(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년11월09일
(11) 등록번호 10-1917501
(24) 등록일자 2018년11월05일

(51) 국제특허분류(Int. Cl.)
G06Q 30/02 (2012.01) *G06F 17/30* (2006.01)
G06Q 50/00 (2018.01)
(52) CPC특허분류
G06Q 30/02 (2013.01)
G06F 17/30312 (2013.01)
(21) 출원번호 10-2017-0030733
(22) 출원일자 2017년03월10일
심사청구일자 2017년03월10일
(65) 공개번호 10-2018-0018263
(43) 공개일자 2018년02월21일
(30) 우선권주장
1020160103059 2016년08월12일 대한민국(KR)
(56) 선행기술조사문헌
US20150032759 A1
KR1020090065317 A
시그니처 트리를 사용한 의미적 유사성 검색 기법
온톨로지 기반의 개인화된 여행 추천 시스템의 구현

(73) 특허권자
명지대학교 산학협력단
경기도 용인시 처인구 명지로 116 (남동, 명지대학교)
(72) 발명자
한승철
경기도 과천시 광창1로 7 (과천동)
한동호
경기도 과천시 광창1로 7 (과천동)
남기원
서울특별시 서초구 신반포로15길 33, 401호(반포동,반포파크빌)
(74) 대리인
송인호, 최관락

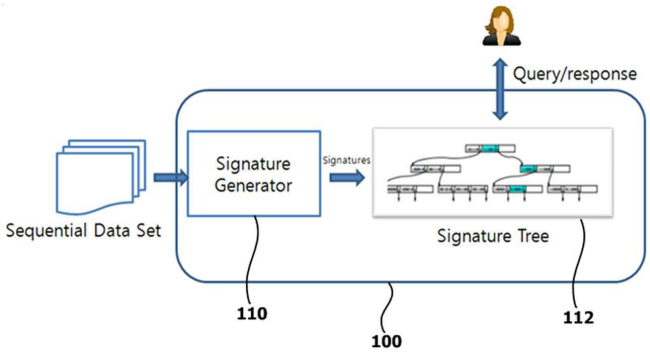
전체 청구항 수 : 총 6 항 심사관 : 홍경희

(54) 발명의 명칭 **시그니처 트리를 이용한 시퀀셜 데이터 클러스터링 방법 및 시스템**

(57) 요 약

시그니처 트리를 이용하여 유사 데이터 검색을 효율적으로 수행할 수 있고 특정 액션의 추천 및 예측 기능을 제공하는 시퀀셜 데이터 클러스터링 방법 및 시스템이 개시된다. 상기 시스템은 시퀀셜 데이터(Sequential data)를 구성하는 원소들의 시그니처들(Signature)을 생성하는 시그니처 생성부 및 상기 생성된 시그니처들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하는 시그니처 트리부를 포함한다. 여기서, 상기 시그니처 트리의 최하위 계층은 상기 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함한다.

대 표 도



(52) CPC특허분류
G06Q 30/0201 (2013.01)
G06Q 30/0202 (2013.01)
G06Q 50/01 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2015-0409
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 기본연구지원사업
연구과제명 뇌인지 과학과 ICT 기술을 융합한 차세대 보안 시스템
기 여 율 50/100
주관기관 명지대학교 산학협력단
연구기간 2015.11.01 ~ 2016.10.31
이 발명을 지원한 국가연구개발사업
과제고유번호 2017R1A2B1005285
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 이공분야기초연구사업>중견연구자지원사업
연구과제명 봇넷탐지를 위한 빅데이터 3D 시각화 시스템
기 여 율 50/100
주관기관 명지대학교
연구기간 2017.03.01 ~ 2018.02.28

명 세 서

청구범위

청구항 1

시퀀셜 데이터(Sequential data)를 구성하는 원소들의 시그니처들(Signature)을 생성하는 시그니처 생성부; 및

상기 생성된 시그니처들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하는 시그니처 트리부를 포함하되,

상기 시그니처 트리의 최하위 계층은 상기 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함하되,

상기 시그니처들은 0 또는 가중치가 부여된 양의 정수로 표현되는 원소들로 구성되고,

상기 시그니처 트리부는,

입력된 특정 시그니처와 유사한 시그니처를 포함하는 노드를 검색하고, 상기 검색된 노드가 상기 특정 시그니처를 포함할 때 디멘젼(dimension)을 초과하는 경우, 상기 검색된 노드의 시그니처들과 상기 특정 시그니처를 포함하는 그룹에서 유사도가 가장 먼 2개의 시그니처들을 각기 새롭게 생성된 2개의 노드들로 분리하여 배열하고, 상기 그룹의 나머지 시그니처들을 유사도에 따라 상기 새롭게 생성된 2개의 노드들로 분리하여 배열하고, 상기 특정 시그니처가 배열됨에 따라 상기 검색된 노드의 상위 노드의 원소를 해당 하위 계층 노드들의 모든 원소들의 논리합으로 가변하는 것을 특징으로 하는 시퀀셜 데이터 클러스터링 시스템.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서, 특정 하위 계층 노드에 해당하는 상위 노드의 일 원소는 상기 특정 하위 계층 노드의 모든 원소들의 논리합인 것을 특징으로 하는 시퀀셜 데이터 클러스터링 시스템.

청구항 5

제1항에 있어서, 상기 시퀀셜 데이터 클러스터링 시스템은 쇼핑몰 시스템, 마케팅 시스템, 금융 정보 분석 시스템, 고객 취향 분석 시스템, SNS 관계 분석 시스템, 패턴 인식 시스템, 멀티미디어 시스템 또는 보안 시스템에 적용되는 것을 특징으로 하는 시퀀셜 데이터 클러스터링 시스템.

청구항 6

제1항에 있어서,

특정 시그니처와 유사한 노드를 상기 시그니처 트리로부터 검색하는 검색부;

상기 검색 결과에 따라 상기 특정 시그니처의 사용자에게 특정 액션을 추천하는 추천부; 및

상기 검색 결과에 따라 상기 사용자의 특정 액션을 예측하는 예측부를 더 포함하는 것을 특징으로 하는 시퀀셜 데이터 클러스터링 시스템.

청구항 7

시퀀셜 데이터 클러스터링 시스템이 수행하는 시퀀셜 데이터 클러스터링 방법에 있어서,

시퀀셜 데이터(Sequential data)를 구성하는 원소들의 시그니처들(Signature)을 생성하는 단계; 및

상기 생성된 시그니처들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하는 단계를 포함하되,

상기 시그니처 트리의 최하위 계층은 상기 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함하고,

상기 시그니처들은 0 또는 가중치가 부여된 양의 정수로 표현되는 원소들로 구성되고,

상기 시그니처 트리를 생성하는 단계는,

특정 시그니처를 입력하는 단계;

상기 입력된 특정 시그니처와 유사한 시그니처를 포함하는 노드를 검색하는 단계; 및

상기 검색된 노드가 상기 특정 시그니처를 포함할 때 디멘젼(dimension)을 초과하는 경우, 상기 검색된 노드의 시그니처들과 상기 특정 시그니처를 2개의 노드들로 분리하여 배열하는 단계를 포함하되,

상기 특정 시그니처가 배열됨에 따라 상기 검색된 노드의 상위 노드의 원소가 해당 하위 계층 노드들의 모든 원소들의 논리합으로 가변되고,

상기 2개의 노드들로 분리하여 배열하는 단계는,

상기 검색된 노드의 시그니처들과 상기 특정 시그니처를 포함하는 그룹에서 유사도가 가장 먼 2개의 시그니처들을 각기 새롭게 생성된 2개의 노드들로 분리하여 배열하는 단계; 및

상기 그룹의 나머지 시그니처들을 유사도에 따라 상기 새롭게 생성된 2개의 노드들로 분리하여 배열하는 단계를 포함하는 것을 특징으로 하는 시퀀셜 데이터 클러스터링 방법.

청구항 8

제7항에 있어서, 특정 하위 계층 노드에 해당하는 상위 노드의 일 원소는 상기 특정 하위 계층 노드의 모든 원소들의 논리합인 것을 특징으로 하는 시퀀셜 데이터 클러스터링 방법.

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 시그니처 트리를 이용한 시퀀셜 데이터 클러스터링 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 종래에는 시퀀셜 데이터를 체계적으로 분류하는 기술이 없었다. 따라서, 유사 데이터 검색을 위해서는 데이터베이스 및 SQL을 사용하여 여러 단계의 질의를 반복적으로 수행하여야 했다.

[0003] 기존의 데이터베이스는 전체적인 유사도 분석이 어려웠으며, 특정 데이터와 유사한 데이터를 추출하기 위해서는 모든 데이터와 일일이 비교하여야 해서 연산량이 상당히 많을 수밖에 없다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) KR 2017-0010883 A

발명의 내용

해결하려는 과제

[0005] 본 발명은 시그니처 트리를 이용하여 유사 데이터 검색을 효율적으로 수행할 수 있고 특정 액션의 추천 및 예측 기능을 제공하는 시퀀셜 데이터 클러스터링 방법 및 시스템을 제공하는 것이다.

과제의 해결 수단

[0006] 상기한 바와 같은 목적을 달성하기 위하여, 본 발명의 일 실시예에 따른 시퀀셜 데이터 클러스터링 시스템은 시퀀셜 데이터(Sequential data)를 구성하는 원소들의 시그니처들(Signature)을 생성하는 시그니처 생성부; 및 상기 생성된 시그니처들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하는 시그니처 트리부를 포함한다. 여기서, 상기 시그니처 트리의 최하위 계층은 상기 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함한다.

[0007] 본 발명의 일 실시예에 따른 시퀀셜 데이터 클러스터링 방법은 시퀀셜 데이터(Sequential data)를 구성하는 원소들의 시그니처들(Signature)을 생성하는 단계; 및 상기 생성된 시그니처들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하는 단계를 포함한다. 여기서, 상기 시그니처 트리의 최하위 계층은 상기 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함한다.

[0008] 본 발명의 일 실시예에 따른 시그니처 트리는 루트 노드; 및 상기 루트 노드의 하위 계층으로 구성되는 복수의 하위 계층 노드들을 포함한다. 여기서, 상기 시그니처 트리의 최하위 계층은 시그니처들을 유사도에 따라 그룹핑함에 의해 형성되는 복수의 하위 계층 노드들을 포함한다.

발명의 효과

[0009] 본 발명에 따른 시퀀셜 데이터 클러스터링 방법 및 시스템은 시퀀셜 데이터에 의해 생성된 시그니처들을 계층적으로 클러스터링하여 시그니처 트리를 구성하므로, 시그니처 트리를 이용하여 유사 데이터 검색을 효율적으로 수행할 수 있고 사용자에게 특정 액션을 추천하거나 사용자의 특정 액션을 예측할 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일 실시예에 따른 시그니처 트리를 이용한 시퀀셜 데이터 클러스터링 시스템을 도시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 시퀀셜 데이터 클러스터링 방법을 도시한 순서도이다.

도 3 내지 도 9는 본 발명의 일 실시예에 따른 시그니처 트리 생성 과정을 도시한 도면들이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0014] 본 발명은 시그니처 트리(Signature tree)를 이용한 시퀀셜 데이터(Sequential data)의 클러스터링 방법 및 시스템에 관한 것으로서, 시그니처의 유사도에 따라 시그니처들을 계층적으로 클러스터링하여 시그니처 트리를 생성한다. 결과적으로, 상기 시스템은 상기 시그니처 트리를 이용하여 특정 데이터와 유사한 특성을 가진 데이터를 효율적으로 검색할 수 있고, 나아가 유사도에 따라 추천 및 예측 기능도 제공할 수 있다.

[0015] 예를 들어, 쇼핑몰 고객의 상품 주문 내역을 물건별 또는 카테고리별로 분류하여 시그니처들을 생성하고 상기 생성된 시그니처들을 유사도에 따라 클러스터링하여 시그니처 트리로 구성하면, 고객들의 쇼핑 습관을 용이하게 분석할 수 있고, 유사한 쇼핑 습관을 가지는 다른 고객들이 구매하는 상품을 특정 고객에게 추천하거나 특정 고객의 특정 상품의 구매를 예측할 수 있다.

[0016] 일 실시예에 따르면, 본 발명의 클러스터링 시스템은 유사도에 따라 시그니처들을 계층적으로 클러스터링하여 시그니처들을 그룹핑하여 분류할 수 있다.

[0017] 종래 기술에서는, 데이터베이스와 SQL을 사용하여 유사 데이터를 검색하기 위해서는 여러 단계의 질의를 반복해야 하였고 유사도 분석이 어려웠으나, 본 발명의 시스템은 유사도에 따라 시그니처들을 계층적으로 클러스터링함에 의해 생성된 시그니처 트리를 이용하므로, 여러 단계의 질의 과정이 필요없고 유사도 분석이 용이하다.

[0018] 또한, 종래 기술에서는 특정 데이터와 유사한 데이터를 추출하기 위해서는 모든 데이터와 상기 특정 데이터를 일일이 비교하여야 해서 연산량이 상당히 많았으나, 본 발명의 시스템은 유사도에 따라 시그니처들을 계층적으로 분류한 시그니처 트리를 사용하므로 특정 데이터와 유사한 데이터를 포함하는 노드(그룹)를 용이하게 검색할 수 있다. 결과적으로, 본 발명의 시스템에서는 종래 기술에 비하여 연산량이 상당히 감소할 수 있다.

[0019] 본 발명의 시스템은 쇼핑몰 시스템뿐만 아니라 마케팅 시스템, 금융 정보 분석 시스템, 고객 취향 분석 시스템, SNS 관계 분석 시스템, 패턴 인식 시스템, 멀티미디어 시스템, 보안 시스템 등 다양한 분야에 확장 적용될 수 있다.

[0021] 이하, 본 발명의 다양한 실시예들을 첨부된 도면을 참조하여 상술하겠다. 다만, 설명의 편의를 위하여 본 발명의 시스템을 쇼핑몰 시스템으로 가정하겠다. 한편, 이하에서 사용되는 클러스터링(Clustering)은 집합 S의 원소들을 유사한 성질을 가지는 원소들을 갖는 원소들끼리 가까이 배치/저정하는 것을 의미한다.

[0022] 도 1은 본 발명의 일 실시예에 따른 시그니처 트리를 이용한 시퀀셜 데이터 클러스터링 시스템을 도시한 도면이고, 도 2는 본 발명의 일 실시예에 따른 시퀀셜 데이터 클러스터링 방법을 도시한 순서도이다.

[0023] 도 1을 참조하면, 본 실시예의 시퀀셜 데이터 클러스터링 시스템(100)은 시그니처 생성부(110) 및 시그니처 트리부(112)를 포함할 수 있다. 또한, 시퀀셜 데이터 클러스터링 시스템(100)은 도시하지는 않았지만 검색부, 추천부 및 예측부를 더 포함할 수 있다.

[0024] 시그니처 생성부(110)는 하기 수학식 1로 표현되는 시퀀셜 데이터들을 이용하여 적어도 하나의 시그니처를 생성할 수 있다.

수학식 1

[0025] $$\text{set } S = \{C_1, C_2, C_3, \ldots, C_n\},\ C_i=\{x_{i1}, x_{i2}, \ldots, x_{im}\},\ 1 \le i \le n$$

[0026] 여기서, x_{ij}는 특징, 속성, 디멘젼(dimension) 또는 값일 수 있다.

[0027] 예를 들어, 12명의 고객이 있고 상품 종류가 6가지 있다고 하자. C가 고객이고, x가 고객의 상품 주문 내역이며, 1이 상품의 구매를 나타내고, 0의 상품의 비구매를 나타낸다고 가정하면, 시퀀셜 데이터는 하기 수학식 2와 같이 표현될 수 있다. 이 경우, 시그니처는 하기 수학식 3과 같이 표현될 수 있다. 즉, 시퀀셜 데이터는 상품을 주문한 고객에 대한 데이터이며, 시그니처는 고객별 주문 상품 목록일 수 있다.

수학식 2

[0028] $S=\{C_1, C_2, C_3, \ldots\ldots, C_{12}\}$, $C_1=\{0,0,1,0,1,0\}$, $C_2=\{0,1,1,0,0,0\}, \ldots\ldots,$
$C_{12}=\{0,0,0,1,1,1\}$

수학식 3

[0029]
$$sig(C_1) = 001010$$
$$sig(C_2) = 011000$$
$$\ldots\ldots$$
$$sig(C_{12}) = 000111$$

[0030] 한편, 시그니처의 원소는 이진수로서 "0"과 "1"로만 표현되었지만, 정수로 표현될 수도 있다.

[0031] 또한, 시그니처의 원소는 가중치가 부여되어 표현될 수 있다. 예를 들어, 특정 상품을 3개 산 경우, "3"으로 표현될 수도 있다.

[0032] 시그니처 트리부(112)는 유사도에 따라 시그니처들을 계층적으로 클러스터링하여 시그니처 트리를 생성한다. 여기서, 시그니처 트리의 최하위 계층은 유사도에 따라 그룹화된 시그니처들을 포함하는 복수의 노드들로 구성될 수 있다. 시그니처 트리 생성 방법에 대한 자세한 설명은 후술하겠다.

[0033] 상기 검색부는 특정 데이터(시그니처)와 유사한 적어도 하나의 데이터(시그니처)를 상기 시그니처 트리에서 검색할 수 있다. 예를 들어, 상기 검색부는 특정 고객의 상품 구매 취향과 유사한 취향을 가지는 고객들을 상기 시그니처 트리에서 검색할 수 있다.

[0034] 상기 추천부는 특정 시그니처와 유사한 시그니처를 포함하는 그룹을 분석하여 상기 특정 시그니터의 사용자에게 특정 액션, 예를 들어 특정 상품의 구매 등을추천할 수 있다.

[0035] 상기 예측부는 특정 시그니처와 유사한 시그니처를 포함하는 그룹을 분석하여 상기 특정 시그니처에 해당하는 사용자의 특정 액션, 예를 들어 특정 상품의 구매 등을 예측할 수 있다.

[0036] 정리하면, 본 발명의 시퀀셜 데이터 클러스터링 시스템은 시퀀셜 데이터들을 유사도에 따라 계층적으로 클러스터링하여 시그니처 트리를 생성하고, 상기 생성된 시그니처 트리를 이용하여 특정 시그니처와 유사한 시그니처 또는 상기 시그니처를 포함하는 노드(그룹)를 검색하거나 검색 결과에 따라 특정 액션을 추천하거나 사용자의 특정 액션을 예측할 수 있다.

[0038] 이하, 이러한 시퀀셜 데이터 클러스터링 시스템에서의 동작을 도 2를 참조하여 살펴보겠다.

[0039] 도 2를 참조하면, 시그니처 생성부(110)는 시퀀셜 데이터들을 이용하여 시그니처를 생성한다(S200).

[0040] 이어서, 시그니처 트리부(112)는 상기 생성된 시그니처들을 유사도에 따라 클러스터링하여 시그니처 트리를 생성한다(S202). 이 경우, 시그니처들은 유사도에 따라 그룹핑된다. 여기서, 그룹핑이란 유사한 시그니처들이 하나의 노드 혹은 근처의 노드에 저장됨을 의미한다.

[0041] 계속하여, 검색부는 사용자의 요청에 따라 특정 시그니처와 유사한 시그니처를 검색하며, 즉 유사 특성 데이터를 검색한다(S204). 여기서, 상기 검색부는 상기 특정 시그니처와 유사한 시그니처를 포함한 그룹을 검색할 수도 있다.

[0042] 이어서, 추천부는 상기 검색된 그룹의 시그니처들을 분석하여 사용자에게 특정 액션을 추천할 수 있고, 예측부는 상기 검색된 그룹의 시그니처들을 분석하여 사용자의 다음 액션을 예측할 수 있다(S206).

[0044] 이하, 시그니처 트리 생성을 위한 알고리즘을 살펴보겠다.

[0045] Function insert(Node n, Entry e): Node splitnode

[0046] if n is a leaf node then // n이 리프노드 경우

[0047] insert e into n;

[0048] if n overflows then // n에 저장된 signature수가 d(차원)개를 초과 경우

[0049] newnode = split(n); // 새로운 노드를 만들고, signature들을 n과 새로운 노드에 분배

[0050] return newnode;

[0051] // if n overflows

[0052] else // if n is not leaf node // n이 루트 혹은 내부노드 경우

[0053] next = choose_subtree(n,e); // e와 가장 유사한 signature의 subtree 선택

[0054] splitnode = insert(next,e); //subtree에 재귀적으로 insert 호출

[0055] if splitnode ≠ null then // subtree에서 splitnode가 발생 경우

[0056] insert new entry pointing to splitnode in n;

[0057] if n overflows then // n에 d개 초과의 signature가 발생경우

[0058] newnode = split(n); //n을 2개의 노드로 분리하고 signature들을 분배

[0059] return newnode;

[0060] //if splitnode ≠ null

[0061] return null;

[0062] }

[0064] Function split(Node n)

[0065] n의 signature 중 가장 다른 signature A,B를 고르고 // tie발생하면 임의로 선택 나머지 signature들은 A,B 중 가까운 쪽으로 순서대로 배치한다.

[0066] 각각의 A와 B를 중심으로 묶인 signature들의 논리합을 구해서 묶음A의 signature와 묶음B의 signature를 노드 n의 signature로 하고 자식노드로 한다.

[0068] 이하, 이러한 알고리즘에 따른 시그니처 트리 생성 방법을 실제 예를 들어 상술하겠다.

[0069] 도 3 내지 도 9는 본 발명의 일 실시예에 따른 시그니처 트리 생성 과정을 도시한 도면들이다.

[0070] 이하, 설명의 편의를 위하여 쇼핑몰에서의 고객 주문에 따른 시퀀셜 데이터 및 시그니처들을 하기 표 1로서 정의하겠다. overflow에 해당하는 디멘젼(d, 차원)은 3으로 하겠다. 즉, 노드(그룹)의 원소의 개수가 3개를 초과하면 원소들을 별도의 그룹으로 분리시키도록 한다.

표 1

[0071]

	item1	item2	item3	item4	item5	item6
고객1	1	0	1	0	0	0
고객2	1	1	1	0	0	0
고객3	0	0	1	0	1	1
고객4	0	0	1	0	0	1
고객5	0	0	1	1	1	0
고객6	0	0	0	1	1	1
고객7	1	1	0	0	0	0
고객8	0	1	1	0	0	0
고객9	0	0	1	0	1	0
고객10	0	0	1	1	1	0
고객11	0	0	0	1	0	1

[0072] 표 1을 참조하면, 시퀀셜 데이터의 원소(고객)별 시그니처들이 표시되어 있다.

[0073] 우선, 도 3에 도시된 바와 같이 처음의 시그니처들(고객1 및 고객2)을 하나의 노드로 구성한다. 결과적으로, {101000, 111000}이 하나의 노드로 형성된다.

[0074] 이어서, 다음의 시그니처(고객3)가 입력되면 {101000, 111000, 001011}이 하나의 노드로 형성되며, 다음 시그니처(고객4)가 입력되면 하나의 노드에 포함된 시그니처들의 수가 d(=3)를 초과하기 때문에, split 함수가 호출되며, 즉 노드 분할 과정이 수행된다.

[0075] 일 실시예에 따르면, 고객1 내지 고객4에 해당하는 4개의 시그니처들 중 가장 유사도가 먼 2개의 시그니처들{111000, 001001}을 선택하고, 선택된 시그니처들을 새로 생성된 2개의 노드들로 분리시키며, 다른 2개의 시그니처들{101000, 001011}을 유사도에 따라 노드들에 삽입한다.

[0076] 구체적으로는, 시그니처{101000}는 시그니처{001001}보다 시그니처{111000}와 더 유사하기 때문에 시그니처{111000}의 노드로 삽입되고, 시그니처{001011}는 시그니처{111000}보다 시그니처{001001}와 더 유사하기 때문에 시그니처{001001}의 노드로 삽입된다. 결과적으로, 시그니처들{111000, 101000}을 포함하는 노드와 시그니처들{001001, 001011}을 포함하는 노드가 생성된다.

[0077] 이 때, 2개의 노드들은 루트 노드인 상위 노드의 제 1 하위 계층 노드로 정의되되, 상기 상위 노드의 원소들은 각 제 1 하위 계층 노드들의 원소들(시그니처들)의 논리합에 해당한다. 결과적으로, 루트 노드는 도 4에 도시된 바와 같이 제 1 하위 계층 노드의 시그니처들{111000, 101000}의 논리합에 해당하는 원소{111000}와 제 1 하위 계층 노드의 시그니처들{001001, 001011}의 논리합에 해당하는 원소{001011}를 포함할 수 있다.

[0078] 즉, 고객1 내지 고객4의 시그니처들은 도 4에 도시된 바와 같이 2개의 계층들로 클러스터링된다.

[0079] 계속하여, 고객5의 시그니처{001110}가 제 1 하위 계층 노드들 중 유사도가 높은 시그니처가 존재하는 노드로 삽입된다. 고객5의 시그니처{001110}는 제 1 하위 계층 노드의 시그니처들{001001, 001011}과 더 유사하므로, 시그니처들{001001, 001011}을 포함하는 제 1 하위 계층 노드로 삽입된다. 결과적으로, 해당 제 1 하위 계층 노드는 원소들{001001, 001011, 001110}을 포함하게 된다.

[0080] 이 경우, choose_subtree가 호출되어 루트 노드의 시그니처{001011}가 반환되고, 시그니처들{001001, 001011, 001110}의 논리합에 해당하는 원소{001111}가 {001011}의 subtree에 삽입된다. 결과적으로, 도 5에 도시된 바와 같은 트리가 완성된다.

[0081] 이어서, 고객6의 시그니처{000111}가 제 1 하위 계층 노드들 중 유사도가 높은 시그니처가 존재하는 노드로 삽입된다. 고객6의 시그니처{000111}는 제 1 하위 계층 노드의 시그니처들{001001, 001011, 001110}과 더 유사하므로, 시그니처들{001001, 001011, 001110}을 포함하는 제 1 하위 계층 노드로 삽입되어야 하나, overflow가 발생하므로 노드 분할 과정이 수행된다.

[0082] 여기서, 노드 분할 과정은 시그니처들{001001, 001011, 001110, 000111} 중 가장 유사도가 먼 시그니처들{001110, 001001}을 별개의 노드로 분리시키고 나머지 시그니처들{001011, 000111}을 유사도에 따라 해당 노드들로 분리하여 배열시킨다. 결과적으로, 시그니처들{001110, 000111}을 포함하는 제 1 하위 계층 노드와 시그니처들{001001, 001011}을 포함하는 제 1 하위 계층 노드가 생성된다.

[0083] 전체적으로 보면, 3개의 하위 계층 노드들이 도 6에 도시된 바와 같이 생성되며, 이에 따라 루트 노드의 원소들이 가변된다. 상기 루트 노드의 원소들은 각 하위 계층 노드들의 모든 원소들의 논리합이다.

[0084] 계속하여, 고객7의 시그니처{110000}, 고객8의 시그니처{011000} 및 고객9의 시그니처{001110}가 제 1 하위 계층 노드들에 유사도에 따라 순차적으로 배열된다. 이 때, 제 1 하위 계층 노드들에 overflow가 발생되지 않으므로, 노드 분할 과정은 수행되지 않는다.

[0085] 고객7의 시그니처{110000}, 고객8의 시그니처{011000} 및 고객9의 시그니처{001110}가 배열됨에 따라, 도 7에 도시된 바와 같이 시그니처들{111000, 101000, 110000}을 포함하는 제 1 하위 계층 노드, 시그니처들{001110, 000111, 011000}을 포함하는 제 1 하위 계층 노드 및 시그니처들{001001, 001011, 001010}을 포함하는 제 1 하위 계층 노드가 형성된다. 이 경우, 루트 노드의 원소들은 각 제 1 하위 계층 노드들의 시그니처들의 논리합으로 변경된다.

[0086] 이어서, 고객10의 시그니처{001110}를 시그니처 트리에 삽입시, 고객10의 시그니처{001110}가 시그니처들{001110, 000111, 01100}을 포함하는 제 1 하위 계층 노드와 가장 유사하여 해당 제 1 하위 계층 노드로 삽입되

어야 하는데, overflow가 발생되므로 2개의 노드들로 분리된다. 결과적으로, 시그니처들{001110, 0001111}을 포함하는 노드와 시그니처들{011000, 001110}을 포함하는 노드가 생성된다.

[0087] 이 경우, 4개의 제 1 하위 계층 노드들이 존재하게 되고, 그 결과 루트 노드에 overflow가 발생되게 되므로, 시그니처 트리가 도 8에 도시된 바와 같이 세개의계층들로 변경된다. 즉, 루트 노드, 제 1 하위 계층 노드 및 제 2 하위 계층 노드를 포함하는 시그니처 트리로 변경된다. 결과적으로, 제 1 하위 계층 노드들이 제 2 하위 계층 노드들로 변경되고, 제 2 하위 계층 노드들의 원소들의 논리합에 해당하는 원소들로 제 1 하위 계층 노드들이 형성되며, 제 1 하위 계층 노드들의 원소들의 논리합으로 루트 노드가 구성된다. 결과적으로, 도 8에 도시된 바와 같은 시그니처 트리가 형성된다.

[0088] 계속하여, 고객11의 시그니처{000101}가 유사도에 따라 제 2 하위 계층 노드들로 삽입되, 그 결과 도 9에 도시된 바와 같은 시그니처 트리가 구성된다.

[0089] 즉, 최종적으로 고객1 내지 고객11의 시그니처들은 도 9에 도시된 바와 같은 시그니처 트리를 구성한다.

[0090] 도 9를 참조하면, 고객1 내지 고객11의 시그니처들은 최하위 계층에서 5개의 그룹들로 구분된다.

[0091] 여기서, 그룹1과 그룹2에 속한 원소들(시그니처들)은 그룹3, 그룹4 및 그룹5에 속한 원소들보다 유사하다. 이는 그룹1과 그룹2가 공통 조상을 가지기 때문이다. 즉, 공통 조상을 어느 계층에서 갖느냐에 따라서 유사 정도를 비교할 수 있다.

[0092] 한편, 고객X의 시그니처가 {101010}이면, 고객X의 시그니처가 그룹1과 가장 유사하다는 것을 시그니처 트리 검색을 통하여 확인할 수 있다. 구체적으로는, 시스템은 고객X의 시그니처와 유사한 시그니처를 포함하는 그룹을 검색하기 위하여, 루트 노드로부터 하위 계층으로 순차적으로 검색한다.

[0093] 이 경우, 상기 시그니처 트리가 계층적으로 클러스터링되어 있어서 유사 노드들을 따라서 검색이 진행되므로, 모든 데이터와 특정 데이터를 유사 비교하는 종래 기술에 비하여 검색 시간이 상당히 짧아지게 된다. 구체적으로는, 알고리즘의 time complexity는 O(logn)에 불과하며, 따라서 상기 시스템은 검색 측면에서 기존 기술에 비하여 상당히 효율적이다.

[0094] 또한, 고객X의 시그니처가 {101010}이면, 고객X가 그룹1의 고객들과 유사한 취향을 가짐을 시그니처 트리를 통하여 확인할 수 있다. 그룹1의 고객들의 구입 아이템들을 살펴보면, 고객X가 구입하지 않은 item2를 그룹1의 고객들이 구입했음을 확인할 수 있다. 따라서, 시스템은 고객X에게 item2를 추천할 수도 있고, 유사 취향을 가지므로 고객X가 item2를 구입할 것이라 예측할 수도 있다.

[0095] 정리하면, 본 발명의 시퀀셜 데이터 클러스터링 시스템은 시그니처들을 유사도에 따라 계층적으로 클러스터링한 시그니처 트리를 생성하고, 상기 생성된 시그니처 트리를 통하여 유사 데이터 검색, 특정 액션의 추천 또는 예측 기능을 수행할 수 있다.

[0097] 한편, 전술된 실시예의 구성 요소는 프로세스적인 관점에서 용이하게 파악될 수 있다. 즉, 각각의 구성 요소는 각각의 프로세스로 파악될 수 있다. 또한 전술된 실시예의 프로세스는 장치의 구성 요소 관점에서 용이하게 파악될 수 있다.

[0098] 또한 앞서 설명한 기술적 내용들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예들을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 하드웨어 장치는 실시예들의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

산업상 이용가능성

[0099] 상기한 본 발명의 실시예는 예시의 목적을 위해 개시된 것이고, 본 발명에 대한 통상의 지식을 가지는 당업자라

면 본 발명의 사상과 범위 안에서 다양한 수정, 변경, 부가가 가능할 것이며, 이러한 수정, 변경 및 부가는 하기의 특허청구범위에 속하는 것으로 보아야 할 것이다.

부호의 설명

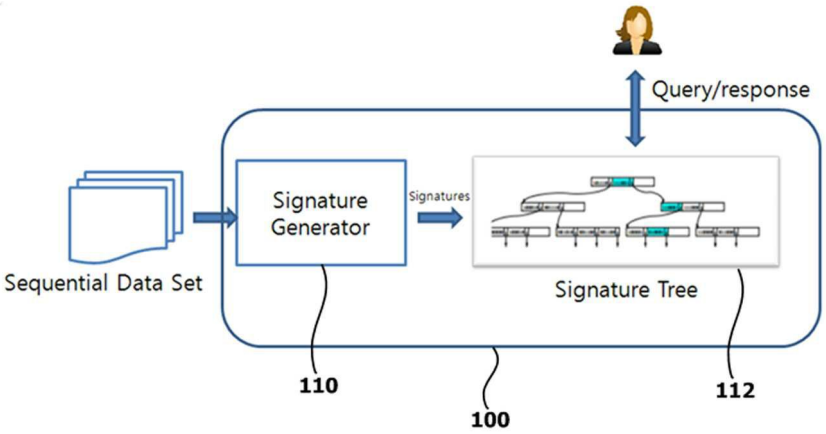
[0100] 100 : 시퀀셜 데이터 클러스터링 시스템

110 : 시그니처 생성부

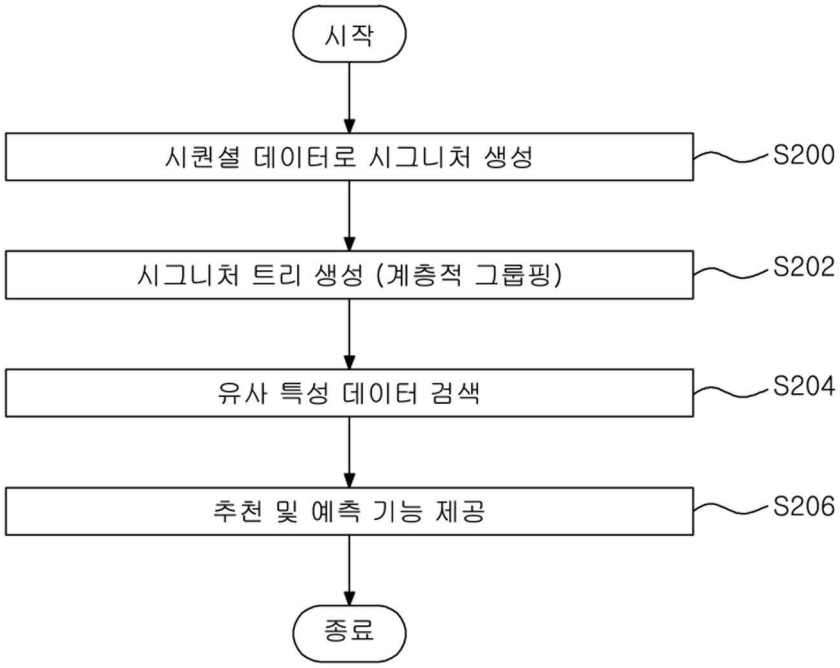
112 : 시그니처 트리부

도면

도면1

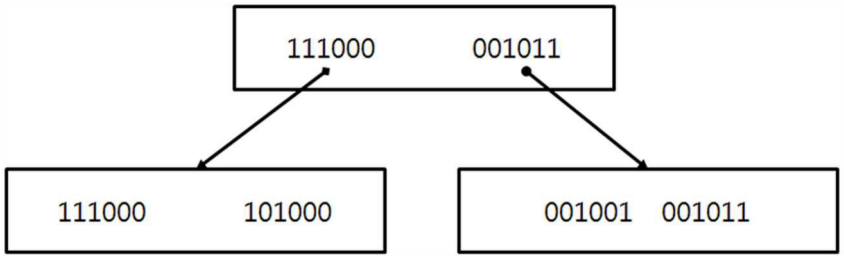


도면2

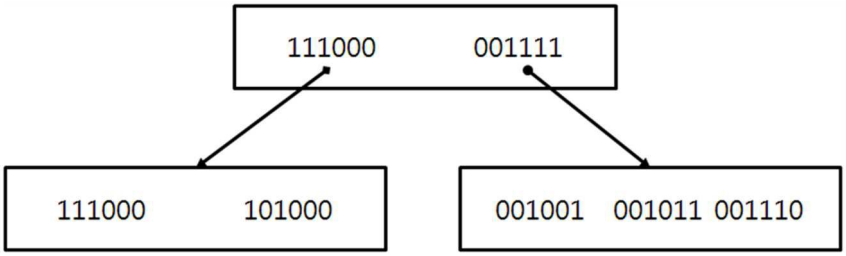


도면3

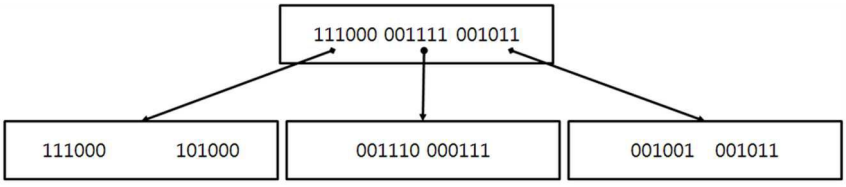
도면4



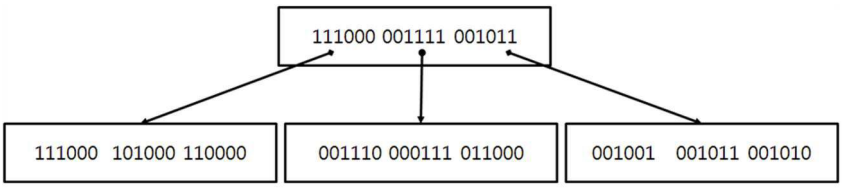
도면5



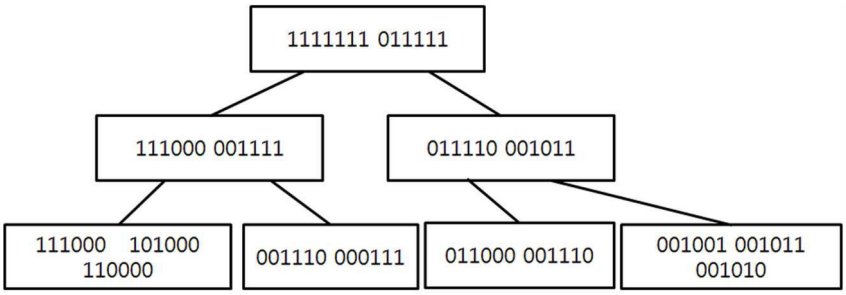
도면6



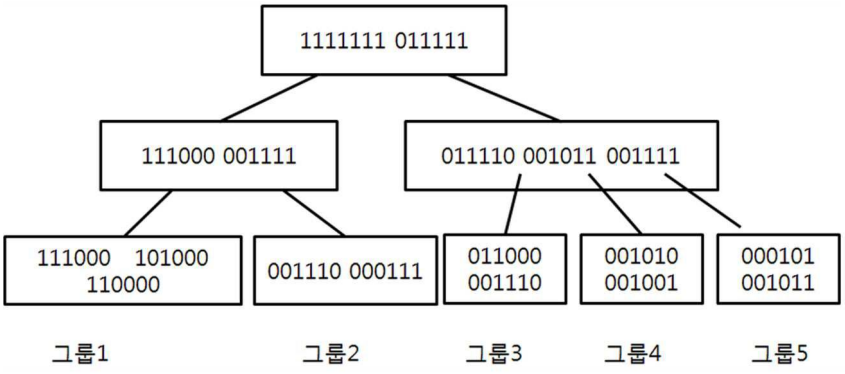
도면7



도면8


1111111 011111
111000 001111
011110 001011
111000 101000
110000
001110 000111
011000 001110
001001 001011
001010

도면9


1111111 011111
111000 001111
011110 001011 001111
111000 101000
110000
001110 000111
011000
001110
001010
001001
000101
001011
그룹1
그룹2
그룹3
그룹4
그룹5