



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0070852
(43) 공개일자 2023년05월23일

(51) 국제특허분류(Int. Cl.)
G06F 16/22 (2019.01) G06F 16/215 (2019.01)
G06F 16/25 (2019.01) G06F 16/28 (2019.01)
(52) CPC특허분류
G06F 16/2282 (2019.01)
G06F 16/215 (2019.01)
(21) 출원번호 10-2021-0156851
(22) 출원일자 2021년11월15일
심사청구일자 2021년11월15일

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
반효경
서울특별시 관악구 관악로30길 12, 107동 2001호 (봉천동, 관악우성아파트)
조경운
서울특별시 서초구 방배천로2안길 118, 3층(방배동)
(74) 대리인
특허법인 무한

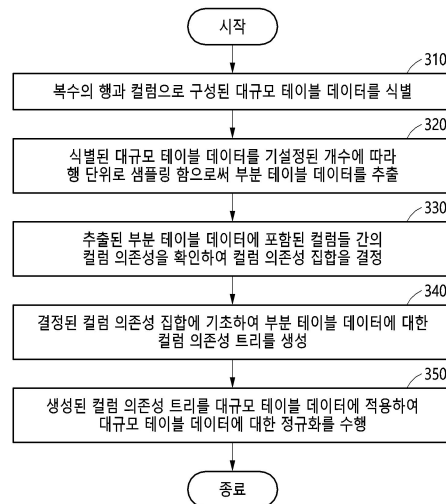
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 대규모 테이블 데이터 정규화 방법 및 장치

(57) 요약

대규모 테이블 데이터 정규화 방법 및 장치가 개시된다. 대규모 테이블 데이터 정규화 방법은 복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별하는 단계; 상기 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하는 단계; 상기 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 결정하는 단계; 상기 결정된 컬럼 의존성에 기초하여 상기 부분 테이블 데이터의 컬럼 의존성 트리를 생성하는 단계; 및 상기 생성된 컬럼 의존성 트리를 상기 대규모 테이블 데이터에 적용하여 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 단계를 포함할 수 있다.

대표도 - 도3



(52) CPC특허분류

G06F 16/2246 (2019.01)

G06F 16/258 (2019.01)

G06F 16/284 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711117117
과제번호	2020-0-00121-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발(R&D)
연구과제명	데이터 품질 평가기반 데이터 고도화 및 데이터셋 보정 기술 개발
기 여 율	1/1
과제수행기관명	부산대학교산학협력단
연구기간	2020.04.01 ~ 2020.12.31

명세서

청구범위

청구항 1

복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별하는 단계;

상기 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하는 단계;

상기 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 결정하는 단계;

상기 결정된 컬럼 의존성에 기초하여 상기 부분 테이블 데이터의 컬럼 의존성 트리를 생성하는 단계; 및

상기 생성된 컬럼 의존성 트리를 상기 대규모 테이블 데이터에 적용하여 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 단계

를 포함하는 대규모 테이블 데이터 정규화 방법.

청구항 2

제1항에 있어서,

상기 부분 테이블 데이터는,

상기 대규모 테이블 데이터를 구성하는 복수의 행들 중 랜덤하게 선택된 행들로 구성되는 대규모 테이블 데이터 정규화 방법.

청구항 3

제1항에 있어서,

상기 컬럼 의존성을 결정하는 단계는,

상기 부분 테이블 데이터를 구성하는 적어도 하나 이상의 컬럼을 선택하는 단계;

상기 선택된 컬럼을 구성하는 행의 값들 중 동일한 값을 가지는 행들을 부분 집합으로 하는 복수의 행집합(Row Set, RS)들을 식별하는 단계; 및

상기 식별된 복수의 행집합들 중 독립 컬럼에 대응하는 제1 행집합의 부분 집합들과 종속 컬럼에 대응하는 제2 행집합의 부분 집합들 사이의 포함 관계에 기초하여 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합을 결정하는 단계

를 포함하는 대규모 테이블 데이터 정규화 방법.

청구항 4

제3항에 있어서,

상기 컬럼 의존성 집합은,

최소 크기의 독립 컬럼으로 구성된 컬럼 의존성들의 집합인 대규모 테이블 데이터 정규화 방법.

청구항 5

제3항에 있어서,

상기 부분 테이블 데이터를 추출하는 단계는,

상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합이 변동되지 않을 때까지 반복하여 수행되는 대규모 테이블 데이터 정규화 방법.

청구항 6

제1항에 있어서,

상기 컬럼 의존성 트리를 생성하는 단계는,

상기 결정된 컬럼 의존성에 기초하여 독립 컬럼과 종속 컬럼을 나타내는 노드 간 순환이 발생하는 경우, 순환이 발생된 노드들을 하나의 통합 노드로 생성하는 대규모 테이블 데이터 정규화 방법.

청구항 7

제1항에 있어서,

상기 정규화를 수행하는 단계는,

상기 생성된 컬럼 의존성 트리가 다중 트리로 형성된 경우, 다중 트리의 루트 노드들을 통합하는 대규모 테이블 데이터 정규화 방법.

청구항 8

제1항에 있어서,

상기 정규화를 수행하는 단계는,

상기 생성된 컬럼 의존성 트리에 포함되지 않은 모든 컬럼들을 상기 컬럼 의존성 트리의 최상위 노드에 포함시키는 대규모 테이블 데이터 정규화 방법.

청구항 9

제1항에 있어서,

상기 정규화를 수행하는 단계는,

상기 컬럼 의존성 트리를 구성하는 노드별 컬럼 정보를 기반으로 정규화된 테이블 데이터를 생성하는 대규모 테이블 데이터 정규화 방법.

청구항 10

대규모 테이블 데이터 정규화 장치에 있어서,

상기 대규모 테이블 데이터 정규화 장치는 프로세서를 포함하고,

상기 프로세서는,

복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별하고, 상기 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하며, 상기 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 결정하고, 상기 결정된 컬럼 의존성에 기초하여 상기 부분 테이블 데이터의 컬럼 의존성 트리를 생성하며, 상기 생성된 컬럼 의존성 트리를 상기 대규모 테이블 데이터에 적용하여 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 대규모 테이블 데이터 정규화 장치.

청구항 11

제10항에 있어서,

상기 부분 테이블 데이터는,

상기 대규모 테이블 데이터를 구성하는 복수의 행들 중 랜덤하게 선택된 행들로 구성되는 대규모 테이블 데이터 정규화 장치.

청구항 12

제10항에 있어서,

상기 프로세서는,

상기 부분 테이블 데이터를 구성하는 적어도 하나 이상의 컬럼을 선택하고, 상기 선택된 컬럼을 구성하는 행의 값들 중 동일한 값을 가지는 행들을 부분 집합으로 하는 복수의 행집합(Row Set, RS)들을 식별하며, 상기 식별된 복수의 행집합들 중 독립 컬럼에 대응하는 제1 행집합의 부분 집합들과 종속 컬럼에 대응하는 제2 행집합의 부분 집합들 사이의 포함 관계에 기초하여 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합을 결정하는 대규모 테이블 데이터 정규화 장치.

청구항 13

제12항에 있어서,

상기 컬럼 의존성 집합은,

최소 크기의 독립 컬럼으로 구성된 컬럼 의존성들의 집합인 대규모 테이블 데이터 정규화 장치.

청구항 14

제12항에 있어서,

상기 프로세서는,

상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합이 변동되지 않을 때까지 부분 테이블 데이터를 반복하여 추출하는 대규모 테이블 데이터 정규화 장치.

청구항 15

제10항에 있어서,

상기 프로세서는,

상기 결정된 컬럼 의존성에 기초하여 독립 컬럼과 종속 컬럼을 나타내는 노드 간 순환이 발생하는 경우, 순환이 발생된 노드들을 하나의 통합 노드로 생성함으로써 컬럼 의존성 트리를 생성하는 대규모 테이블 데이터 정규화 장치.

청구항 16

제10항에 있어서,

상기 프로세서는,

상기 생성된 컬럼 의존성 트리가 다중 트리로 형성된 경우, 다중 트리의 루트 노드들을 통합함으로써 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 대규모 테이블 데이터 정규화 장치.

청구항 17

제10항에 있어서,

상기 프로세서는,

상기 생성된 컬럼 의존성 트리에 포함되지 않은 모든 컬럼들을 상기 컬럼 의존성 트리의 최상위 노드에 포함시킴으로써 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 대규모 테이블 데이터 정규화 장치.

청구항 18

제10항에 있어서,

상기 프로세서는,

상기 컬럼 의존성 트리를 구성하는 노드별 컬럼 정보를 기반으로 정규화된 테이블 데이터를 생성하는 대규모 테이블 데이터 정규화 장치.

발명의 설명

기술 분야

[0001] 본 발명은 대규모 테이블 데이터 정규화 방법 및 장치에 관한 것으로, 보다 구체적으로는 일정 규모 이상의 행들을 포함한 테이블 데이터를 일정 기준에 따라 행단위로 샘플링하여 컬럼 간 종속성을 파악함으로써 정규화 작업에 걸리는 시간을 단축시킬 수 있는 방법 및 장치에 관한 것이다.

배경 기술

[0002] 테이블 데이터는 컬럼(열)과 행으로 구성된 데이터이다. 일반적으로는 관계형 데이터베이스에서 주로 쓰는 표현 형식이며, 다수의 속성을 가지는 많은 개수의 항목들을 표시하는 방법이다. 테이블의 각 행은 데이터 항목이고, 컬럼은 각 항목이 가지는 속성으로서 테이블의 각 셀은 항목의 속성값으로 볼 수 있다.

[0003] 다양한 정보들을 하나의 테이블 데이터로 요약하게 되면, 각 컬럼들은 중복된 데이터가 많이 발생하게 된다. 예를 들어, 학생들이 수강하는 과목과 강사들의 정보를 단일 테이블로 표시하게 되면, 컬럼 정보는 학생ID, 과목명, 강사명으로 구성될 수 있다. 이때, 동일한 과목에 대한 강사명의 관계들이 모든 학생들 행 정보에 반복되는 문제가 발생한다. 이처럼 정규화되지 않은 테이블은 정합성과 데이터의 관계 모델을 명확히 파악하기 힘들다는 문제가 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명은 대규모 테이블 데이터 정규화 방법 및 장치에 관한 것으로, 보다 구체적으로는 일정 규모 이상의 행들을 포함한 테이블 데이터를 일정 기준에 따라 행단위로 샘플링하여 컬럼 간 종속성을 파악함으로써 정규화 작업에 걸리는 시간을 단축시킬 수 있는 방법 및 장치를 제공한다.

과제의 해결 수단

[0005] 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 방법은 복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별하는 단계; 상기 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하는 단계; 상기 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 결정하는 단계; 상기 결정된 컬럼 의존성에 기초하여 상기 부분 테이블 데이터의 컬럼 의존성 트리를 생성하는 단계; 및 상기 생성된 컬럼 의존성 트리를 상기 대규모 테이블 데이터에 적용하여 상기 대규모 테이블 데이터에 대한 정규화를 수행하는 단계를 포함할 수 있다.

[0006] 상기 부분 테이블 데이터는 상기 대규모 테이블 데이터를 구성하는 복수의 행들 중 랜덤하게 선택된 행들로 구성될 수 있다.

[0007] 상기 컬럼 의존성을 결정하는 단계는 상기 부분 테이블 데이터를 구성하는 적어도 하나 이상의 컬럼을 선택하는 단계; 상기 선택된 컬럼을 구성하는 행의 값들 중 동일한 값을 가지는 행들을 부분 집합으로 하는 복수의 행집합(Row Set, RS)들을 식별하는 단계; 및 상기 식별된 복수의 행집합들 중 독립 컬럼에 대응하는 제1 행집합의 부분 집합들과 종속 컬럼에 대응하는 제2 행집합의 부분 집합들 사이의 포함 관계에 기초하여 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합을 결정하는 단계를 포함할 수 있다.

[0008] 상기 컬럼 의존성 집합은 최소 크기의 독립 컬럼으로 구성된 컬럼 의존성들의 집합일 수 있다.

[0009] 상기 부분 테이블 데이터를 추출하는 단계는 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합이 변동되지 않을 때까지 반복하여 수행될 수 있다.

[0010] 상기 컬럼 의존성 트리를 생성하는 단계는 상기 결정된 컬럼 의존성에 기초하여 독립 컬럼과 종속 컬럼을 나타내는 노드 간 순환이 발생하는 경우, 순환이 발생된 노드들을 하나의 통합 노드로 생성할 수 있다.

[0011] 상기 정규화를 수행하는 단계는 상기 생성된 컬럼 의존성 트리가 다중 트리로 형성된 경우, 다중 트리의 루트 노드들을 통합할 수 있다.

[0012] 상기 정규화를 수행하는 단계는 상기 생성된 컬럼 의존성 트리에 포함되지 않은 모든 컬럼들을 상기 컬럼 의존성 트리의 최상위 노드에 포함시킬 수 있다.

[0013] 상기 정규화를 수행하는 단계는 상기 컬럼 의존성 트리를 구성하는 노드별 컬럼 정보를 기반으로 정규화된 데이

블 데이터를 생성할 수 있다.

- [0014] 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 장치는 프로세서를 포함하고, 상기 프로세서는 복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별하고, 상기 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하며, 상기 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 결정하고, 상기 결정된 컬럼 의존성에 기초하여 상기 부분 테이블 데이터의 컬럼 의존성 트리를 생성하며, 상기 생성된 컬럼 의존성 트리를 상기 대규모 테이블 데이터에 적용하여 상기 대규모 테이블 데이터에 대한 정규화를 수행할 수 있다.
- [0015] 상기 부분 테이블 데이터는 상기 대규모 테이블 데이터를 구성하는 복수의 행들 중 랜덤하게 선택된 행들로 구성될 수 있다.
- [0016] 상기 프로세서는 상기 부분 테이블 데이터를 구성하는 적어도 하나 이상의 컬럼을 선택하고, 상기 선택된 컬럼을 구성하는 행의 값들 중 동일한 값을 가지는 행들을 부분 집합으로 하는 복수의 행집합(Row Set, RS)들을 식별하며, 상기 식별된 복수의 행집합들 중 독립 컬럼에 대응하는 제1 행집합의 부분 집합들과 종속 컬럼에 대응하는 제2 행집합의 부분 집합들 사이의 포함 관계에 기초하여 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합을 결정할 수 있다.
- [0017] 상기 컬럼 의존성 집합은 최소 크기의 독립 컬럼으로 구성된 컬럼 의존성들의 집합일 수 있다.
- [0018] 상기 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합이 변동되지 않을 때까지 부분 테이블 데이터를 반복하여 추출할 수 있다.
- [0019] 상기 프로세서는 상기 결정된 컬럼 의존성에 기초하여 독립 컬럼과 종속 컬럼을 나타내는 노드 간 순환이 발생하는 경우, 순환이 발생된 노드들을 하나의 통합 노드로 생성함으로써 컬럼 의존성 트리를 생성할 수 있다.
- [0020] 상기 프로세서는 상기 생성된 컬럼 의존성 트리가 다중 트리로 형성된 경우, 다중 트리의 루트 노드들을 통합함으로써 상기 대규모 테이블 데이터에 대한 정규화를 수행할 수 있다.
- [0021] 상기 프로세서는 상기 생성된 컬럼 의존성 트리에 포함되지 않은 모든 컬럼들을 상기 컬럼 의존성 트리의 최상위 노드에 포함시킴으로써 상기 대규모 테이블 데이터에 대한 정규화를 수행할 수 있다.
- [0022] 상기 프로세서는 상기 컬럼 의존성 트리를 구성하는 노드별 컬럼 정보를 기반으로 정규화된 테이블 데이터를 생성할 수 있다.

발명의 효과

- [0023] 본 발명의 일실시예에 의하면, 일정 규모 이상의 행들을 포함한 테이블 데이터를 일정 기준에 따라 행단위로 샘플링하여 컬럼 간 종속성을 파악함으로써 정규화 작업에 걸리는 시간을 단축시킬 수 있다.

도면의 간단한 설명

- [0024] 도 1은 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 장치를 나타낸 도면이다.
- 도 2는 본 발명의 일실시예에 따른 컬럼 의존성의 결정 알고리즘을 나타낸 도면이다.
- 도 3은 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 방법을 플로우차트로 나타낸 도면이다.
- 도 4는 본 발명의 일실시예에 따른 컬럼 의존성 트리를 생성하는 제1 방법을 나타낸 도면이다.
- 도 5는 본 발명의 일실시예에 따른 컬럼 의존성 트리를 생성하는 제2 방법을 나타낸 도면이다.
- 도 6은 본 발명의 일실시예에 따른 컬럼 의존성 트리를 이용하여 테이블 데이터를 정규화하는 제1 방법을 나타낸 도면이다.
- 도 7은 본 발명의 일실시예에 따른 컬럼 의존성 트리를 이용하여 테이블 데이터를 정규화하는 제2 방법을 나타낸 도면이다.
- 도 8은 본 발명의 일실시예에 따른 컬럼 의존성 트리를 이용하여 테이블 데이터를 정규화하는 제3 방법을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0026] 도 1은 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 장치를 나타낸 도면이다.
- [0027] 데이터 배포시 많이 사용되는 CSV(Comma-Separated Values) 데이터와 같은 경우에는 정규화된 다수의 테이블 데이터들을 배포의 편의성을 위해 비정규화(Denormalization) 하는 경우가 일반적이다. 이처럼 정규화되지 않은 테이블 데이터는 다수의 항목들이 혼합되어 있어 항목들 간의 관계 모델을 파악하기 힘든 문제가 있다.
- [0028] 본 발명에서 제공하는 대규모 테이블 데이터 정규화 장치(100)의 프로세서(110)는 이와 같은 비정규화된 특성을 가지는 대규모 테이블 데이터(120)에 대해 항목들 간의 관계를 보다 명확히 파악할 수 있도록 논리적인 부분 테이블 데이터(130)로 분리하기 위한 정규화 방법을 제공할 수 있다.
- [0029] 이를 위해 대규모 테이블 데이터 정규화 장치(100)의 프로세서(110)는 대규모 테이블 데이터(120)를 구성하는 항목들 간의 컬럼 의존성을 파악해야 된다. 본 발명에서 컬럼 의존성은 $(X) \rightarrow (Y)$ 로 표현할 수 있다. 이는 X 행에 해당하는 컬럼값이 Y 행에 해당하는 컬럼값에 의존적임을 나타낸다.
- [0030] 한편, 컬럼 의존성을 나타내는 $(X) \rightarrow (Y)$ 는 $f(X)=Y$ 의 함수 형식으로도 표현할 수 있으며, 본 발명에서는 (X)를 독립 컬럼, (Y)를 종속 컬럼으로 표현하기로 한다. 이때, 종속 컬럼은 항상 1개이나 독립 컬럼은 $f(X, W, Z)=Y$ 와 같이 다수개의 조합으로 표현될 수도 있다.
- [0031] 도 2는 본 발명의 일실시예에 따른 컬럼 의존성의 결정 알고리즘을 나타낸 도면이다.
- [0032] 대규모 테이블 데이터 정규화 장치(100)의 프로세서(110)는 테이블 데이터를 구성하는 적어도 하나 이상의 컬럼을 선택하고, 선택된 컬럼을 구성하는 행의 컬럼값들 중 동일한 컬럼값을 가지는 행들을 부분 집합으로 하는 복수의 행집합(Row Set, RS)들을 식별할 수 있다.
- [0033] 일례로, 도 2를 참고하면, 프로세서(110)는 테이블 데이터를 구성하는 컬럼들 중 매출시각(A)에 대응하는 컬럼이 선택된 경우, 매출시각(A)의 컬럼값들 중 동일한 컬럼값을 가지는 행들을 부분 집합으로 하는 행집합 RS(A)를 결정할 수 있다.
- [0034] 보다 구체적으로, 프로세서(110)는 매출시각(A)의 1행과 2행의 컬럼값이 11:30으로 동일하므로 1행과 2행을 하나의 부분 집합으로 설정하고, 4행과 5행의 컬럼값이 12:22으로 동일하므로 4행과 5행을 하나의 부분 집합으로 설정할 수 있다. 그리고, 프로세서(110)는 동일한 컬럼값을 가지는 행이 없는 3행을 별도의 부분 집합으로 설정함으로써 행집합 RS(A)를 결정할 수 있다.
- [0035] 마찬가지로, 프로세서(110)는 매출액(B)의 3행, 4행 및 5행의 컬럼값이 2000으로 동일하므로 3행, 4행 및 5행을 하나의 부분 집합으로 설정하고, 동일한 컬럼값을 가지는 행이 없는 1행과 2행을 각각 별도의 부분 집합으로 설정함으로써 행집합 RS(B)를 결정할 수 있다.
- [0036] 상기의 예에서는 하나의 컬럼에 대응하는 행집합을 결정하는 구성을 개시하였으나, 프로세서(110)는 하나 이상의 컬럼들을 선택하고, 선택된 컬럼들에 대한 행집합을 결정할 수 있다.
- [0037] 일례로, 프로세서(110)는 테이블 데이터를 구성하는 컬럼들 중 매출시각(A), 매출액(B) 및 수량(E)이 선택된 경우, 선택된 컬럼들의 컬럼값들이 모두 동일한 컬럼값을 가지는 행들을 부분 집합으로 하는 행집합 RS(A, B, E)를 식별할 수 있다. 이때, 도 2의 예서는 매출시각(A), 매출액(B) 및 수량(E)의 행들 중 3개 항목의 컬럼값이 모두 동일한 행이 없으므로 각각의 행을 별도의 부분 집합으로 설정함으로써 행집합 RS(A, B, E)를 결정할 수 있다.
- [0038] 이때, 프로세서(110)는 테이블 데이터를 구성하는 컬럼들의 개수가 N 개인 경우, 총 $2^N - 1$ 개의 조합에 대한 행집합들을 결정할 수 있으며, 아래의 식 1은 도 2의 테이블 데이터에 대한 행집합들의 일부 예를 나타낸 것이다.
- [0039] <식 1>
- [0040] $RS(A)=\{\{1, 2\}, \{3\}, \{4, 5\}\}$
- [0041] $RS(B)=\{\{1\}, \{2\}, \{3, 4, 5\}\}$
- [0042] $RS(C)=\{\{1, 3\}, \{2, 5\}, \{4\}\}$

- [0043] $RS(D)=\{\{1, 3, 4\}, \{2, 5\}\}$
- [0044] $RS(E)=\{\{1,5\}, \{2\}, \{3, 4\}\}$
- [0045] $RS(A, B, E)=\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$
- [0046] 이때, 프로세서(110)는 복수의 행집합들 중 독립 컬럼에 대응하는 행집합의 부분 집합들과 종속 컬럼에 대응하는 행집합의 부분 집합들 사이의 포함 관계에 기초하여 테이블 데이터를 구성하는 컬럼들 간의 컬럼 의존성을 나타내는 컬럼 의존성 집합을 결정할 수 있다. 보다 구체적으로 프로세서(110)는 독립 컬럼에 대응하는 행집합의 모든 부분 집합들이 종속 컬럼에 대응하는 행집합의 부분 집합들 중 어느 하나의 부분 집합에 포함되는 경우 컬럼 의존성이 있다고 판단하고, 이를 토대로 컬럼 의존성 집합을 결정할 수 있다.
- [0047] 예를 들어, 독립 컬럼 A와 종속 컬럼 C가 선택된 경우, 프로세서(110)는 독립 컬럼 A의 행집합 $RS(A)$ 에 포함된 부분 집합들과 종속 컬럼 C의 행집합 $RS(C)$ 에 포함된 부분 집합들 사이의 포함 관계를 판단함으로써 컬럼 의존성을 확인할 수 있다. 그러나, 독립 컬럼 A의 행집합 $RS(A)$ 에 포함된 부분 집합들은 종속 컬럼 C의 행집합 $RS(C)$ 에 포함된 부분 집합들 어디에도 포함되지 않으므로, 프로세서(110)는 독립 컬럼 A와 종속 컬럼 C 간에 컬럼 의존성이 없다고 판단할 수 있다.
- [0048] 이와는 달리 독립 컬럼 C와 종속 컬럼 D가 선택된 경우, 프로세서(110)는 독립 컬럼 C의 행집합 $RS(C)$ 에 포함된 모든 부분 집합들이 종속 컬럼 D의 행집합 $RS(D)$ 에 포함된 부분 집합들 중 어느 하나의 부분 집합에 포함되므로 독립 컬럼 C와 종속 컬럼 D 간에 컬럼 의존성이 있다고 판단할 수 있다. 이후 프로세서(110)는 컬럼 의존성이 있다고 판단된 독립 컬럼 C와 종속 컬럼 D 간의 컬럼 의존성을 나타내기 위하여 $(C) \rightarrow (D)$ 와 같이 컬럼 의존성 집합을 표현할 수 있다.
- [0049] 프로세서(110)는 종속 컬럼은 1개만 고려하고, 다양한 조합의 독립 컬럼들에 대해 컬럼 의존성을 결정할 수 있다. 일례로, A, B, C, D, E 5개의 컬럼에 대해 A를 종속 컬럼으로 하는 컬럼 의존성을 찾고자 하는 경우, 프로세서(110)는 B, C, D, E, BC, BD, BE, CD, CE, DE, BCD, BCE, BDE, CDE, BCDE의 총 15개 독립 컬럼에 대한 컬럼 의존성을 확인할 수 있다.
- [0050] 이때, 프로세서(110)는 컬럼 의존성이 확인된 독립 컬럼의 상위 집합에 대해서는 추가적인 컬럼 의존성 확인을 수행하지 않을 수 있다. 일례로, 컬럼 의존성 집합 $(B) \rightarrow (C)$ 가 발견된 경우, 프로세서(110)는 나머지 조합, 일례로 $(B, C) \rightarrow (A)$ 나 $(B, C, D, E) \rightarrow (A)$ 등에 대해서는 컬럼 의존성 확인을 수행하지 않을 수 있다. 즉, 컬럼 의존성 집합은 최소 크기의 독립 컬럼으로 구성된 컬럼 의존성들의 집합일 수 있다.
- [0051] 도 3은 본 발명의 일실시예에 따른 대규모 테이블 데이터 정규화 방법을 플로우차트로 나타낸 도면이다.
- [0052] 단계(310)에서, 대규모 테이블 데이터 정규화 장치(100)의 프로세서(110)는 복수의 행과 컬럼으로 구성된 대규모 테이블 데이터를 식별할 수 있다. 이때, 테이블 데이터의 각 행은 항목을 나타내고, 컬럼은 각 항목이 가지는 속성으로써 테이블의 각 셀은 항목의 속성값을 나타낼 수 있다.
- [0053] 단계(320)에서, 프로세서(110)는 식별된 대규모 테이블 데이터를 기설정된 개수에 따라 행 단위로 샘플링 함으로써 부분 테이블 데이터를 추출하고, 단계(330)에서 추출된 부분 테이블 데이터에 포함된 컬럼들 간의 컬럼 의존성을 확인하여 컬럼 의존성 집합을 결정할 수 있다.
- [0054] 위에서 언급한 컬럼 의존성을 이용하여 테이블 데이터를 정규화하는 경우, 임의의 컬럼들 간의 조합에 대한 컬럼 의존성을 판단해야 하므로 테이블 데이터의 크기가 클수록 많은 시간이 소요되는 문제가 발생될 수 있다. 이에 프로세서(110)는 대규모 테이블 데이터에서 랜덤하게 샘플링하고, 샘플링된 데이터 내에서 컬럼 의존성을 확인하여 컬럼 의존성 집합을 결정함으로써 정규화 작업에 소요되는 시간을 단축시킬 수 있다.
- [0055] 프로세서(110)는 샘플링을 통해 추출된 부분 테이블 데이터에 대한 컬럼 의존성 집합을 결정하고, 결정된 컬럼 의존성 집합이 변동되지 않을 때까지 반복적으로 샘플링을 통한 부분 테이블 데이터 추출 작업을 수행할 수 있다. 이때, 샘플링을 통해 추출되는 행의 개수는 변경 가능하며, 이전 샘플링을 통해 추출된 행이 반복적인 샘플링 과정에서 배제되거나, 재추출될 수도 있다.
- [0056] 이때, 프로세서(110)는 최초의 컬럼 의존성 확인 과정에서는 모든 조합의 독립 컬럼들에 대해 컬럼 의존성 집합을 결정할 수 있으나, 2회 시행부터는 전회에서 결정된 컬럼 의존성 집합들에 대해서만 컬럼 의존성 만족 여부를 판단할 수 있다. 만약 컬럼 의존성이 만족되지 않는 경우, 프로세서(110)는 컬럼 의존성 집합에서 해당하는 컬럼 의존성을 삭제할 수 있다.

- [0057] 단계(340)에서, 프로세서(110)는 결정된 컬럼 의존성 집합에 기초하여 부분 테이블 데이터에 대한 컬럼 의존성 트리를 생성할 수 있다. 보다 구체적으로 프로세서(110)는 부분 테이블 데이터에 대한 반복적인 컬럼 의존성 확인 결과 결정된 컬럼 의존성 집합으로부터 컬럼 의존성 트리를 생성할 수 있다.
- [0058] 프로세서(110)는 컬럼 의존성 집합의 독립 컬럼과 종속 컬럼을 각각 노드로 설정하고, 화살표를 통해 노드들을 서로 연결함으로써 컬럼 의존성 트리를 생성할 수 있다. 도 4의 (a)를 참고하면, 컬럼 의존성 집합이 (A)→(B), (B)→(C) 이므로 (A) 노드에서 (B) 노드로 화살표를 연결하고, (B) 노드에서 (C) 노드로 화살표를 연결함으로써 컬럼 의존성 트리를 생성할 수 있다.
- [0059] 이때, 생성되는 컬럼 의존성 트리는 도 4의 (b)와 같이 자식노드 또는 부모노드가 복수개 존재하거나, 루트가 여러 개인 다중 트리로 구성될 수도 있다.
- [0060] 한편, 컬럼 의존성 간 순환이 발생하는 경우, 프로세서(110)는 순환되는 노드들을 합쳐 통합 노드를 생성할 수 있다. 일례로, 도 5를 참고하면, (A) 노드, (B) 노드 및 (C) 노드는 컬럼 의존성 간 순환이 발생하였고, 프로세서(110)는 이와 같이 순환이 발생된 노드들을 'I'를 통해 합침으로써 통합 노드를 생성할 수 있다. 이때, 순환되는 노드들의 하위 노드들은 통합 노드의 하위 노드에 포함될 수 있으며, 상위 노드가 다중 항목으로 구성된 경우, 독립 컬럼의 정보를 명시적으로 표시될 수 있다.
- [0061] 단계(350)에서, 프로세서(110)는 생성된 컬럼 의존성 트리를 대규모 테이블 데이터에 적용하여 대규모 테이블 데이터에 대한 정규화를 수행할 수 있다. 먼저, 프로세서(110)는 컬럼 의존성 트리가 도 6의 (a)와 같이 다중 트리로 생성된 경우, 모든 트리의 루트 노드들을 합쳐 통합 노드로 전환할 수 있다. 이때, 통합 노드의 하위 노드들은 모두 독립 컬럼과 종속 컬럼의 정보가 명시적으로 표시될 수 있다.
- [0062] 한편, 도 6의 (b)와 같이 컬럼 의존성 트리를 구성하는 어느 하나의 트리에 대한 종속 컬럼이 다른 트리의 독립 컬럼에 포함되는 경우, 프로세서(110)는 해당 종속 컬럼에 대한 컬럼 의존성을 삭제할 수 있다. 일례로, (B) 노드는 (A) 노드에 대한 종속 컬럼이면서, (D) 노드에 대한 독립 컬럼임을 알 수 있다. 이때, 프로세서(110)는 (B) 노드를 종속 컬럼으로 하는 컬럼 의존성 (A)→(B)를 삭제하여 컬럼 의존성 트리를 생성할 수 있다. 이를 통해 본 발명은 약간의 종속성 정보를 잃더라도 중복성을 제거하여 보다 효율적인 컬럼 의존성 트리를 생성할 수 있다.
- [0063] 다음으로 프로세서(110)는 도 7과 같이 컬럼 의존성 트리에 포함되지 않은 모든 컬럼들을 컬럼 의존성 트리의 최상위 노드에 포함시킬 수 있다.
- [0064] 마지막으로 프로세서(110)는 도 8과 같이 컬럼 의존성 트리를 구성하는 각 노드 별 컬럼 정보를 기반으로 정규화된 테이블 데이터를 생성할 수 있다.
- [0065] 한편, 본 발명에 따른 방법은 컴퓨터에서 실행될 수 있는 프로그램으로 작성되어 마그네틱 저장매체, 광학적 판독매체, 디지털 저장매체 등 다양한 기록 매체로도 구현될 수 있다.
- [0066] 본 명세서에 설명된 각종 기술들의 구현들은 디지털 전자 회로조직으로, 또는 컴퓨터 하드웨어, 펌웨어, 소프트웨어로, 또는 그들의 조합들로 구현될 수 있다. 구현들은 데이터 처리 장치, 예를 들어 프로그램가능 프로세서, 컴퓨터, 또는 다수의 컴퓨터들의 동작에 의한 처리를 위해, 또는 이 동작을 제어하기 위해, 컴퓨터 프로그램 제품, 즉 정보 캐리어, 예를 들어 기계 판독가능 저장 장치(컴퓨터 판독가능 매체) 또는 전파 신호에서 유형적으로 구체화된 컴퓨터 프로그램으로서 구현될 수 있다. 상술한 컴퓨터 프로그램(들)과 같은 컴퓨터 프로그램은 컴파일된 또는 인터프리트된 언어들을 포함하는 임의의 형태의 프로그래밍 언어로 기록될 수 있고, 독립형 프로그램으로서 또는 모듈, 구성요소, 서브루틴, 또는 컴퓨팅 환경에서의 사용에 적절한 다른 유닛으로서 포함하는 임의의 형태로 전개될 수 있다. 컴퓨터 프로그램은 하나의 사이트에서 하나의 컴퓨터 또는 다수의 컴퓨터들 상에서 처리되도록 또는 다수의 사이트들에 걸쳐 분배되고 통신 네트워크에 의해 상호 연결되도록 전개될 수 있다.
- [0067] 컴퓨터 프로그램의 처리에 적절한 프로세서들은 예로서, 범용 및 특수 목적 마이크로프로세서들 둘 다, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서들을 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 다로부터 명령어 및 데이터를 수신할 것이다. 컴퓨터의 요소들은 명령어들을 실행하는 적어도 하나의 프로세서 및 명령어 및 데이터를 저장하는 하나 이상의 메모리 장치들을 포함할 수 있다. 일반적으로, 컴퓨터는 데이터를 저장하는 하나 이상의 대량 저장 장치들, 예를 들어 자기, 자기-광 디스크들, 또는 광 디스크들을 포함할 수 있거나, 이것들로부터 데이터를 수신하거나 이것들에 데이터를 송신하거나 또는 양쪽으로 되도록 결합될 수도 있다. 컴퓨터 프로그램 명령어 및 데이터를 구체화하는데 적절한 정

보 캐리어들은 예로서 반도체 메모리 장치들, 예를 들어, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM(Compact Disk Read Only Memory), DVD(Digital Video Disk)와 같은 광 기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 롬(ROM, Read Only Memory), 램(RAM, Random Access Memory), 플래시 메모리, EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM) 등을 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로조직에 의해 보충되거나, 이에 포함될 수 있다.

[0068] 또한, 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용매체일 수 있고, 컴퓨터 저장매체 및 전송매체를 모두 포함할 수 있다.

[0069] 본 명세서는 다수의 특정한 구현물의 세부사항들을 포함하지만, 이들은 어떠한 발명이나 청구 가능한 것의 범위에 대해서도 제한적인 것으로서 이해되어서는 안되며, 오히려 특정한 발명의 특정한 실시형태에 특유할 수 있는 특징들에 대한 설명으로서 이해되어야 한다. 개별적인 실시형태의 문맥에서 본 명세서에 기술된 특정한 특징들은 단일 실시형태에서 조합하여 구현될 수도 있다. 반대로, 단일 실시형태의 문맥에서 기술한 다양한 특징들 역시 개별적으로 혹은 어떠한 적절한 하위 조합으로도 복수의 실시형태에서 구현 가능하다. 나아가, 특징들이 특정한 조합으로 동작하고 초기에 그와 같이 청구된 바와 같이 묘사될 수 있지만, 청구된 조합으로부터의 하나 이상의 특징들은 일부 경우에 그 조합으로부터 배제될 수 있으며, 그 청구된 조합은 하위 조합이나 하위 조합의 변형물로 변경될 수 있다.

[0070] 마찬가지로, 특정한 순서로 도면에서 동작들을 묘사하고 있지만, 이는 바람직한 결과를 얻기 위하여 도시된 그 특정한 순서나 순차적인 순서대로 그러한 동작들을 수행하여야 한다거나 모든 도시된 동작들이 수행되어야 하는 것으로 이해되어서는 안 된다. 특정한 경우, 멀티태스킹과 병렬 프로세싱이 유리할 수 있다. 또한, 상술한 실시형태의 다양한 장치 컴포넌트의 분리는 그러한 분리를 모든 실시형태에서 요구하는 것으로 이해되어서는 안되며, 설명한 프로그램 컴포넌트와 장치들은 일반적으로 단일의 소프트웨어 제품으로 함께 통합되거나 다중 소프트웨어 제품에 패키징 될 수 있다는 점을 이해하여야 한다.

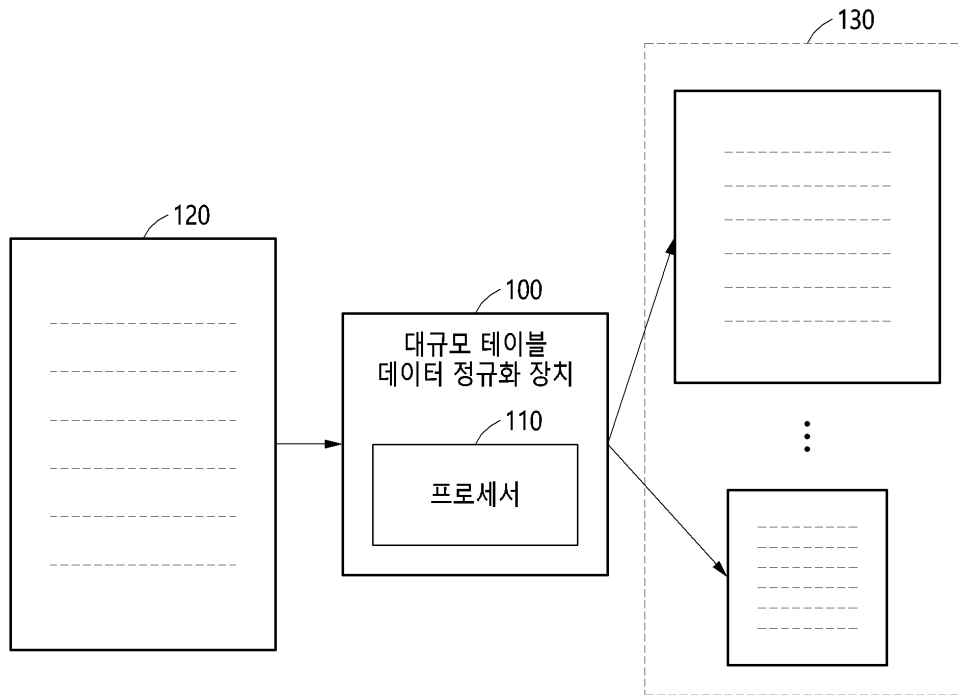
[0071] 한편, 본 명세서와 도면에 개시된 본 발명의 실시 예들은 이해를 돕기 위해 특정 예를 제시한 것에 지나지 않으며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 자명한 것이다.

부호의 설명

[0072] 100 : 대규모 테이블 데이터 정규화 장치
110 : 프로세서
120 : 대규모 테이블 데이터
130 : 부분 테이블 데이터

도면

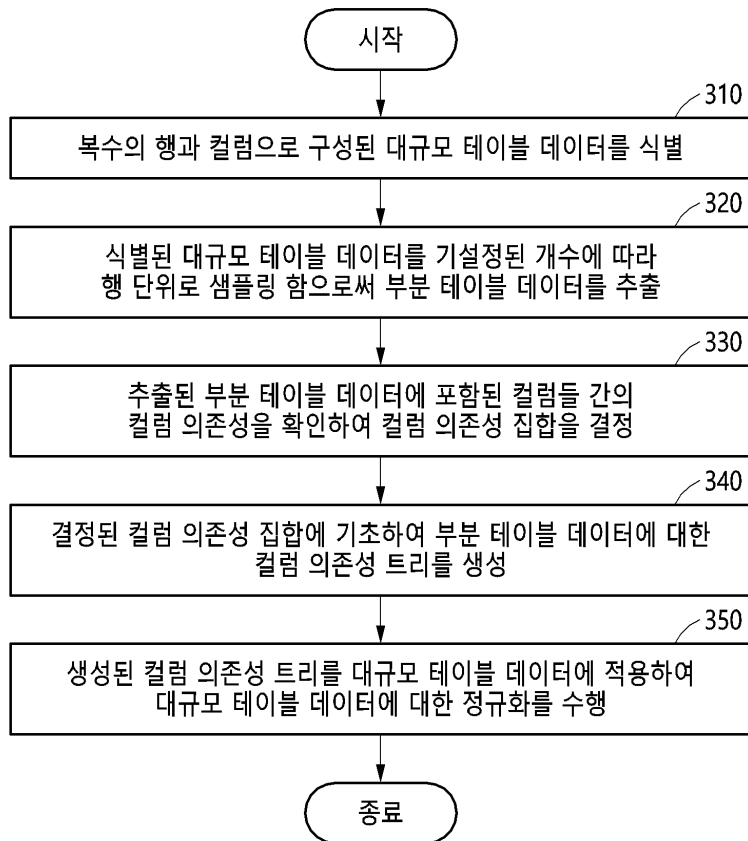
도면1



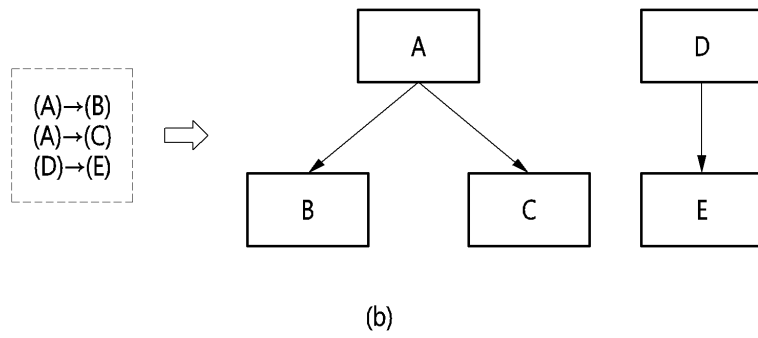
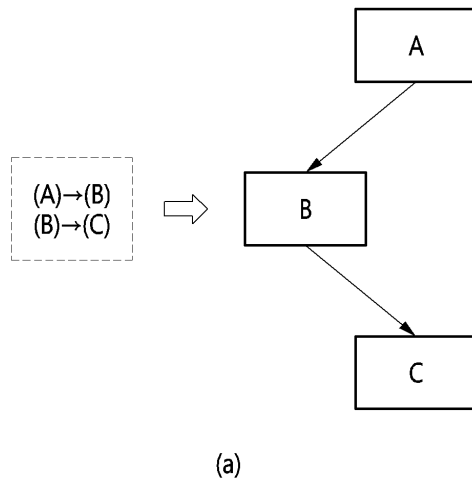
도면2

행번호 (가상)	매출시각 (A)	매출액 (B)	상품명 (C)	단가 (D)	수량 (E)
1	11:30	4000	포도	2000	2
2	11:30	3000	사과	1000	3
3	12:11	2000	포도	2000	1
4	12:22	2000	바나나	2000	1
5	12:22	2000	사과	1000	2

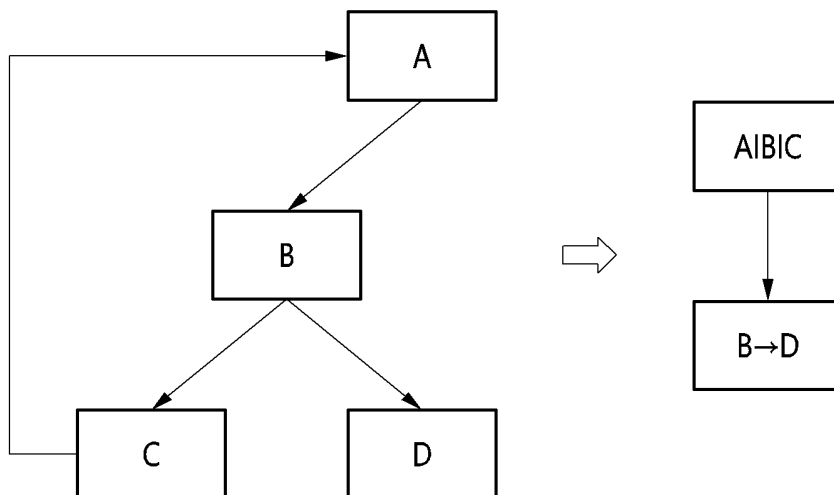
도면3



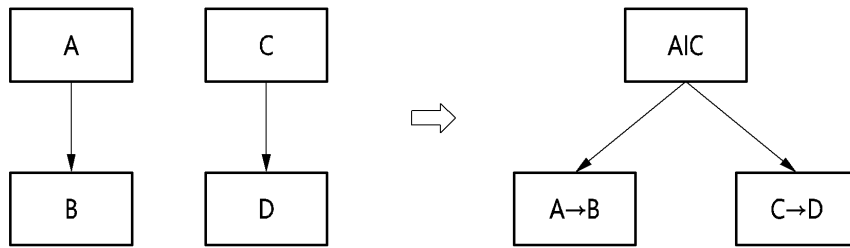
도면4



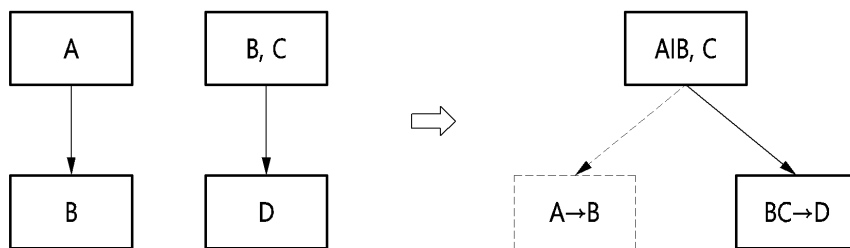
도면5



도면6

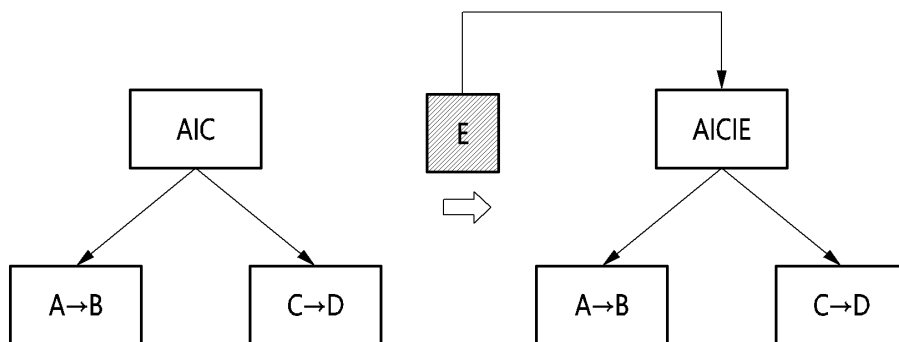


(a)



(b)

도면7



도면8

