하고 루트노드는 엔트리들이 보조 클러스터인 인덱스노드에 해당할 수 있다. 또한, CF 트리는 $node_2, \dots, node_6$ 를 리프노드로 포함하고 각 리프노드는 엔트리들이 미세 클러스터로 구성될 수 있으며, 각 리프노드들은 next라는 링크를 통해 연결될 수 있다. 리프노드 $node_2, \dots, node_6$ 는 중심 c와 반지름 r을 포함할 수 있고, 이는 해당 리프노드에 속한 미세 클러스터들의 중심과 반지름을 통해 산출될 수 있으며, 해당 리프노드의 부모인 보조 클러스터에 관한 정보로서 사용될 수 있다.
- [0089] 도 5는 CF 트리 구축 과정에서 사용되는 미세 클러스터 간의 거리를 설명하는 예시도이다.
- [0090] 도 5를 참조하면, 데이터 클러스터링 장치(130)는 CF 트리를 구축하고 데이터 클러스터링을 수행하는 과정에서

미세 클러스터 간의 거리를 활용할 수 있고, 미세 클러스터 간의 거리(Inter-Microcluster Distance, IMD)는 미세 클러스터의 중심과 평균 반지름을 활용하여 산출할 수 있다. 보다 구체적으로, 데이터 클러스터링 장치(130)는 미세 클러스터 세그먼트 생성부(230)를 통해 두 미세 클러스터의 중심 간의 유클리디언 거리(Euclidian Distance, ED)에서 각 미세 클러스터의 평균 반지름을 뺀 값을 미세 클러스터 간의 거리로서 산출할 수 있다.

[0091] 도 5에서, 두 미세 클러스터 MC_1 및 MC_2 의 중심과 평균 반지름은 각각 $c(MC_1)$, $c(MC_2)$, $ar(MC_1)$, $ar(MC_2)$ 에 해당하고 두 미세 클러스터 간의 거리 $IMD(MC_1, MC_2)$ 는 두 미세 클러스터의 중심 간의 유클리디언 거리인 $ED(c(MC_1), c(MC_2))$ 에서 두 미세 클러스터의 평균 반지름의 합인 $ar(MC_1) + ar(MC_2)$ 을 뺀 값을 통해 산출할 수 있다.

[0093] 도 6은 CF 트리 구축 과정에서 사용되는 미세 클러스터 세그먼트를 설명하는 예시도이다.

[0094] 도 6을 참조하면, 데이터 클러스터링 장치(130)는 미세 클러스터 세그먼트 생성부(230)를 통해 CF 트리의 리프 노드를 오른쪽 방향으로 순차탐색하면서 미세 클러스터 간의 거리를 기초로 미세 클러스터 세그먼트를 생성할 수 있다. 보다 구체적으로, 미세 클러스터 세그먼트 생성부(230)는 리프노드의 next 포인터를 통해 모든 미세 클러스터를 순차탐색할 수 있고, 각 미세 클러스터에 대해 오른쪽 형제인 미세 클러스터와의 거리가 임계값 T보다 작은지 확인할 수 있다. 만약 T보다 작을 경우, 미세 클러스터 세그먼트 생성부(230)는 두 미세 클러스터를 동일한 미세 클러스터 세그먼트로 묶을 수 있다.

[0095] 도 6에서, 미세 클러스터 세그먼트 MCS_1 의 중심은 $MCS_1.c$ 에 해당하고, MCS_1 에 포함된 미세 클러스터들의 중심의 평균값으로 산출될 수 있다. 또한, MCS_1 의 반지름은 $MCS_1.r$ 에 해당하고, $MCS_1.c$ 에서 각 미세 클러스터까지의 거리와 각 미세 클러스터의 반지름의 합이 가장 큰 값으로 산출될 수 있다. $MCS_1.c$ 와 MC_2 의 중심 간의 거리와 MC_2 의 반지름의 합이 제일 크기 때문에 $MCS_1.r$ 은 $ED(MCS_1.c, c(MC_2)) + ar(MC_2)$ 에 해당할 수 있다.

[0097] 도 7은 CF 트리 구축 과정에서 사용되는 미세 클러스터 세그먼트 간의 연결 관계를 설명하는 예시도이다.

[0098] 도 7을 참조하면, 미세 클러스터 세그먼트 간의 연결관계가 존재할 수 있다. 보다 구체적으로, 미세 클러스터 세그먼트 간의 모든 쌍에 대한 IMD의 값이 T보다 작을 경우 두 미세 클러스터 세그먼트가 연결되었다고 정의할 수 있다. 또한, 미세 클러스터 세그먼트 간의 연결관계는 직접 연결 또는 간접 연결로 분류할 수 있다. 도 7에서, MCS_2 의 경우 MCS_1 및 MCS_6 과 직접 연결되어 있고, 나머지 MCS들과는 간접적으로 연결되어 있다.

[0100] 도 8은 도 1에 있는 데이터 클러스터링 장치에서 미세 클러스터 세그먼트에 대한 범위 질의를 수행하는 과정을 설명하는 예시도이다.

[0101] 도 8을 참조하면, 데이터 클러스터링 장치(130)는 클러스터 집합 생성부(250)를 통해 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성할 수 있다. 데이터 클러스터링 장치(130)는 모든 미세 클러스터 간의 거리 계산을 피하기 위해서, CF 트리를 이용한 범위 질의를 수행할 수 있으며, 이를 위해서 CF 트리 구축부(210)를 통해 CF 트리를 구축한 이후 각 보조 클러스터에 대해 반지름을 산출하는 과정을 수행할 수 있다.

[0102] CF 트리의 노드의 엔트리는 보조 클러스터 또는 미세 클러스터의 형태에 해당할 수 있다. 도 4에서, 보조 클러스터 SC_1 은 미세 클러스터 $MC_1 \sim MC_5$ 를 자식으로 가지고 있다. CF 트리 구축부(210)는 자식이 미세 클러스터인 경우 해당 미세 클러스터의 평균 반지름을 사용하고, 이외의 경우 반지름을 사용할 수 있다. 즉, CF 트리 구축부(210)는 자식이 보조 클러스터인 경우 다시 자식의 자식을 찾아가 반지름을 구할 수 있고, 이를 자식이 미세 클러스터일 때까지 재귀적으로 반복 수행할 수 있다. 따라서, CF 트리 구축부(210)는 전체 트리를 한번 탐색하여 트리 내 전체 보조 클러스터들에 대한 반지름을 구할 수 있다. 이렇게 구해진 반지름은 자신의 모든 자손에 해당하는 미세 클러스터들을 포함하는 원의 형태가 되며, 형태는 도 6의 미세 클러스터 세그먼트와 비슷한 형태가 될 수 있다.

[0103] 데이터 클러스터링 장치(130)는 클러스터 집합 생성부(250)를 통해 모든 미세 클러스터 세그먼트에 대해 범위 질의를 수행할 수 있다. 클러스터 집합 생성부(250)는 미세 클러스터 세그먼트 MCS_a 의 중심인 $MCS_a.c$ 를 중심으로

하고, MCS_a의 반지름 MCS_a.r과 임계값 T의 합을 반경으로 하는 가상의 원과 겹쳐지는 모든 미세 클러스터들을 후보군으로 뽑을 수 있다. 도 8에서, 미세 클러스터 MC₅, MC₇ 및 MC₁₀이 미세 클러스터 후보군으로 결정될 수 있다. 여기에서, 미세 클러스터 세그먼트 MCS_i와 미세 클러스터 MC_y에 대해 미세 클러스터 세그먼트 MCS_i와 미세 클러스터 MC_y 간의 거리(Inter-Cluster Distance, ICD)는 다음과 같이 정의될 수 있다.

$$ICD_{i,y} = \min_{MC \in MCS_i} IMD(MC, MC_y)$$

[0104]

[0106]

이후, 클러스터 집합 생성부(250)는 미세 클러스터 세그먼트 MCS_a의 모든 미세 클러스터들(MC₁, MC₂, MC₃)과 후보군 미세 클러스터와의 IMD가 T보다 작은지 확인할 수 있다. IMD가 T보다 작을 경우, 겹쳐진 미세 클러스터를 가지고 있는 미세 클러스터 세그먼트는 검색의 키인 MCS_a와 연결될 수 있다. 도 8에서, MCS_a(MC₂ 보유)와 MCS_c(MC₇ 보유)가 연결될 수 있고, 또한 MCS_a(MC₁ 보유)와 MCS_d(MC₁₀ 보유)가 연결될 수 있다. 클러스터 집합 생성부(250)는 이렇게 연결된 미세 클러스터 세그먼트들을 하나의 집합으로 병합하여 부분 클러스터 집합을 생성할 수 있다.

[0107]

또한, 부분 클러스터 집합은 범위 질의의 키에 해당하는 미세 클러스터 세그먼트와 직접적으로 연결된 미세 클러스터 세그먼트들 만을 포함하고 있으므로, 클러스터 집합 생성부(250)는, 도 7에서, MCS₂와 MCS₅의 간접적인 연결과 같은 간접 연결을 찾기 위한 합병 작업을 수행할 수 있다. 클러스터 집합 생성부(250)는 모든 부분 클러스터 집합에 대해 동일한 미세 클러스터 세그먼트를 공유하는 집합들을 선별하여 하나로 병합하는 과정을 수행함으로써 최종적인 클러스터 집합을 생성할 수 있다.

[0109]

도 9는 도 1에 있는 데이터 클러스터링 장치에서 미세 클러스터 세그먼트에 대해 범위 질의를 수행하여 클러스터 집합을 생성하는 과정의 알고리즘을 설명하는 도면이다.

[0110]

도 9를 참조하면, 데이터 클러스터링 장치(130)는 클러스터 집합 생성부(250)를 통해 CF 트리를 활용하여 범위 질의 기반의 클러스터링 기법을 수행할 수 있다. 특히, 클러스터 집합 생성부(250)는 도 9의 알고리즘을 사용하여 미세 클러스터 세그먼트에 대해 범위 질의를 수행 및 클러스터링을 수행할 수 있다. 도 9에서, 입력된 데이터 중 미세 클러스터 세그먼트 MCS의 집합 D는 미세 클러스터 세그먼트 생성부(230)에 의해 생성된 미세 클러스터 세그먼트들에 관한 집합에 해당할 수 있고, RQ는 범위 질의에 해당한다. 범위 질의 RQ는 질의의 키로 입력된 미세 클러스터 세그먼트 MCS와 직접 연결된 미세 클러스터 세그먼트들의 집합, 즉 부분 클러스터 집합을 결과로서 제공할 수 있다.

[0112]

상기에서는 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

[0114]

100: CF 트리를 활용한 범위 질의 기반의 데이터 클러스터링 시스템

110: 사용자 단말 130: 데이터 클러스터링 장치

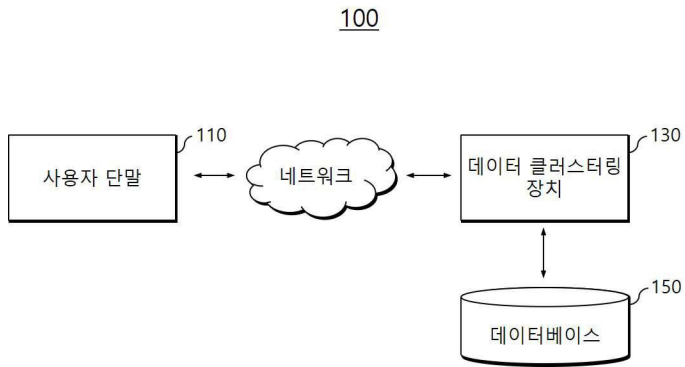
150: 데이터베이스

210: CF 트리 구축부 230: 미세 클러스터 세그먼트 생성부

250: 클러스터 집합 생성부 270: 제어부

도면

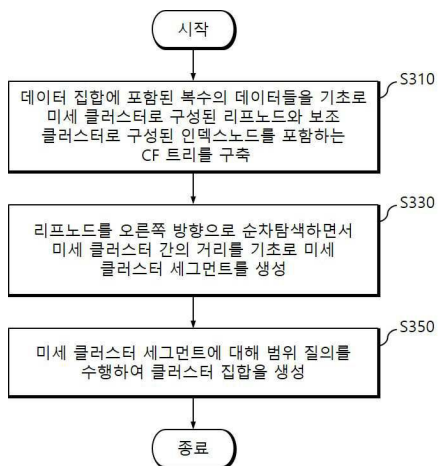
도면1



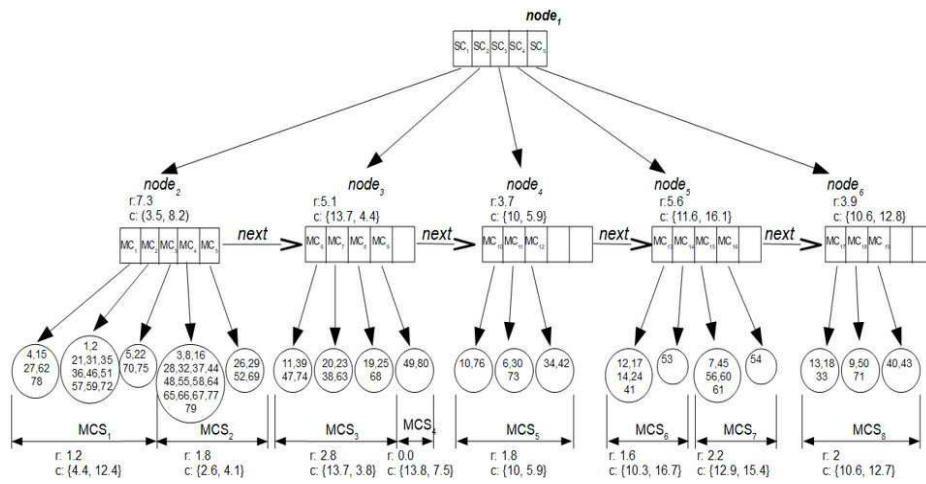
도면2



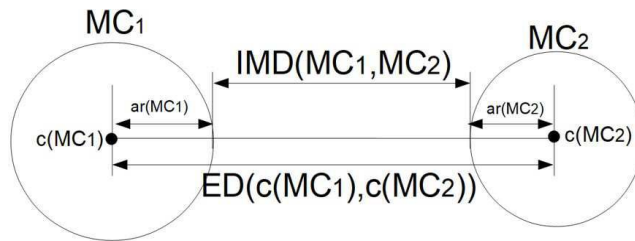
도면3



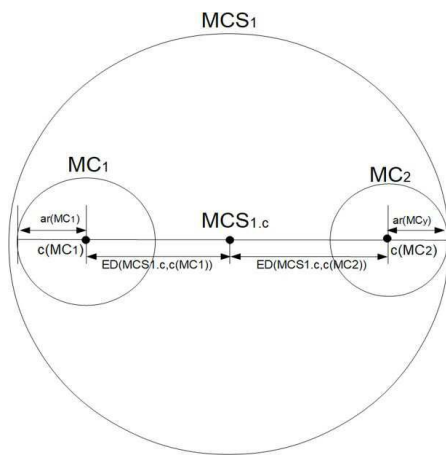
도면4



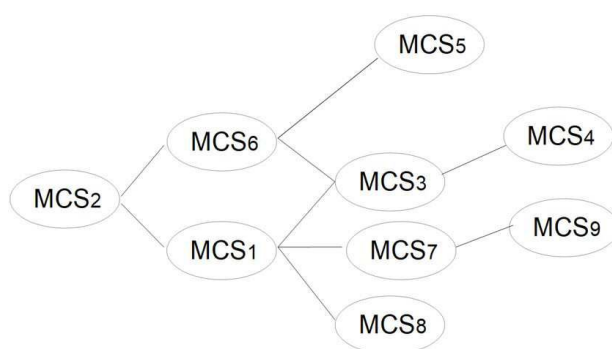
도면5



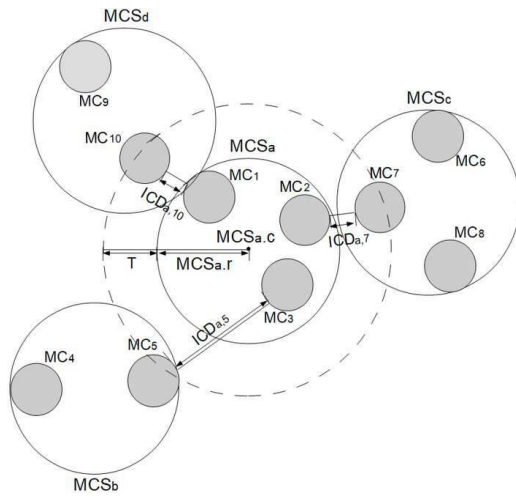
도면6



도면7



도면8



도면9

```

algorithm Refinement( $\Omega$ , tree,  $\mathcal{D}$ , T)
input: A label set  $\Omega$  of microclusters, CF-tree tree, A
set  $\mathcal{D}$  of microcluster segments, Threshold value T
output: A set  $\mathcal{R}$  of final clusters

for  $MCS_a \in \mathcal{D}$  do
     $\psi_x := MCS_a \cup RQ(\Omega, tree, MCS_a, T)$ ;
     $\Psi := \Psi \cup \psi_x$ ;
idx := 1;
while  $|\Psi| > 0$  do
    if  $\psi_a \in \Psi$  then
         $R_{idx} := \psi_a, \Psi := \Psi \setminus \psi_a$ ;
        for  $MCS_b \in R_{idx}$  do
            if  $\psi_b \in \Psi$  then
                 $R_{idx} := R_{idx} \cup \psi_b, \Psi := \Psi \setminus \psi_b$ ;
         $\mathcal{R} := \mathcal{R} \cup R_{idx}$ ;
        idx := idx + 1;
    
```