



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2015년11월03일

(11) 등록번호 10-1565715

(24) 등록일자 2015년10월28일

(51) 국제특허분류(Int. Cl.)

G06F 17/30 (2006.01)

(21) 출원번호 10-2014-0187250

(22) 출원일자 2014년12월23일

심사청구일자 2014년12월23일

(56) 선행기술조사문헌

KR1020130102930 A

(73) 특허권자

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)

(72) 발명자

이영구

대전광역시 서구 청사서로 11, 106동 302호(월평동, 무지개아파트)

한용구

경기도 용인시 기흥구 덕영대로 1732, 경희대학교 국제캠퍼스전자정보관 349호(서천동)

박기성

경기도 부천시 원미구 원미로 123-4, 백송연립 E동 4호(원미동)

(74) 대리인

송인호, 민영준, 최관락

전체 청구항 수 : 총 10 항

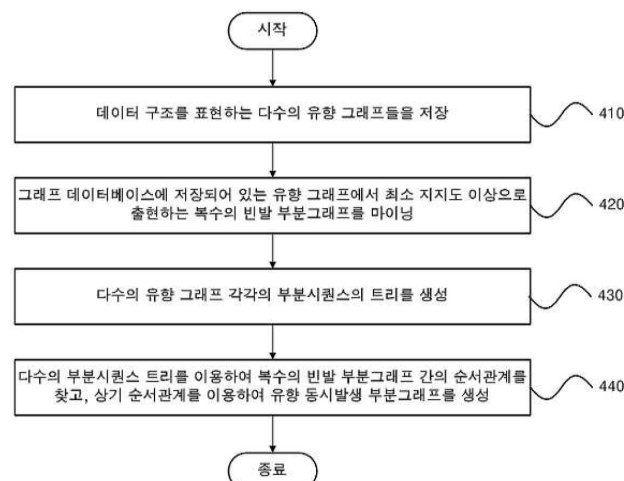
심사관 : 이복현

(54) 발명의 명칭 **유형 동시발생 부분그래프 생성 장치 및 방법**

(57) 요약

유형 동시발생 부분그래프 생성 장치 및 방법이 개시된다. 개시된 유형 동시발생 부분그래프 생성 방법은 프로세서가 포함된 장치에서, 그래프 데이터베이스에 저장되어 있는 다수의 유형 그래프에 대한 유형 동시발생 부분그래프를 생성하는 방법으로서, 상기 다수의 유형 그래프를 대상으로 복수의 빈발 부분그래프를 생성하는 단계; 상기 다수의 유형 그래프 각각의 부분시퀀스의 트리(다수의 부분시퀀스 트리)를 생성하는 단계 - 상기 부분시퀀스는 유형 그래프에 포함되는 복수의 버텍스가 순차적으로 에지로 연결되어 있는 데이터 시퀀스이고, 상기 부분시퀀스 트리는 상기 유형 그래프에 존재하는 부분시퀀스 간의 순서를 나타내는 데이터 트리임 - ; 및 상기 다수의 부분시퀀스 트리를 이용하여 상기 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 상기 유형 동시발생 부분그래프를 생성하는 단계;를 포함한다.

대표도 - 도4



이 발명을 지원한 국가연구개발사업

과제고유번호 2012R1A2A2A01047478

부처명 교육과학기술부

연구관리전문기관 한국연구재단

연구사업명 중견연구자지원사업(핵심)

연구과제명 행위 데이터베이스에서 행위패턴 모델링 및 마이닝 기술연구

기 여 율 1/1

주관기관 경희대학교산학협력단

연구기간 2013.09.01 ~ 2014.08.31

명세서

청구범위

청구항 1

프로세서가 포함된 장치에서, 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 생성하는 방법에 있어서,

상기 다수의 유향 그래프를 대상으로 복수의 빈발 부분그래프를 생성하는 단계;

상기 다수의 유향 그래프 각각의 부분시퀀스의 트리(다수의 부분시퀀스 트리)를 생성하는 단계 - 상기 부분시퀀스는 유향 그래프에 포함되는 복수의 버텍스가 순차적으로 에지로 연결되어 있는 데이터 시퀀스이고, 상기 부분시퀀스 트리는 상기 유향 그래프에 존재하는 부분시퀀스 간의 순서를 나타내는 데이터 트리임 - ; 및

상기 다수의 부분시퀀스 트리를 이용하여 상기 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 상기 유향 동시발생 부분그래프를 생성하는 단계;를 포함하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법.

청구항 2

제1항에 있어서,

상기 부분시퀀스 트리를 구성하는 노드 각각은 하나의 부분시퀀스와 대응되고,

상기 부분시퀀스 트리를 생성하는 단계는, 루트노드를 기준으로 하여 순차적으로 노드를 생성해 가면서 상기 부분시퀀스 트리를 생성하되, 상기 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신함으로써 상기 부분시퀀스 트리를 생성하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법.

청구항 3

제2항에 있어서,

상기 부분시퀀스 트리를 생성하는 단계는,

상기 노드 중 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 0인 경우, 상기 제1 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신하지 않고 확정하며, 새로운 노드를 생성하지 않고,

상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 1인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 하나의 버텍스를 상기 제1 노드와 대응되는 부분시퀀스에 추가하여 갱신하고,

상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 2인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 2개의 버텍스를 검색하고, 상기 2개의 버텍스 중 하나의 버텍스를 상기 제1 노드의 자식노드인 제2 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, 상기 2개의 버텍스 중 다른 버텍스를 상기 제1 노드의 또 다른 자식노드인 제3 노드와 대응되는 부분시퀀스에 추가하여 갱신하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법.

청구항 4

제3항에 있어서,

상기 부분시퀀스의 트리를 생성하는 단계는,

상기 유향 그래프의 버텍스들 중 내향 에지가 0인 버텍스를 검색하고, 상기 내향 에지가 0인 버텍스를 상기 부분시퀀스 트리의 루트노드의 자식노드의 첫번째 버텍스로 추가하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법.

청구항 5

제1항에 있어서,

상기 유향 동시발생 부분그래프를 생성하는 단계는,

상기 복수의 빈발 부분그래프에 대해, 빈발 부분그래프 A와 빈발 부분그래프 B로 구성된 빈발 부분그래프 쌍들을 생성하고,

상기 빈발 부분그래프 쌍들 각각을 대상으로 하여, 빈발 부분그래프 쌍과 상기 다수의 부분시퀀스 트리 각각을 비교하여, 빈발 부분그래프 A와 빈발 부분그래프 B의 순서관계가 만족되는 부분그래프가 최소 지지도 이상 출현하는 경우 상기 빈발 부분그래프 쌍을 상기 유향 동시발생 부분그래프로 판단함으로써 상기 유향 동시발생 부분그래프를 생성하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법.

청구항 6

제1항 내지 제5항 중 어느 한 항의 방법을 수행하는 프로그램을 기록한 컴퓨터 판독 가능 기록 매체.

청구항 7

그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 생성하는 장치에 있어서,

상기 다수의 유향 그래프를 대상으로 복수의 빈발 부분그래프를 생성하는 빈발 부분그래프 마이닝부;

상기 다수의 유향 그래프 각각의 부분시퀀스의 트리(다수의 부분시퀀스 트리)를 생성하는 부분시퀀스 트리 생성부 - 상기 부분시퀀스는 유향 그래프에 포함되는 복수의 버텍스가 순차적으로 에지로 연결되어 있는 데이터 시퀀스이고, 상기 부분시퀀스 트리는 상기 유향 그래프에 존재하는 부분시퀀스 간의 순서를 나타내는 데이터 트리임 -; 및

상기 다수의 부분시퀀스 트리를 이용하여 상기 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 상기 유향 동시발생 부분그래프를 생성하는 유향 동시발생 부분그래프 생성부;를 포함하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 장치.

청구항 8

제7항에 있어서,

상기 부분시퀀스 트리를 구성하는 노드 각각은 하나의 부분시퀀스와 대응되고,

상기 부분시퀀스 트리 생성부는 루트노드를 기준으로 하여 순차적으로 노드를 생성해 가면서 상기 부분시퀀스 트리를 생성하되, 상기 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신함으로써 상기 부분시퀀스 트리를 생성하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 장치.

청구항 9

제8항에 있어서,

상기 부분시퀀스 트리 생성부는,

상기 노드 중 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 0인 경우, 상기 제1 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신하지 않고 확정하며, 새로운 노드를 생성하지 않고,

상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 1인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 하나의 버텍스를 상기 제1 노드와 대응되는 부분시퀀스에 추가하여 갱신하고,

상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 2인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 2개의 버텍스를 검색하고, 상기 2개의 버텍스 중 하나의 버텍스를 상기 제1 노드의 자식노드인 제2 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, 상기 2개의 버텍스 중 다른 버텍스를 상기 제1 노드의 또 다른 자식노드인 제3 노드와 대응되는 부분시퀀스에 추가하여 갱신하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 장치.

청구항 10

제7항에 있어서,

상기 유향 동시발생 부분그래프 생성부는,

상기 복수의 빈발 부분그래프에 대해, 빈발 부분그래프 A와 빈발 부분그래프 B로 구성된 빈발 부분그래프 쌍들을 생성하고,

상기 빈발 부분그래프 쌍들 각각을 대상으로 하여, 빈발 부분그래프 쌍과 상기 다수의 부분시퀀스 트리 각각을 비교하여, 빈발 부분그래프 A와 빈발 부분그래프 B의 순서관계가 만족되는 부분그래프가 최소 지지도 이상 출현하는 경우 상기 빈발 부분그래프 쌍을 상기 유향 동시발생 부분그래프로 판단함으로써 상기 유향 동시발생 부분그래프를 생성하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 장치.

발명의 설명

기술 분야

[0001] 본 발명의 실시예들은 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 생성하는 장치 및 방법에 관한 것이다.

배경 기술

[0002] 그래프(graph)는 소셜 네트워크, XML문서, 화학 복합물, web 등과 같이 객체와 객체간의 관계를 가지고 있는 데이터를 표현한다. 그래프는 고차원적인 특성과 복잡한 구조를 갖기 때문에, 그래프 데이터로부터 유용한 정보를 마이닝(mining)하는 것은 많은 수행시간을 요구한다. 그래프 데이터베이스(DB)로부터 유용한 정보를 찾기 위하여 많은 그래프 마이닝 기술들이 제안되었다.

[0003] 특히, 유향 그래프(directed graph)는 방향성 있는 에지로 연결된 그래프로서, 에지의 방향에 따라 버텍스들의 순서가 정해진다. 유향 그래프는 플로우 차트, 웹 그래프, 소셜 네트워크, 행위 그래프 등과 같은 그래프의 구조적인 성질과 순서정보를 표현하기 위해 많이 사용되고 있다.

[0004] 또한, 동시발생 부분그래프(co-occurrence subgraph)는 그래프 데이터베이스(DB)에서 최소 지지도 이상으로 함께 출현하는 두 개 이상의 부분그래프들 나타내는 패턴을 의미한다. 동시발생 부분그래프는 그래프 데이터베이스의 특성을 나타낼 수 있는 유용한 패턴으로, 그래프의 군집화, 분류, 인덱싱 등에서 활용할 수 있다.

[0005] 유향 그래프 데이터베이스에서의 유향 동시발생 부분그래프는 부분그래프들 간의 순서 정보를 함께 고려해야 한다. 실세계 데이터를 유향 그래프로 모델링할 때 에지의 방향은 중요한 정보이며, 그래프 내에서 부분그래프들을 연결하는 에지의 방향에 따라 부분그래프들의 순서정보가 결정된다. 따라서 동시발생 부분그래프에서 부분그래프들 간의 순서정보가 손실되어서는 안 된다.

[0006] 도 1은 유향 그래프에서의 동시발생 부분그래프의 예를 도시한 도면이다.

[0007] f_1 과 f_2 는 그래프 G_1 과 G_2 에서 빈발하게 나타나는 부분그래프들이다. 부분그래프들 간의 순서를 고려하지 않으면 f_1 과 f_2 의 동시발생 지지도는 2가 된다. 그러나 순서를 고려하면 동시발생 $f_1 \rightarrow f_2$ 와 $f_2 \rightarrow f_1$ 의 지지도는 각각 1이 된다. 따라서 유향 그래프 데이터베이스에서 정확한 의미의 동시발생 패턴을 마이닝하기 위해서는 부분그래프들 간의 순서를 고려해야 한다.

[0008] 다양한 그래프 연구 분야에서 동시발생 부분그래프를 마이닝 기술이 종래에 제안되었다. 그러나 종래의 기술들은 무향 그래프 데이터베이스에서 제안되었기 때문에 유향 그래프 데이터베이스에서는 에지의 방향이라는 중요한 정보를 잃어버리게 되는 단점이 있다.

[0009] 또한, 최근 유향 그래프 데이터베이스의 순서정보를 고려한 빈발 부분그래프 마이닝 기술이 제안되기는 했지만, 방향성을 고려한 동시발생 빈발부분그래프 마이닝 기술은 아직 제안되지 않았다.

발명의 내용

해결하려는 과제

[0010] 상기한 바와 같은 종래기술의 문제점을 해결하기 위해, 본 발명에서는 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 효율적으로 생성할 수 있는 장치 및 방법을 제안하고자 한다.

[0011] 본 발명의 다른 목적들은 하기의 실시예를 통해 당업자에 의해 도출될 수 있을 것이다.

과제의 해결 수단

[0012] 상기한 목적을 달성하기 위해 본 발명의 바람직한 일 실시예에 따르면, 프로세서가 포함된 장치에서, 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 생성하는 방법에 있어서, 상기 다수의 유향 그래프를 대상으로 복수의 빈발 부분그래프를 생성하는 단계; 상기 다수의 유향 그래프 각각의 부분시퀀스의 트리(다수의 부분시퀀스 트리)를 생성하는 단계 - 상기 부분시퀀스는 유향 그래프에 포함되는 복수의 버텍스가 순차적으로 에지로 연결되어 있는 데이터 시퀀스이고, 상기 부분시퀀스 트리는 상기 유향 그래프에 존재하는 부분시퀀스 간의 순서를 나타내는 데이터 트리임 - ; 및 상기 다수의 부분시퀀스 트리를 이용하여 상기 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 상기 유향 동시발생 부분그래프를 생성하는 단계;를 포함하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 방법이 제공된다.

[0013] 상기 부분시퀀스 트리를 구성하는 노드 각각은 하나의 부분시퀀스와 대응되고, 상기 부분시퀀스 트리를 생성하는 단계는, 루트노드를 기준으로 하여 순차적으로 노드를 생성해 가면서 상기 부분시퀀스 트리를 생성하되, 상기 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신함으로써 상기 부분시퀀스 트리를 생성할 수 있다.

[0014] 상기 부분시퀀스 트리를 생성하는 단계는, 상기 노드 중 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 0인 경우, 상기 제1 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신하지 않고 확정하며, 상기 새로운 노드를 생성하지 않고, 상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 1인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 하나의 버텍스를 상기 제1 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, 상기 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 2인 경우, 상기 마지막 버텍스의 외향 에지를 수신하는 2개의 버텍스를 검색하고, 상기 2개의 버텍스 중 하나의 버텍스를 상기 제1 노드의 자식노드인 제2 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, 상기 2개의 버텍스 중 다른 버텍스를 상기 제1 노드의 또 다른 자식노드인 제3 노드와 대응되는 부분시퀀스에 추가하여 갱신할 수 있다.

[0015] 상기 부분시퀀스의 트리를 생성하는 단계는, 상기 유향 그래프의 버텍스들 중 내향 에지가 0인 버텍스를 검색하고, 상기 내향 에지가 0인 버텍스를 상기 부분시퀀스 트리의 루트노드의 자식노드의 첫번째 버텍스로 추가할 수 있다.

[0016] 상기 유향 동시발생 부분그래프를 생성하는 단계는, 상기 복수의 빈발 부분그래프에 대해, 빈발 부분그래프 A와 빈발 부분그래프 B로 구성된 빈발 부분그래프 쌍들을 생성하고, 상기 빈발 부분그래프 쌍들 각각을 대상으로 하여, 빈발 부분그래프 쌍과 상기 다수의 부분시퀀스 트리 각각을 비교하여, 빈발 부분그래프 A와 빈발 부분그래프 B의 순서관계가 만족되는 부분그래프가 최소 지지도 이상 출현하는 경우 상기 빈발 부분그래프 쌍을 상기 유향 동시발생 부분그래프로 판단함으로써 상기 유향 동시발생 부분그래프를 생성할 수 있다.

[0017] 또한 본 발명의 다른 실시예에 따르면, 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 생성하는 장치에 있어서, 상기 다수의 유향 그래프를 대상으로 복수의 빈발 부분그래프를 생성하는 빈발 부분그래프 마이닝부; 상기 다수의 유향 그래프 각각의 부분시퀀스의 트리(다수의 부분시퀀스 트리)를 생성하는 부분시퀀스 트리 생성부 - 상기 부분시퀀스는 유향 그래프에 포함되는 복수의 버텍스가 순차적으로 에지로 연결되어 있는 데이터 시퀀스이고, 상기 부분시퀀스 트리는 상기 유향 그래프에 존재하는 부분시퀀스 간의 순서를 나타내는 데이터 트리임 -; 및 상기 다수의 부분시퀀스 트리를 이용하여 상기 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 상기 유향 동시발생 부분그래프를 생성하는 유향 동시발생 부분그래프 생성부;를 포함하는 것을 특징으로 하는 유향 동시발생 부분그래프 생성 장치가 제공된다.

발명의 효과

[0018] 본 발명에 따르면, 그래프 데이터베이스에 저장되어 있는 다수의 유향 그래프에 대한 유향 동시발생 부분그래프를 효율적으로 생성할 수 있게 된다.

도면의 간단한 설명

[0019] 도 1은 유향 그래프에서의 동시발생 부분그래프의 예를 도시한 도면이다.

도 2는 유향 그래프 내의 부분그래프 순서의 개념을 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 유향 동시발생 부분그래프 생성 장치의 개략적인 구성을 도시한 도면이다.

도 4는 본 발명의 일 실시예에 따른 유향 동시발생 부분그래프 생성 방법의 전체적인 흐름을 도시한 순서도이다.

도 5는 본 발명의 일 실시예에 따른 부분시퀀스 트리의 생성 방법의 개념을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0021] 이하에서, 본 발명에 따른 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다.

[0022] 본 발명의 설명에 앞서, 본 발명에서 정의하고 있는 용어에 대해 간략하게 설명한다.

[0023] 1. 부분그래프 순서(subgraph order)

[0024] 부분그래프 순서는 그래프 G 내의 부분 그래프들 간의 순서를 의미한다. G 에 두 부분그래프 f 와 f' 가 존재할 때, f 의 버텍스들 중에서 f' 의 버텍스로 경로가 존재한다면, f 가 f' 보다 낮은 순서를 갖고, $f < f'$ 로 표기한다. f 와 f' 의 버텍스들 간에 서로 경로가 존재하면, 순환 관계(cycle)로 순서가 없다고 정의한다. 또한 f 와 f' 내에 포함관계가 존재한다면 순서가 없다고 정의한다.

[0025] 도 2는 유향 그래프 내의 부분그래프 순서의 개념을 설명하기 위한 도면이다. 그래프 내에서 점선으로 표기한 원은 부분그래프를 나타낸다. 예를 들어, 부분그래프 f_1 의 버텍스 v_1 에서 f_2 의 v_5 로 향하는 경로(v_1, v_4, v_5)가 존재하기 때문에, f_1 은 f_2 보다 더 낮은 순서를 갖는다. 그리고 f_1 과 f_3 의 경우는 버텍스들 간의 어떠한 경로도 존재하지 않으므로 순서를 가지지 않는다.

[0026] 2. 유향 동시발생 부분그래프(directed co-occurrence subgraph)

[0027] 유향 동시발생 부분그래프는 동시발생 부분그래프 $f_1 f_2 \dots f_m$ 이 존재할 때, 부분그래프 간의 순서가 $f_1 < f_2 < \dots < f_m$ 을 만족하는 동시발생 부분그래프이다.

[0028] 도 2에서 동시발생 부분그래프 $f_1 f_2$ 는 부분그래프간의 순서가 $f_1 < f_2$ 를 만족하므로, $f_1 f_2$ 는 유향 동시발생 부분그래프이다.

[0029] 3. 부분시퀀스(subsequence)

[0030] 부분시퀀스는 유향 레이블 그래프 $G=(V, E, L, l)$ (여기서, V 는 버텍스 집합, E 는 에지의 집합, L 은 레이블 집합, l 은 버텍스와 에지에 레이블을 할당하는 함수를 각각 의미함)가 존재할 때, $V' \subseteq V$ 를 만족하는 V' 내의 복수의 버텍스들이 순차적으로 에지로 연결되어 있는 데이터 시퀀스를 말한다.

[0031] 유향 그래프는 다수의 부분시퀀스들로 구성된다. 이러한 부분시퀀스들은 그래프 내에서 외향 에지가 1인 버텍스의 외향 에지를 따라 탐색한 버텍스들의 시퀀스로 구성된다. 예를 들어, 도 2에서 부분시퀀스는 $\{v_2, v_3\}$, $\{v_4, v_5, v_6\}$ 이 해당된다.

[0032] 4. 부분시퀀스 트리(subsequence tree)

[0033] 부분시퀀스 트리는 유향 레이블 그래프 $G=(V, E, L, l)$ 의 부분 시퀀스 집합 $S=\{s_1, s_2, \dots, s_k\}$ 가 존재할 때, 부분 시퀀스 간의 순서를 나타낸 데이터 트리를 의미한다. 트리의 노드는 부분 시퀀스를 의미하며, 다음 순서의 부분 시퀀스를 자식노드로 생성한다.

[0034] 유향 그래프에서 부분시퀀스 트리는 유향 그래프내의 버텍스들 간의 순서를 모두 보존하면서 그래프를 축약 표현한 구조이다. 이러한 부분시퀀스 트리는 유향 그래프를 탐색하며 추출한 부분시퀀스들을 노드로 생성하며 구

측된다.

[0035] 이하, 상기에서 정의된 용어를 참조하여, 본 발명의 일 실시예를 설명하기로 한다.

[0036] 도 3은 본 발명의 일 실시예에 따른 유향 동시발생 부분그래프 생성 장치의 개략적인 구성을 도시한 도면이다.

[0037] 도 3을 참조하면, 본 발명의 일 실시예에 따른 유향 동시발생 부분그래프 생성 장치(300)는 그래프 데이터베이스(310), 빈발 부분그래프 마이닝부(320), 부분시퀀스 트리 생성부(330) 및 유향 동시발생 부분그래프 생성부(340)를 포함할 수 있다.

[0038] 그리고, 도 4는 본 발명의 일 실시예에 따른 유향 동시발생 부분그래프 생성 방법의 전체적인 흐름을 도시한 순서도이다.

[0039] 이하, 도 3 및 도 4를 참조하여, 각 구성 요소 별 기능 및 각 단계별로 수행되는 과정을 설명하기로 한다.

[0040] 먼저, 단계(410)에서, 그래프 데이터베이스(310)는 데이터 구조를 표현하는 다수의 유향 그래프들을 저장한다.

[0041] 본 발명의 일 실시예에 따르면, 다수의 그래프는 클래스 레이블이 결정된 그래프인 유향 레이블 그래프(labeled graph) 일 수 있다. 여기서, 클래스 레이블은 특정 대상에 대한 그래프의 활성 또는 비활성의 성질을 나타내는 정보를 의미한다. 예를 들어 암의 치료 물질의 경우, 암이라는 특정 대상에서 항암 성질을 가지는 치료물질의 화학 성분 그래프는 활성 클래스로 레이블되고, 항암 성질을 가지지 않는 치료물질의 화학 성분 그래프는 비활성 클래스로 레이블된다. 레이블 그래프는 $G=(V, E, L, l)$ 로 표시될 수 있는데, 이는 4가지 튜플들로 표현된다. 여기서, V 는 버텍스 집합, E 는 에지의 집합, L 은 레이블 집합, l 은 버텍스와 에지에 레이블을 할당하는 함수를 각각 의미한다. 두 연결 그래프 $G'=(V', E', L', l')$ 과 $G=(V, E, L, l)$ 이 존재할 때, V' 내의 모든 버텍스가 V 내에 속하고 E' 내의 모든 에지가 E 에 속한다면 G' 은 G 의 부분그래프라고 한다.

[0042] 또한, 유향 그래프에서, 임베딩은 그래프 G 에서 부분그래프 f 의 위치를 말한다. 예를 들어 도 1에서, 부분그래프 f_1 의 임베딩은 "버텍스 v_1 "과 버텍스 " v_2 " 및 v_1 과 v_2 를 연결하는 "에지"를 의미한다.

[0043] 다음으로, 단계(420)에서, 빈발 부분그래프 마이닝부(320)는 그래프 데이터베이스(310)에 저장되어 있는 유향 그래프에서 최소 지지도 이상으로 출현하는 복수의 빈발 부분그래프를 마이닝(mining)한다. 여기서, 빈발 부분그래프(FG)는 아래의 수학적 식 1와 같이 표현될 수 있다.

수학적 식 1

$$FG = \frac{|g \subseteq G|}{|GD|} > MinSup$$

[0044]

[0045] 여기서, $|g \subseteq G|$ 는 부분그래프(g)를 가지는 그래프($G = \{g_0, g_1, \dots, g_n\}$)의 개수, $|GD|$ 는 그래프 데이터베이스(300)에 존재하는 그래프의 개수, MinSup는 최소 지지도를 의미한다. 이 때, 최소 지지도는 부분그래프가 그래프 데이터베이스(310)에서 나타나는 정도를 의미한다.

[0046] 본 발명의 일 실시예에 따르면, 빈발 부분그래프 마이닝부(320)는 깊이 우선 탐색(depth-first search)과 사전식 순서(lexicographic order)를 이용하여 빈도가 높은 부분그래프를 검색하는 gSpan 방식을 이용하여 빈발 부분그래프를 마이닝할 수 있다. gSpan 방식은 당업자에게 자명한 것이므로, 보다 상세한 설명은 생략하기로 한다.

[0047] 계속하여, 단계(430)에서, 부분시퀀스 트리 생성부(330)는 다수의 유향 그래프 각각의 부분시퀀스의 트리를 생성한다. 즉, 부분시퀀스 트리 생성부(330)는 그래프 데이터베이스(310)로부터 다수의 부분시퀀스 트리를 생성한다. 이 때, 부분시퀀스 트리를 구성하는 노드 각각은 하나의 부분시퀀스와 대응될 수 있다.

[0048] 본 발명의 일 실시예에 따르면, 단계(430)에서, 부분시퀀스 트리 생성부(330)는 루트노드를 기준으로 하여 순차

적으로 노드를 생성해 가면서 부분시퀀스 트리를 생성하되, 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신함으로써 부분시퀀스 트리를 생성할 수 있다.

[0049] 보다 상세하게, 부분시퀀스 트리 생성부(330)는 유한 그래프의 버텍스들 중 내향 에지가 0인 버텍스를 검색하고, 내향 에지가 0인 버텍스를 부분시퀀스 트리의 루트노드의 자식노드의 첫번째 버텍스로 추가할 수 있다.

[0050] 그리고, 부분시퀀스 트리 생성부(330)는 노드 각각에 대해(이하, 설명의 편의를 위해, 다수의 노드 중 해당 노드를 "제1 노드"로 호칭하기로 한다), i) 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 0인 경우, 제1 노드와 대응되는 부분시퀀스에 포함되는 버텍스를 갱신하지 않고 확정하며, 새로운 노드를 생성하지 않고, ii) 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 1인 경우, 마지막 버텍스의 외향 에지를 수신하는 하나의 버텍스를 제1 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, iii) 제1 노드와 대응되는 부분시퀀스의 마지막 버텍스의 외향 에지의 개수가 2인 경우, 마지막 버텍스의 외향 에지를 수신하는 2개의 버텍스를 검색하고, 2개의 버텍스 중 하나의 버텍스를 제1 노드의 자식노드인 제2 노드와 대응되는 부분시퀀스에 추가하여 갱신하고, 2개의 버텍스 중 다른 버텍스를 제1 노드의 또 다른 자식노드인 제3 노드와 대응되는 부분시퀀스에 추가하여 갱신할 수 있다.

[0051] 즉, 유한 그래프에서 내향 에지가 0인 버텍스 v_s 를 찾고 부분시퀀스 트리의 루트노드에 자식노드로 추가한다. 그리고, v_s 의 외향 에지 정도가 1인 경우 인접한 버텍스도 v_s 와 같은 부분시퀀스에 속하므로 트리의 현재 노드의 부분시퀀스를 갱신한다. 반대의 경우, 인접한 버텍스가 같은 부분시퀀스에 속하지 않으므로 부분시퀀스 트리에 새로운 노드를 추가하고 인접한 버텍스를 새로운 노드의 부분시퀀스로 추가한다. 그리고 만약, 탐색중인 버텍스가 다른 부분시퀀스에 속하는 경우, 트리에 새로운 노드를 추가하여 탐색중인 버텍스에 대한 부분시퀀스를 새롭게 갱신한다. 이와 같은 과정을 반복하여 유한 그래프의 모든 버텍스에 대하여 탐색하며 부분시퀀스 트리의 노드를 구축한다.

[0052] 예를 들어, 도 5를 참조하면, 우선, 내향 에지가 0인 버텍스 v_1 을 찾고, 부분시퀀스 트리의 루트노드의 자식노드에 v_1 을 추가한다. 그리고, v_1 과 인접한 v_2 를 루트노드의 자식노드에 추가하여 부분시퀀스를 갱신한다. v_2 의 외향 에지 정도가 2이므로 트리에 새로운 노드 v_2 를 추가한 후, v_2 와 인접한 v_3 를 v_2 에 추가한다. 이와 같은 과정을 반복하면 트리에 노드 n_1 , n_2 , n_3 가 생성된다. 마지막으로 v_7 을 탐색할 때, v_7 이 n_2 와 n_3 의 부분시퀀스에 해당하므로 새로운 노드를 생성하여 v_7 을 추가한다.

[0053] 표 1은 본 발명의 일 실시예에 따른 부분시퀀스 트리의 생성 알고리즘을 나타내고 있다.

❧ 1

Algorithm 2. GenerateSubsequenceTree(G, T)

Input :

G : a graph $G = (V, L, E, l)$

Output :

T : a subsequence tree

1. **WHILE** all vertices are not marked
2. V_s = find the 0-in edge degree vertex
3. **IF** V_s is $\emptyset$
4. choose the vertex v in not marked vertices
5. V_s is marked
6. Maintain the vertex V_s as the node n of the tree T
7. TraverseAllVertices(G, T, V_s)
8. **RETURN** T

Subprocedure 1. TraverseAllVertices(G, T, V_s)

1. **FOR** $v \in$ neighbour vertices of V_s
 2. **IF** out edge degree of v is one
 3. add v to n
 4. **ELSE**
 5. make a node n' and set n' to child of n
 6. Maintain the v as the node n'
 7. v is marked
 8. TraverseAllVertices(G, T, v)
 9. **RETURN** T
-

[0054]

[0055]

표 1을 참조하면, 알고리즘은 그래프 G 를 입력 받고 출력 값으로 부분 시퀀스 트리 T 를 반환한다. 그래프 내에서 내향 에지 정도(in edge degree)가 0인 버텍스를 시작 버텍스 v_s 로 설정한다(line2-4). 시작 버텍스 v_s 와 인접한 버텍스를 깊이 우선으로 탐색하며 그래프 G 에 대한 부분시퀀스 트리 T 를 생성한다. 버텍스의 외향 에지 정도가 1인 경우 시퀀스를 의미하므로 현재 트리의 노드에 버텍스를 추가한다(subprocedure 1, line 2-3). 반대의 경우, 트리에 새로운 노드를 추가하고 현재 노드의 자식으로 설정한다(subprocedure 1, line 4-7). 이와 같은 과정을 더 이상 탐색할 노드가 없을 때까지 반복하여 트리를 생성한다.

[0056]

마지막으로, 단계(440)에서, 유향 동시발생 부분그래프 생성부(340)는 다수의 부분시퀀스 트리를 이용하여 복수의 빈발 부분그래프 간의 순서관계를 찾고, 상기 순서관계를 이용하여 유향 동시발생 부분그래프를 생성한다. 본 발명의 일 실시예에 따르면, 유향 동시발생 부분그래프 생성부(340)는, 복수의 빈발 부분그래프에 대해, 빈발 부분그래프 A와 빈발 부분그래프 B로 구성된 빈발 부분그래프 쌍들을 생성한다. 그리고, 빈발 부분그래프 쌍들 각각을 대상으로 하여, 빈발 부분그래프 쌍과 다수의 부분시퀀스 트리 각각을 비교하여, 빈발 부분그래프 A와 빈발 부분그래프 B의 순서관계가 만족되는 부분그래프가 최소 지지도 이상 출현하는 경우, 빈발 부분그래프 쌍을 유향 동시발생 부분그래프로 판단함으로써 유향 동시발생 부분그래프를 생성한다.

[0057]

보다 상세하게, 빈발 부분그래프의 순서를 효율적으로 비교하기 위해 유향 그래프 데이터베이스(310)의 모든 그래프에 대해서 부분 시퀀스 트리를 생성한다. 그리고 빈발 부분그래프 마이닝 기술을 사용하여 모든 그래프부

터 최소 지지도 이상으로 출현하는 빈발 부분그래프를 마이닝한다. 이 때, 빈발 부분그래프는 임베딩 정보에 따라 순서가 다르게 표현되기 때문에 모든 임베딩 정보를 유지해야 한다. 부분 시퀀스 트리는 그래프의 순서 정보를 유지하면서 그래프의 노드와 에지를 최대한으로 축약하였으므로, 부분 시퀀스 트리의 노드를 통하여 부분그래프의 임베딩 정보를 표현하면 많은 양의 저장 공간과 연결성 검사를 위한 많은 비용을 절감할 수 있다.

[0058] 또한, 빈발 부분그래프를 마이닝할 때, 빈발 부분그래프의 임베딩 정보를 부분 시퀀스 트리의 노드를 통해 표현한다. 빈발 부분그래프의 모든 버텍스에 대해서, 빈발 부분그래프를 가지고 있는 그래프의 부분시퀀스 트리에 대응하는 노드를 탐색한다. 이와 같은 과정을 모든 빈발 부분그래프를 가지고 있는 모든 그래프들에 대해서 반복하여 임베딩 정보를 유지한다.

[0059] 그리고, 순서 정보를 고려하여 동시발생 부분그래프를 마이닝하는 방법은 다음과 같다. 빈발 부분그래프에서 두 빈발 부분그래프 A와 B를 선택한다. 그리고 빈발 부분그래프 A와 B를 동시에 가지고 있는 그래프의 수가 최소 지지도를 만족하는지 확인한다. 순서 정보를 고려하기 위하여, 빈발 부분그래프 A와 B를 동시에 가지고 있는 그래프에 해당하는 부분시퀀스 트리와 빈발 부분그래프 A와 B의 임베딩 정보를 통해 순서를 비교한다. 빈발 부분그래프 A와 B가 $A < B$ 의 순서를 만족하는 부분그래프가 최소 지지도 이상 출현할 경우 유향 동시발생 부분그래프로 마이닝한다.

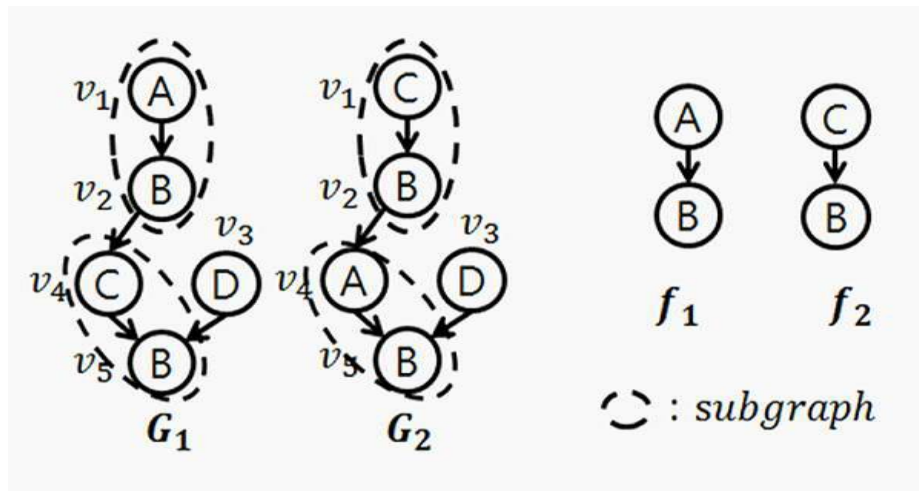
[0060] 정리하면, 부분그래프 간의 순서를 계산하기 위한 비용을 줄이기 위하여, 유향 그래프를 순서를 고려하여 여러 개의 부분 영역으로 나누어 산출한다. 즉, 빈발 부분그래프의 임베딩 정보를 부분 영역을 통해 압축 표현한다면, 임베딩 정보를 적재하기 위한 메모리 사용량을 절약할 수 있으며, 부분영역간의 연결성 검사를 통해 연산량을 줄일 수 있다.

[0061] 또한, 본 발명의 실시예들은 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 본 발명을 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), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 본 발명의 일 실시예들의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

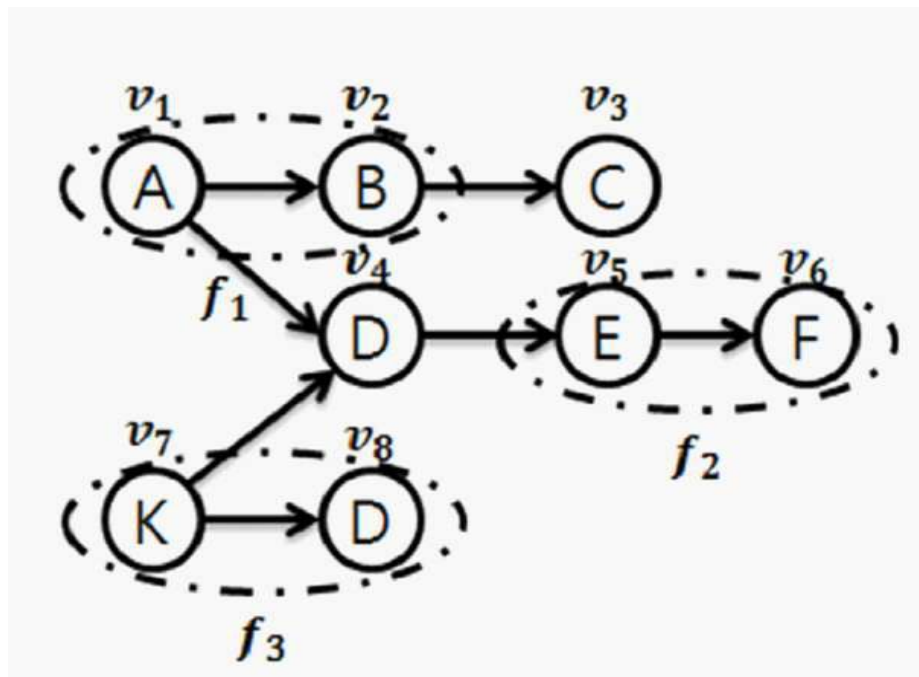
[0062] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나 이는 본 발명의 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다. 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

도면

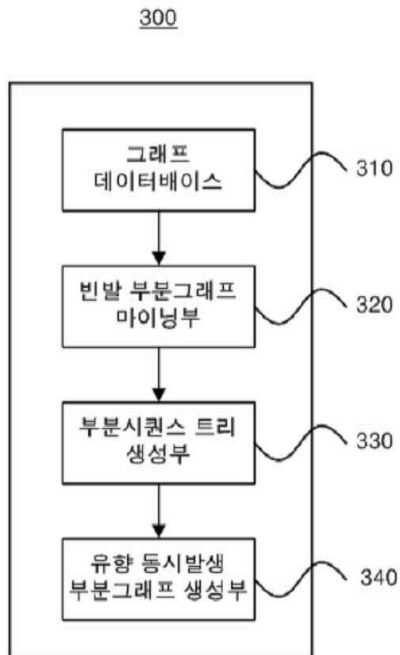
도면1



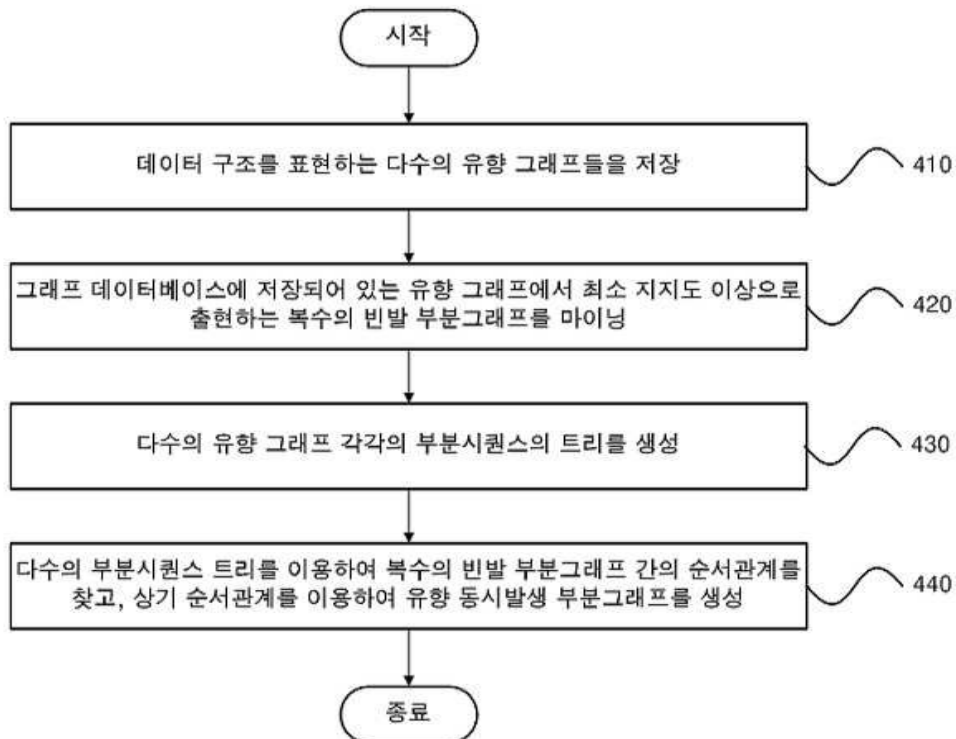
도면2



도면3



도면4



도면5

