(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0079377
(43) 공개일자 2015년07월08일

(51) 국제특허분류(Int. Cl.)
G06F 17/00 (2006.01) *G06F 19/00* (2011.01)
(21) 출원번호 10-2014-0091073
(22) 출원일자 2014년07월18일
심사청구일자 2014년07월18일
(30) 우선권주장
1020130165503 2013년12월27일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
이영구
대전 서구 청사서로 11, 106동 302호 (월평동, 무지개아파트)
한용구
경기 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)
박기성
경기 부천시 원미구 원미로 123-4, E동 4호 (원미동, 백송연립)
(74) 대리인
서재승

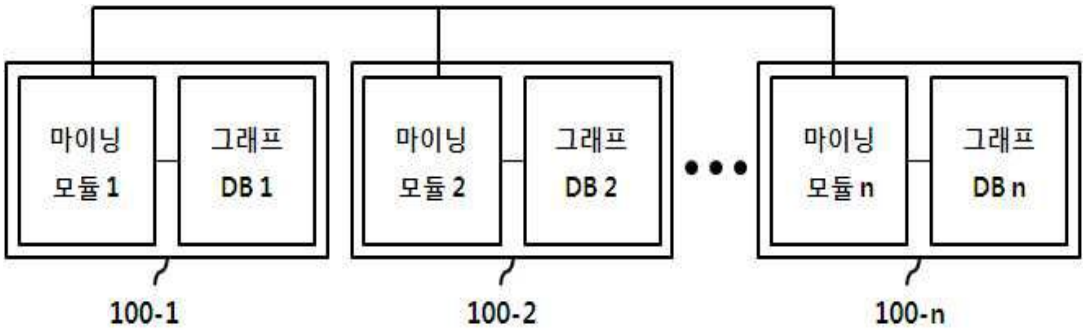
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **맵리듀스를 이용한 빈발 부분 그래프의 마이닝 방법**

(57) 요 약

본 발명은 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법에 관한 것으로, 보다 구체적으로 맵리듀스 기법을 이용하여 대용량의 그래프 데이터베이스를 분산 처리하여 대용량의 그래프 데이터베이스에서 빈발 부분 그래프를 빠르게 마이닝할 수 있으며, 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하고 복원 그래프로부터 에지 확장 가능한 후보 부분 그래프만을 생성하여 후보 부분 그래프를 생성하는데 소요되는 연산량과 시간을 줄일 수 있는 빈발 부분 그래프의 마이닝 방법에 관한 것이다.

대 표 도 - 도4



이 발명을 지원한 국가연구개발사업
과제고유번호 1415123336
부처명 지식경제부
연구관리전문기관 정보통신산업진흥원
연구사업명 대학IT연구
연구과제명 동서신의학 u-라이프케어 기술 연구
기 여 율 1/1
주관기관 경희대학교산학협력단
연구기간 2013.01.01 ~ 2013.12.31

명 세 서

청구범위

청구항 1

대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 장치에 있어서,

다수개의 그래프 그룹으로 분리되어 있는 상기 대용량 그래프 데이터베이스 중 1개의 그래프 그룹이 분산 저장되어 있는 그래프 데이터베이스; 및

상기 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 빈발 부분 그래프의 복원 그래프를 생성하여 상기 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하며, 상기 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 마이닝 모듈을 포함하는 것을 특징으로 하는 그래프 데이터베이스의 마이닝 장치.

청구항 2

제 1 항에 있어서, 상기 마이닝 모듈은

상기 그래프 그룹에서 최소 지지도를 만족하는 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하며, 상기 복원 그래프에 기초하여 상기 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하는 에지 확장부; 및

각 그래프에서 동일한 정규 코드를 가지는 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 생성부를 포함하며,

상기 에지 확장부와 상기 생성부는 상기 확장 빈발 부분 그래프로부터 신규 확장 빈발 부분 그래프가 생성되지 않을 때까지 상기 확장 빈발 부분 그래프의 에지를 확장하여 상기 그래프 데이터베이스의 빈발 부분 그래프를 생성하는 것을 특징으로 하는 그래프 데이터베이스의 마이닝 장치.

청구항 3

제 2 항에 있어서, 상기 에지 확장부는

상기 그래프 그룹에서 최소 지지도를 만족하며 k(k는 자연수)개 에지를 가지는 빈발 부분 그래프가 입력되는 경우, 상기 빈발 부분 그래프와 상기 빈발 부분 그래프가 포함된 그래프 식별자로 출력 데이터를 생성하는 에지 확장 맵퍼; 및

상기 출력 데이터의 그래프 식별자별로 구분하여 상기 빈발 부분 그래프의 복원 그래프를 생성하며, 상기 복원 그래프에 기초하여 상기 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성하는 에지 확장 리듀서를 포함하는 것을 특징으로 하는 그래프 데이터베이스의 마이닝 장치.

청구항 4

제 3 항에 있어서, 상기 에지 확장 리듀서는

동일한 그래프 식별자를 가지는 상기 빈발 부분 그래프에서 상기 빈발 부분 그래프를 구성하는 버텍스와 에지의 연결 관계에 기초하여 복원 그래프를 생성하는 것을 특징으로 하는 그래프 데이터베이스의 마이닝 장치.

청구항 5

제 3 항에 있어서, 상기 에지 확장 리듀서는

상기 복원 그래프와 상기 빈발 부분 그래프의 버텍스 및 에지의 연결 관계를 비교하여, 상기 빈발 부분 그래프의 버텍스로부터 연결된 에지가 존재하는지 여부로 후보 부분 그래프가 존재하는지 판단하는 것을 특징으로 하는 그래프 데이터베이스의 마이닝 장치.

청구항 6

제 3 항에 있어서, 상기 생성부는

상기 후보 부분 그래프를 정규화 연산하여 상기 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단하여 상기 그래프 식별자별 상기 정규 후보 부분 그래프를 판단하는 카운팅 맵퍼; 및

각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 카운팅 리듀서를 포함하는 것을 특징으로 하는 데이터베이스 마이닝 장치.

청구항 7

제 6 항에 있어서, 상기 카운팅 리듀서는

상기 대용량 그래프 데이터베이스 중 다른 그래프 그룹이 분산 저장되어 있는 주변 데이터베이스 마이닝 장치로부터 상기 다른 그래프 그룹의 그래프 식별자별 정규 후보 부분 그래프를 수신하며,

상기 그래프 그룹과 상기 다른 그래프 그룹에서 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지지도를 만족하는 확장 빈발 부분 그래프를 생성하는 것을 특징으로 하는 데이터베이스 마이닝 장치.

청구항 8

대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법에 있어서,

상기 대용량 그래프 데이터베이스는 다수개의 그래프 그룹으로 분리되어 분산되어 있으며,

k(k는 자연수)개의 에지를 가지는 빈발 부분 그래프가 입력되는 경우 상기 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 상기 빈발 부분 그래프의 복원 그래프를 생성하여 상기 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성하는 단계; 및

상기 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 k+1개의 에지를 가지는 확장 빈발 부분 그래프를 생성하는 단계를 포함하며,

상기 확장 빈발 부분 그래프가 입력되어 k+2개의 에지를 가지는 신규 확장 빈발 부분 그래프가 생성되지 않을 때까지 상기 확장 빈발 부분 그래프의 에지를 확장하여 상기 그래프 데이터베이스의 빈발 부분 그래프를 반복 생성하는 것을 특징으로 하는 빈발 부분 그래프의 마이닝 방법.

청구항 9

제 8 항에 있어서, 상기 후보 부분 그래프를 생성하는 단계는

k개의 에지를 가지는 빈발 부분 그래프가 입력되는 단계;

상기 빈발 부분 그래프가 포함된 그래프 식별자를 키 값으로 하여 상기 빈발 부분 그래프를 구비하는 출력 데이터로 생성하는 단계;

동일한 그래프 식별자를 가지는 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하는 단계; 및

상기 복원 그래프에 기초하여 상기 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하는 단계를 포함하는 것을 특징으로 하는 빈발 부분 그래프의 마이닝 방법.

청구항 10

제 9 항에 있어서,

상기 복원 그래프는 동일한 그래프 식별자를 가지는 상기 빈발 부분 그래프에서 상기 빈발 부분 그래프를 구성하는 버텍스와 에지의 연결 관계에 기초하여 생성되는 것을 특징으로 하는 빈발 부분 그래프의 마이닝 방법.

청구항 11

제 10 항에 있어서,

상기 후보 부분 그래프는 상기 복원 그래프와 상기 빈발 부분 그래프의 버텍스 및 에지의 연결 관계를 비교하여, 상기 빈발 부분 그래프의 버텍스로부터 연결된 에지가 존재하는지 경우 생성되는 것을 특징으로 하는

빈발 부분 그래프의 마이닝 방법.

청구항 12

제 9 항에 있어서, 상기 확장 빈발 부분 그래프를 생성하는 단계는

상기 후보 부분 그래프를 정규화 연산하여 상기 후보 부분 그래프의 정규 코드를 계산하는 단계;

상기 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단하는 단계; 및

각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 단계를 포함하는 것을 특징으로 하는 빈발 부분 그래프의 마이닝 방법.

청구항 13

제 12 항에 있어서, 상기 확장 빈발 부분 그래프를 생성하는 단계는

상기 대용량 그래프 데이터베이스 중 다른 그래프 그룹이 분산 저장되어 있는 주변 데이터베이스 마이닝 장치로부터 상기 다른 그래프 그룹의 그래프 식별자별 정규 후보 부분 그래프를 수신하는 단계를 더 포함하며,

상기 그래프 그룹과 상기 다른 그래프 그룹에서 각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는 확장 빈발 부분 그래프를 생성하는 것을 특징으로 하는 빈발 부분 그래프의 마이닝 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법에 관한 것으로, 보다 구체적으로 맵리듀스(MapReduce) 기법을 이용하여 대용량의 그래프 데이터베이스를 분산 처리하여 대용량의 그래프 데이터베이스에서 빈발 부분 그래프를 빠르게 마이닝할 수 있으며, 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하고 복원 그래프로부터 에지 확장 가능한 후보 부분 그래프만을 생성하여 후보 부분 그래프를 생성하는데 소요되는 연산량과 시간을 줄일 수 있는 빈발 부분 그래프의 마이닝 방법에 관한 것이다.

배 경 기 술

[0002] 페타(peta) 바이트 이상의 규모의 대용량 데이터 분석은 다양한 분야에서 연구되고 있다. 특히, 소셜 네트워크, XML 등과 같은 구조적인 정보를 갖는 대용량의 그래프들을 분석하는 기술이 활발히 연구되고 있다. 예를 들어, 소셜 네트워크에서 사용자 식별자는 노드 버텍스로, 사용자 사이의 관계는 노드 버텍스를 연결하는 노드 에지로, 사용자 관계의 긴밀도 또는 거리는 노드 가중치로 표현될 수 있다.

[0003] 이러한 대용량의 그래프데이터로부터 유용한 지식을 찾아내는 그래프 마이닝 기술이 주목받고 있는데, 그래프 마이닝은 화합물, XML, 웹 문서, 소셜 네트워크 등과 같은 그래프 구조를 갖는 데이터를 분류하는 기술이다.

[0004] 그래프 마이닝을 위해 구조적 정보를 갖는 빈발 부분 그래프들이 특징으로 널리 사용되고 있다. 여기서 빈발 부분 그래프란 그래프 데이터베이스에서 최소 지지도 이상으로 발현되는 부분 그래프를 의미하는데, 빈발 부분 그래프는 그래프의 고유한 특성을 나타내어 그래프의 분류, 군집화, 인덱싱 등에서 사용된다. 예를 들어, 생명정보학에서 빈발 부분 그래프는 그래프로 표현된 화학 합성물 데이터 가운데 빈발하게 발생하는 분절을 검색하는데 사용된다.

[0005] 빈발 부분 그래프를 마이닝하는 알고리즘은 그래프의 복잡한 구조적 특성 때문에 오랜 수행 시간이 소요되며, 대용량의 메모리 공간을 요구한다. 따라서 대용량의 그래프 데이터베이스의 경우 크기가 증가할수록 빈발 부분 그래프를 마이닝하는데 더욱 오랜 시간이 소요된다는 문제점을 가진다.

[0006] 따라서 맵리듀스(Map-Reduce)에 기반하여 대용량 그래프 데이터베이스를 다수의 그래프 그룹으로 분할하고 각 그래프 그룹을 분산 처리하는 방식이 제안되고 있다.

[0007] 도 1은 다수의 그래프 그룹으로 이루어진 대용량 그래프 데이터베이스의 일 예를 도시하고 있다.

[0008] 도 1을 참고로 살펴보면, 원으로 표현된 다수의 노드 버텍스와 노드 버텍스 사이를 연결하는 에지로 구성되어 있는 대용량 그래프 데이터베이스가 존재한다. 여기서 노드 버텍스와 노드 버텍스를 연결하는 에지에는 가중치가 부여되어 있는데, 노드 버텍스는 소셜 네트워크의 경우 사용자 이름, 아이디 등과 같은 사용자 식별자이며, 노드 에지는 사용자의 관계를 의미하며, 가중치는 노드와 노드 사이의 거리, 친밀도 등을 의미한다.

[0009] 대용량 그래프 데이터베이스는 다수의 그래프 그룹(GG1, GG2, GG3)으로 구분되어 그래프 그룹 단위로 빈발 부분 그래프를 마이닝하는 장치에 분산 저장되어 있다.

[0010] 도 2는 종래 맵리듀스 기반의 빈발 부분 그래프의 마이닝 방법(이하 종래 기술)을 설명하기 위한 도면이다.

[0011] 종래 기술에 따르면, 대용량의 그래프 데이터베이스는 다수의 그래프 그룹으로 분리되어 분산되어 있으며 각 그래프 그룹별로 빈발 부분 그래프를 분산 마이닝하는데, 도 2와 같이 그래프 그룹1(GG1)에 3개의 그래프(G1, G2, G3)가 존재하는 경우, 그래프 그룹1에서 빈발 부분 그래프를 마이닝한다. 빈발 부분 그래프(frequent subgraph)는 그래프 그룹에서 최소 지지도 이상으로 출현하는 부분 그래프로서, 빈발 부분 그래프는 그래프의 고유한 특성을 나타내어 그래프의 분류, 군집화, 인덱싱 등에서 사용된다. 빈발 부분 그래프 마이닝은 그래프 그룹에서 모든 빈발 부분 그래프를 찾는 기법으로, 복잡한 구조를 가지는 그래프들이 대용량으로 저장되어 있는 그래프 데이터베이스로부터 최소 지지도 이상으로 출현하는 모든 부분 그래프를 검색하는 과정을 의미한다.

[0012] 도 3에 도시되어 있는 바와 같이, 종래기술에서는 그래프 그룹1에서 3개의 빈발 부분 그래프(SG1, SG2, SG3)가 최소 지지도 2로 출현하는 경우, 1개의 에지를 가지는 전체 빈발 부분 그래프 중에서 2개의 빈발 부분 그래프를 선택하여, 즉 (SG1, SG2), (SG1, SG2), (SG2, SG1), (SG2, SG3), (SG3, SG1), (SG3, SG2)로 이루어진 모든 후보 부분 그래프를 생성하기 때문에, 대량의 빈발하지 않은 후보 부분 그래프가 생성된다. 따라서 대량의 빈발하지 않은 후보 부분 그래프를 생성하는데 많은 연산량이 소요되어 대용량의 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는데 오랜 시간이 걸리며 빈발 부분 그래프를 마이닝하는데 많은 비용이 소요된다는 문제점을 가진다.

발명의 내용

해결하려는 과제

[0013] 본 발명은 종래 기술이 가지는 문제점을 해결하기 위한 것으로, 본 발명이 이루고자 하는 목적은 맵리듀스 기법을 이용하여 대용량의 그래프 데이터베이스를 분산 처리하여 대용량의 그래프 데이터베이스에서 빈발 부분 그래프를 빠르게 마이닝할 수 있는 방법을 제공한다.

[0014] 한편 본 발명이 이루고자 하는 다른 목적은 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하고 복원 그래프로부터 에지 확장 가능한 후보 부분 그래프만을 생성하기 때문에, 후보 부분 그래프를 생성하는데 소요되는 연산량과 시간을 줄일 수 있는 빈발 부분 그래프의 마이닝 방법을 제공하는 것이다.

[0015] 본 발명이 이루고자 하는 또 다른 목적은 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프만을 고려하여 확장 빈발 부분 그래프를 카운트하기 때문에, 최소 지지도를 만족하는 확장 빈발 부분 그래프를 정확하고 빠르게 마이닝할 수 있는 빈발 부분 그래프의 마이닝 방법을 제공하는 것이다.

과제의 해결 수단

[0016] 본 발명의 목적을 달성하기 위하여, 본 발명에 따른 빈발 부분 그래프의 마이닝 장치는 다수개의 그래프 그룹으로 분리되어 있는 대용량 그래프 데이터베이스 중 1개의 그래프 그룹이 분산 저장되어 있는 그래프 데이터베이스와, 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 빈발 부분 그래프의 복원 그래프를 생성하여 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하며, 상기 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 상기 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 마이닝 모듈을 포함하는 것을 특징으로 한다.

[0017] 여기서 마이닝 모듈은 그래프 그룹에서 최소 지지도를 만족하는 빈발 부분 그래프로부터 그래프 식별자별로 복

원 그래프를 생성하며, 복원 그래프에 기초하여 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하는 에지 확장부와, 각 그래프에서 동일한 정규 코드를 가지는 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 생성부를 포함한다.

[0018] 바람직하게, 에지 확장부와 생성부는 확장 빈발 부분 그래프로부터 신규 확장 빈발 부분 그래프가 생성되지 않을 때까지 확장 빈발 부분 그래프의 에지를 확장하여 그래프 데이터베이스의 빈발 부분 그래프를 생성하는 것을 특징으로 한다.

[0019] 여기서 에지 확장부는 그래프 그룹에서 최소 지지도를 만족하며 k(k는 자연수)개 에지를 가지는 빈발 부분 그래프가 입력되는 경우, 빈발 부분 그래프와 빈발 부분 그래프가 포함된 그래프 식별자로 출력 데이터를 생성하는 에지 확장 맵퍼와, 출력 데이터의 그래프 식별자별로 구분하여 빈발 부분 그래프의 복원 그래프를 생성하며 복원 그래프에 기초하여 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성하는 에지 확장 리듀서를 포함하는 것을 특징으로 한다.

[0020] 바람직하게 에지 확장 리듀서는 동일한 그래프 식별자를 가지는 빈발 부분 그래프에서 빈발 부분 그래프를 구성하는 버텍스와 에지의 연결 관계에 기초하여 복원 그래프를 생성하는 것을 특징으로 한다.

[0021] 바람직하게 에지 확장 리듀서는 복원 그래프와 빈발 부분 그래프의 버텍스 및 에지의 연결 관계를 비교하여, 빈발 부분 그래프의 버텍스로부터 연결된 에지가 존재하는지 여부로 후보 부분 그래프가 존재하는지 판단하는 것을 특징으로 한다.

[0022] 여기서 생성부는 후보 부분 그래프를 정규화 연산하여 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단하여 그래프 식별자별 정규 후보 부분 그래프를 판단하는 카운팅 맵퍼와, 각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 카운팅 리듀서를 포함하는 것을 특징으로 한다.

[0023] 바람직하게, 카운팅 리듀서는 대용량 그래프 데이터베이스 중 다른 그래프 그룹이 분산 저장되어 있는 주변 데이터베이스 마이닝 장치로부터 다른 그래프 그룹의 그래프 식별자별 정규 후보 부분 그래프를 수신하며, 그래프 그룹과 다른 그래프 그룹에서 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지지도를 만족하는 확장 빈발 부분 그래프를 생성하는 것을 특징으로 한다.

[0024] 한편, 본 발명의 목적을 달성하기 위하여 본 발명에 따른 빈발 부분 그래프의 마이닝 방법은 대용량 그래프 데이터베이스는 다수개의 그래프 그룹으로 분리되어 분산되어 있으며, k(k는 자연수)개의 에지를 가지는 빈발 부분 그래프가 입력되는 경우 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 빈발 부분 그래프의 복원 그래프를 생성하여 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성하는 단계와, 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 k+1개의 에지를 가지는 확장 빈발 부분 그래프를 생성하는 단계를 포함하며, 확장 빈발 부분 그래프가 입력되어 k+2개의 에지를 가지는 신규 확장 빈발 부분 그래프가 생성되지 않을 때까지 확장 빈발 부분 그래프의 에지를 확장하여 그래프 데이터베이스의 빈발 부분 그래프를 반복 생성하는 것을 특징으로 한다.

[0025] 여기서 후보 부분 그래프를 생성하는 단계는 k개의 에지를 가지는 빈발 부분 그래프가 입력되는 단계와, 빈발 부분 그래프가 포함된 그래프 식별자를 키 값으로 하여 빈발 부분 그래프를 구비하는 출력 데이터로 생성하는 단계와, 동일한 그래프 식별자를 가지는 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하는 단계와, 복원 그래프에 기초하여 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하는 단계를 포함한다.

[0026] 여기서 확장 빈발 부분 그래프를 생성하는 단계는 후보 부분 그래프를 정규화 연산하여 후보 부분 그래프의 정규 코드를 계산하는 단계와, 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단하는 단계와, 각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 단계를 포함한다.

[0027] 바람직하게 확장 빈발 부분 그래프를 생성하는 단계는 대용량 그래프 데이터베이스 중 다른 그래프 그룹이 분산 저장되어 있는 주변 데이터베이스 마이닝 장치로부터 다른 그래프 그룹의 그래프 식별자별 정규 후보 부분 그래프를 수신하는 단계를 더 포함하며, 그래프 그룹과 다른 그래프 그룹에서 각 그래프에 존재하는 동일한 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지지도를 만족하는 확장 빈발 부분 그래프를 생성하는 것을

특징으로 한다.

발명의 효과

[0028] 본 발명에 따른 빈발 부분 그래프의 마이닝 방법은 맵리듀스 기법을 이용하여 대용량의 그래프 데이터베이스를 분산 처리함으로써, 대용량의 그래프 데이터베이스에서 빈발 부분 그래프를 빠르게 마이닝할 수 있다.

[0029] 또한, 본 발명에 따른 빈발 부분 그래프의 마이닝 방법은 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하고 복원 그래프로부터 에지 확장 가능한 후보 부분 그래프만을 생성함으로써, 후보 부분 그래프를 생성하는데 소요되는 연산량과 시간을 줄일 수 있다.

[0030] 또한, 본 발명에 따른 빈발 부분 그래프의 마이닝 방법은 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프만을 고려하여 확장 빈발 부분 그래프를 카운트함으로써, 최소 지지도를 만족하는 확장 빈발 부분 그래프를 정확하고 빠르게 마이닝할 수 있다.

도면의 간단한 설명

[0031] 도 1은 다수의 그래프 그룹으로 이루어진 대용량 그래프 데이터베이스의 일 예를 도시하고 있다.

도 2는 종래 맵리듀스 기반의 빈발 부분 그래프의 마이닝 방법을 설명하기 위한 도면이다.

도 3은 빈발 부분 그래프의 일 예를 설명하기 위한 도면이다.

도 4는 본 발명에 따른 빈발 부분 그래프의 마이닝 장치를 설명하기 위한 기능 블록도이다.

도 5는 본 발명의 일 실시예에 따른 마이닝 모듈을 설명하기 위한 기능 블록도이다.

도 6은 본 발명에 따른, 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법을 설명하기 위한 흐름도이다.

도 7은 후보 부분 그래프를 생성하는 단계를 보다 구체적으로 설명하기 위한 흐름도이다.

도 8은 본 발명에 따른 확장 빈발 부분 그래프를 마이닝하는 단계를 보다 구체적으로 설명하기 위한 흐름도이다.

도 9 내지 도 11은 도 2에 도시되어 있는 그래프 그룹에서 본 발명에 따른 빈발 부분 그래프의 마이닝 방법의 일 예를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0032] 이하 첨부한 도면을 참고로 본 발명에 따른, 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법 및 그 장치에 대해 설명한다.

[0033] 도 4는 본 발명에 따른 빈발 부분 그래프의 마이닝 장치를 설명하기 위한 기능 블록도이다.

[0034] 도 4를 참고로 보다 구체적으로 살펴보면, 대용량 그래프 데이터베이스는 다수의 그래프 그룹으로 분류되어 있으며, 다수의 그래프 그룹 중 적어도 1개 이상의 그래프 그룹이 각 마이닝 장치(100-1, 100-2,..., 100-n)의 그래프 데이터베이스에 분산 저장되어 있다.

[0035] 각 마이닝 장치(100)는 각 마이닝 장치의 그래프 데이터베이스(DB)에 분산 저장되어 있는 그래프 그룹에서 후보 부분 그래프를 생성하고 각 마이닝 장치(100)에서 생성한 후보 부분 그래프의 정규 코드에 기초하여 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는데, 각 마이닝 장치(100)의 마이닝 모듈은 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 빈발 부분 그래프의 복원 그래프를 생성하여 k개의 에지를 가지는 빈발 부분 그래프로부터 에지 확장 가능한 k+1개의 에지를 가지는 후보 부분 그래프를 생성하며, 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한, k+1개의 에지를 가지는 확장 빈발 부분 그래프를 마이닝한다.

[0036] 확장 빈발 부분 그래프로부터 다시 k+2개의 에지를 가지는 정규 후보 부분 그래프를 생성하여 정규 후보 부분 그래프의 수가 최소 지지도를 만족하는 경우 확장 빈발 부분 그래프로 마이닝한다. 이를 반복하여 대용량 그래프 데이터베이스에 존재하는 모든 빈발 부분 그래프를 마이닝한다.

[0037] 도 5는 본 발명의 일 실시예에 따른 마이닝 모듈을 설명하기 위한 기능 블록도이다.

[0038] 도 5를 참고로 보다 구체적으로 살펴보면, 마이닝 모듈은 그래프 데이터베이스에 저장되어 있는 그래프 그룹에서 최소 지지도를 만족하는 빈발 부분 그래프로부터 그래프 식별자별로 복원 그래프를 생성하며 복원 그래프에 기초하여 빈발 부분 그래프로부터 에지 확장 가능한 후보 부분 그래프를 생성하는 에지 확장부(110)와, 동일한 정규 코드의 정규 후보 부분 그래프가 발현하는 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성하는 생성부(130)를 포함한다.

[0039] 본 발명의 일 실시예에 따른 에지 확장부(110)를 보다 구체적으로 살펴보면, 에지 확장부(110)는 에지 확장 맵퍼(111)와 에지 확장 리듀서(113)을 구비하는데 에지 확장 맵퍼(111)는 그래프 그룹에서 최소 지지도를 만족하며 k개 에지를 가지는 빈발 부분 그래프가 입력되는 경우 빈발 부분 그래프와 빈발 부분 그래프가 포함된 그래프 식별자로 출력 데이터를 생성한다. 에지 확장 리듀서(113)는 출력 데이터를 그래프 식별자별로 구분하여 빈발 부분 그래프의 복원 그래프를 생성하며, 복원 그래프에 기초하여 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성한다.

[0040] 한편, 본 발명의 일 실시예에 따른 생성부(130)를 보다 구체적으로 살펴보면, 생성부(130)는 카운팅 맵퍼(131)와 카운팅 리듀서(133)를 구비하는데 카운팅 맵퍼(131)는 후보 부분 그래프를 정규화 연산하여 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단하여 그래프 식별자별 정규 후보 부분 그래프를 판단한다. 카운팅 리듀서(133)는 동일한 정규 코드를 가지는 정규 후보 부분 그래프가 발현하는 그래프의 수가 최소 지지도를 만족하는지를 판단하여 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프를 생성한다.

[0041] 여기서 에지 확장부(110)와 생성부(130)는 확장 빈발 부분 그래프로부터 신규 확장 빈발 부분 그래프가 생성되지 않을 때까지 반복하여 확장 빈발 부분 그래프의 에지를 1개씩 확장하여 그래프 데이터베이스의 빈발 부분 그래프를 마이닝한다.

[0042] 본 발명에 따른 마이닝 모듈는 k(k는 자연수)개의 에지를 가지는 모든 빈발 부분 그래프를 2개씩 임의 조합하여 k+1개의 에지를 가지는 모든 후보 부분 그래프를 생성하여 후보 부분 그래프가 최소 지지도를 만족하는지 판단하는 대신, k개의 에지를 가지는 빈발 부분 그래프를 그래프 식별자별로 구분하여 복원 그래프를 생성하고 생성한 복원 그래프에서 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성함으로써, 후보 부분 그래프를 생성하는데 소요되는 연산량과 시간을 줄일 수 있다.

[0043] 도 6은 본 발명에 따른, 대용량 그래프 데이터베이스에서 빈발 부분 그래프를 마이닝하는 방법을 설명하기 위한 흐름도이다.

[0044] 도 6을 참고로 보다 구체적으로 살펴보면, k(k는 자연수)개의 에지를 가지는 빈발 부분 그래프가 입력되는 경우(S110), 그래프 그룹을 구성하는 다수의 그래프에서 그래프 식별자별로 빈발 부분 그래프의 복원 그래프를 생성하고, 생성한 복원 그래프에 기초하여 빈발 부분 그래프로부터 k+1개의 에지로 확장 가능한 후보 부분 그래프를 생성한다(S130).

[0045] 분산되어 있는 다른 그래프 그룹에서 위에서 설명한 방식으로 생성한 후보 부분 그래프에 대한 정보를 수신하며(S150), 생성한 후보 부분 그래프와 수신한 후보 부분 그래프에 기초하여 최소 지지도 이상으로 발현하는 확장 빈발 부분 그래프를 마이닝한다(S170). 여기서 분산되어 있는 다른 그래프 그룹으로부터 수신한 후보 부분 그래프에 대한 정보는 후보 부분 그래프, 후보 부분 그래프가 포함되어 있는 그래프 식별자를 포함하는 것을 특징으로 한다.

[0046] 최소 지지도 이상으로 발현하는 확장 빈발 부분 그래프를 그래프 데이터베이스에서 최소 지지도 이상으로 발현하는 빈발 부분 그래프로 저장하며(S190), 확장 빈발 부분 그래프에 대해 앞서 설명한 S110 내지 S190 단계를 반복한다. 그러나 생성한 후보 부분 그래프와 수신한 후보 부분 그래프에 기초하여 후보 부분 그래프가 최소 지지도 이상으로 발현하지 않는 경우, 그래프 데이터베이스에서 확장 빈발 부분 그래프의 마이닝을 종료한다.

[0047] 도 7은 후보 부분 그래프를 생성하는 단계를 보다 구체적으로 설명하기 위한 흐름도이다.

[0048] 도 7을 참고로 살펴보면, k개의 에지를 가지는 빈발 부분 그래프가 포함된 그래프 식별자를 키 값으로 하여 빈

발 부분 그래프를 구비하는 출력 데이터를 생성한다(S131). 예를 들어, 출력 데이터는 <그래프 식별자, 빈발 부분 그래프> 형식일 수 있다.

[0049] 출력 데이터에 기초하여 동일한 그래프 식별자를 가지는 빈발 부분 그래프로 분류하고, 동일한 그래프 식별자를 가지는 빈발 부분 그래프부터 그래프 식별자별로 복원 그래프를 생성한다(S133). 여기서 복원 그래프는 동일한 그래프 식별자를 가지는 빈발 부분 그래프에서 빈발 부분 그래프를 구성하는 버텍스와 에지의 연결 관계에 기초하여 생성되는데, 예를 들어 동일한 그래프 식별자를 가지는 서로 다른 빈발 부분 그래프가 동일한 버텍스를 가지는 경우, 서로 다른 빈발 부분 그래프에서 동일한 버텍스에 연결되어 있는 버텍스로 에지 확장하여 복원 그래프를 생성한다.

[0050] 복원 그래프에 기초하여 빈발 부분 그래프로부터 k+1개의 에지를 가지도록 에지 확장 가능한 후보 부분 그래프를 생성한다(S135). 여기서 후보 부분 그래프는 복원 그래프와 빈발 부분 그래프의 버텍스 및 에지의 연결 관계를 비교하여, 빈발 부분 그래프의 버텍스로부터 연결된 에지가 복원 그래프에 존재하는 경우 빈발 부분 그래프에서 에지를 확장하여 k+1개의 에지를 가지는 후보 부분 그래프를 생성한다.

[0051] 도 8은 본 발명에 따른 확장 빈발 부분 그래프를 마이닝하는 단계를 보다 구체적으로 설명하기 위한 흐름도이다.

[0052] 도 8을 참고로 보다 구체적으로 살펴보면, 후보 부분 그래프를 정규화 연산하여 후보 부분 그래프의 정규 코드를 계산하고(S151), 후보 부분 그래프 중 정규 코드를 가지는 정규 후보 부분 그래프를 판단한다(S153).

[0053] 그래프 데이터베이스에서 확장 빈발 부분그래프를 효율적으로 마이닝하기 위하여 AGM, FSG, gSpan 등의 다양한 기법들이 연구되고 있다. 이 중에서 성능이 탁월한 것으로 알려진 gSpan이 널리 사용되고 있다. gSpan은 패턴 확장 방법으로 그래프 데이터베이스로부터 빈발 부분그래프를 마이닝하는 알고리즘으로, 깊이우선탐색(depth first search, 이하 DFS) 순서로 방문한 그래프의 에지들을 각 에지의 DFS 코드로 표현한다. 그래프는 에지의 확장 순서로 탐색 시작 버텍스(vertex)에 따라 다양한 DFS 코드가 생성되는데, 빈발 부분그래프를 DFS코드로 표현하고, DFS 코드로 표현된 후보 부분 그래프 중 최소 DFS코드를 가지는 정규 후보 부분 그래프인 경우에만 확장 빈발 부분 그래프로 확장하여 빠르게 빈발 부분그래프를 마이닝하며 최소 DFS 코드를 가지지 않는 나머지 후보 부분 그래프는 중복 그래프로 간주하여 제거한다.

[0054] 후보 부분 그래프의 DFS 코드 중 최소 DFS 코드를 후보 부분 그래프의 정규 코드 형태로 사용하여 후보 부분 그래프를 유일하게 표현한다. 후보 부분 그래프를 정규 코드 형태로 표현하는 것을 정규화 연산(canonized operation)이라 정의하고 정규 코드 형태로 표현되는 후보 부분 그래프를 정규 후보 부분 그래프라 언급한다.

[0055] 정규화 연산을 위하여 후보 부분 그래프의 최소 DFS코드를 찾기 위한 과정은 다음과 같다. 후보 부분 그래프의 에지들을 DFS코드로 표현하고, 각 에지의 DFS코드 중 가장 작게 표현된 값을 탐색 시작 지점으로 설정한다. 탐색 시작 지점부터 깊이우선탐색(DFS) 순서에 따라 버텍스의 식별자를 순차적으로 부여한다. 만약 동일 버텍스가 여러 개의 에지를 가지면, 가장 작은 DFS코드로 표현되는 에지를 선택하여 탐색한다. 만약 가장 작게 표현되는 에지가 여러 개 존재할 경우, 하나의 에지를 선택하여 모두 탐색한 후, 나머지 에지들에 대해서도 재귀적으로 모든 에지를 탐색한다. 탐색하여 얻은 DFS코드 집합 중에서 가장 작게 표현되는 최소 DFS코드를 찾는다.

[0056] 동일한 정규 코드의 정규 후보 부분 그래프를 가지는 그래프의 수가 최소 지지도를 만족하는지를 판단하여(S155), 그래프 데이터베이스에서 정규 후보 부분 그래프를 가지는 그래프의 수가 최소 지지도를 초과하는 경우 정규 후보 부분 그래프를 빈발 부분 그래프로부터 에지 확장한 확장 빈발 부분 그래프로 마이닝한다(S157).

[0057] 도 9 내지 도 11은 도 2에 도시되어 있는 그래프 그룹에서 본 발명에 따른 빈발 부분 그래프의 마이닝 방법의 일 예를 설명하기 위한 도면이다.

[0058] 먼저, 도 9와 도 2를 참고로 살펴보면, 그래프 그룹에서 최소 지지도 2 이상으로 발현하는 1에지 빈발 부분 그래프를 그래프 식별자별로 생성한다. 도 9(a)와 같이 그래프1(G1)에서는 최소 지지도 2를 만족하는 3개의 빈발 부분 그래프가 생성되며, 도 9(b)와 같이 그래프2(G2)에서는 최소 지지도 2를 만족하는 2개의 빈발 부분 그래프가 생성되며, 도 9(c)와 같이 그래프3(G3)에서는 최소 지지도 2를 만족하는 2개의 빈발 부분 그래프가

생성된다.

[0059] 도 10을 참고로 살펴보면, 그래프 식별자별로 생성된 빈발 부분 그래프로부터 복원 그래프를 생성한다. 그래프 식별자별로 동일한 버텍스에 연결되어 있는 에지를 확장하여 복원 그래프를 생성하는데, 예를 들어 그래프1(G1)에서 동일한 버텍스(B), 동일한 버텍스(A)에 연결되어 있는 에지를 확장하여 도 10(a)와 같은 복원 그래프를 생성한다. 이와 동일하게, 그래프2(G2)에 대해서도 도 10(b)와 같은 복원 그래프를 생성하며, 그래프3(G3)에 대해서도 도 10(c)와 같은 복원 그래프를 생성한다.

[0060] 도 11을 참고로 살펴보면, 그래프별의 복원 그래프와 그래프별의 1에지 빈발 부분 그래프의 버텍스 및 에지의 연결 관계를 비교하여, 1에지 빈발 부분 그래프의 버텍스로부터 연결된 에지가 복원 그래프에 존재하는 경우 빈발 부분 그래프에서 에지를 확장하여 2개의 에지를 가지는 후보 부분 그래프를 생성한다. 예를 들어, 그래프1(G1)의 1에지 빈발 부분 그래프와 그래프1(G1)의 복원 그래프 사이의 버텍스 및 에지의 연결 관계를 비교하여 도 11(a)과 같은 후보 부분 그래프를 생성한다.

[0061] 여기서 버텍스(A)에서 버텍스(B)로 연결된 1에지 빈발 부분 그래프와 버텍스(A)에서 버텍스(D)로 연결된 1에지 빈발 부분 그래프로부터 생성되는 후보 부분 그래프는 2가지가 존재하는데, 즉 버텍스(A)에서 순차적으로 버텍스(B), 버텍스(D)로 확장되어 생성되는 후보 부분 그래프와 버텍스(A)에서 순차적으로 버텍스(D), 버텍스(B)로 확장되어 생성되는 후보 부분 그래프가 존재한다. 이 경우, 버텍스(A)에서 순차적으로 버텍스(B)와 버텍스(D)로 확장되어 생성되는 후부 부분 그래프가 정규 코드를 가지는 정규 후보 부분 그래프이며 나머지 후보 부분 그래프는 삭제된다.

[0062] 이와 동일하게, 그래프2(G2)의 1에지 빈발 부분 그래프와 그래프2(G2)의 복원 그래프 사이의 버텍스 및 에지의 연결 관계를 비교하여 도 11(b)과 같은 후보 부분 그래프를 생성하며, 그래프3(G3)의 1에지 빈발 부분 그래프와 그래프3(G3)의 복원 그래프 사이의 버텍스 및 에지의 연결 관계를 비교하여 도 11(c)과 같은 후보 부분 그래프를 생성한다.

[0063] 후보 부분 그래프 중 정규 후보 부분 그래프만을 기준으로 정규 후보 부분 그래프를 포함하는 그래프의 수가 최소 지지도를 만족하는지 여부로 확장 빈발 부분 그래프를 생성한다.

[0064] 한편, 상술한 본 발명의 실시 예들은 컴퓨터에서 실행될 수 있는 프로그램으로 작성 가능하고, 컴퓨터로 읽을 수 있는 기록 매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다.

[0065] 상기 컴퓨터로 읽을 수 있는 기록 매체는 전기 또는 자기식 저장 매체(예를 들어, 롬, 플로피 디스크, 하드디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등) 및 캐리어 웨이브(예를 들면, 인터넷을 통한 전송)와 같은 저장 매체를 포함한다.

[0066] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0067] 100: 마이닝 장치

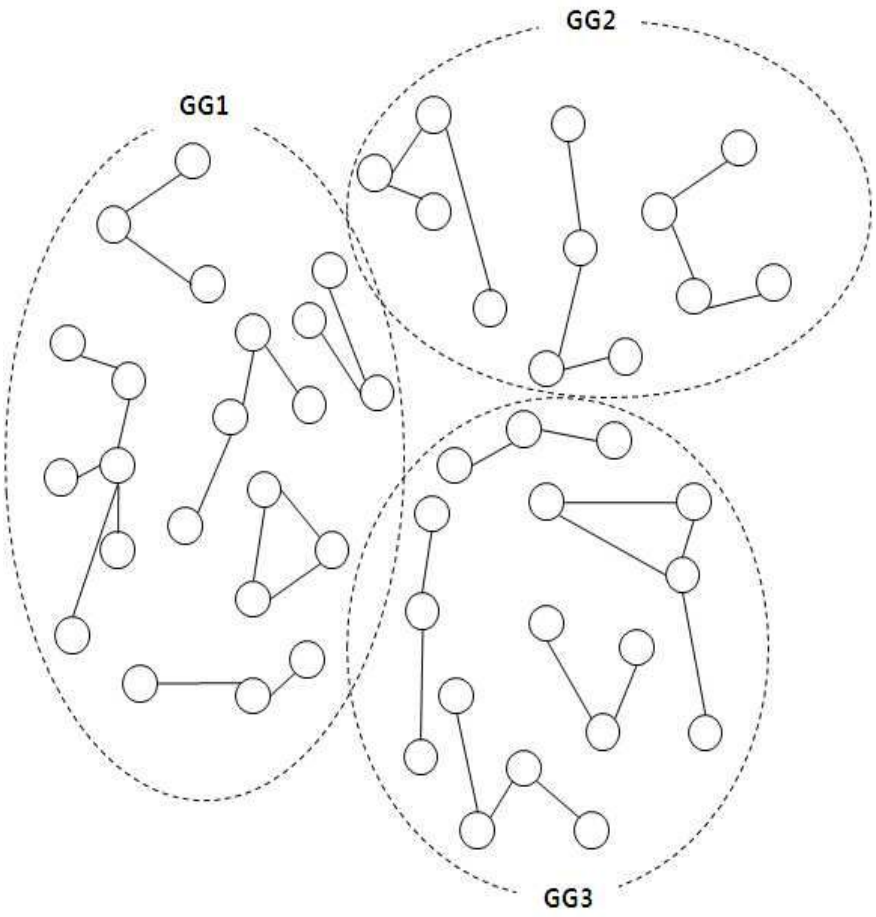
110: 에지 확장부 130: 생성부

111: 에지 확장 맵퍼 113: 에지 확장 리듀서

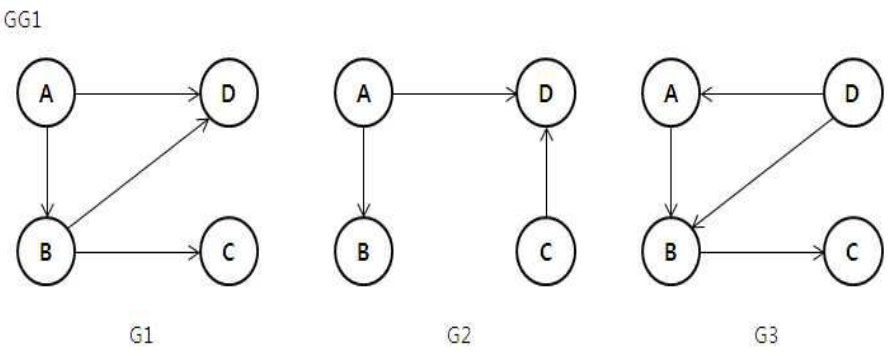
131: 카운팅 맵퍼 133: 카운팅 리듀서

도면

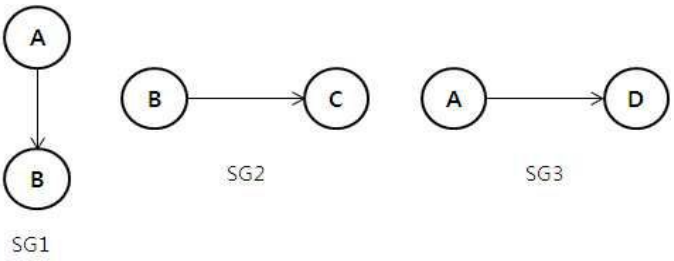
도면1



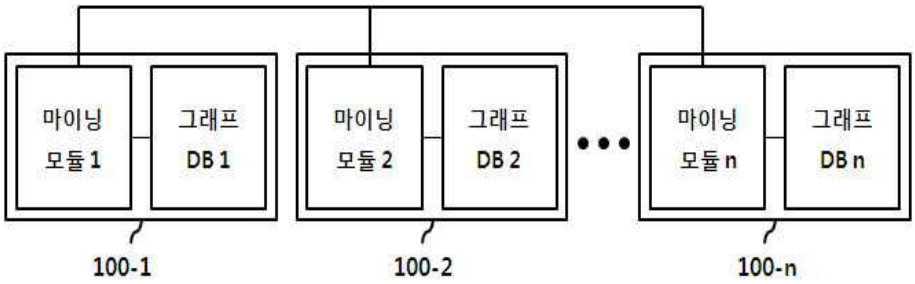
도면2



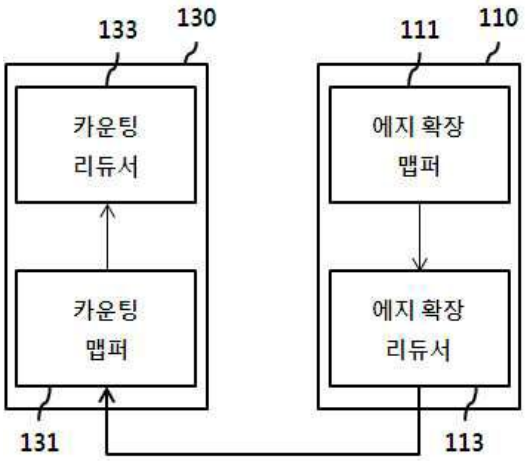
도면3


A
B
SG1
B
C
SG2
A
D
SG3

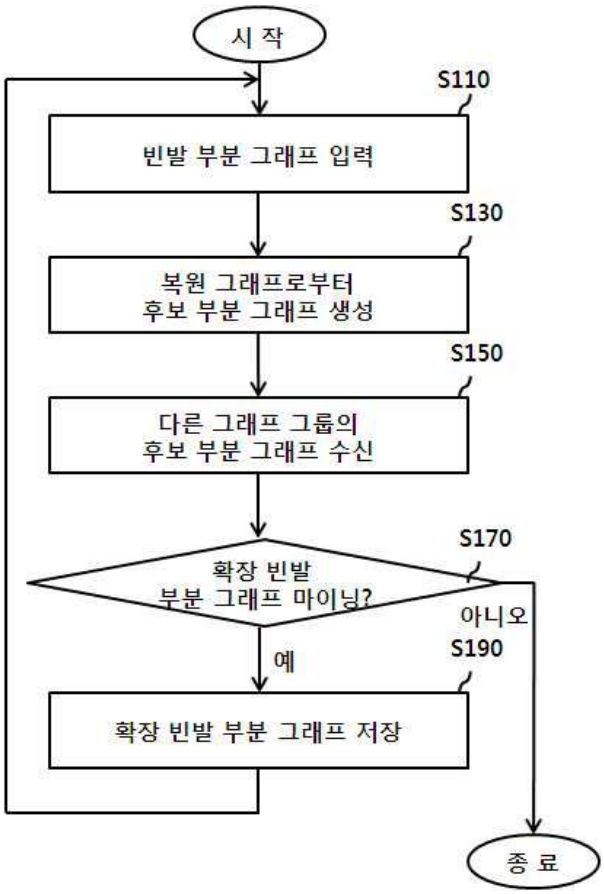
도면4


마이닝 모듈 1
그래프 DB 1
100-1
마이닝 모듈 2
그래프 DB 2
100-2
마이닝 모듈 n
그래프 DB n
100-n

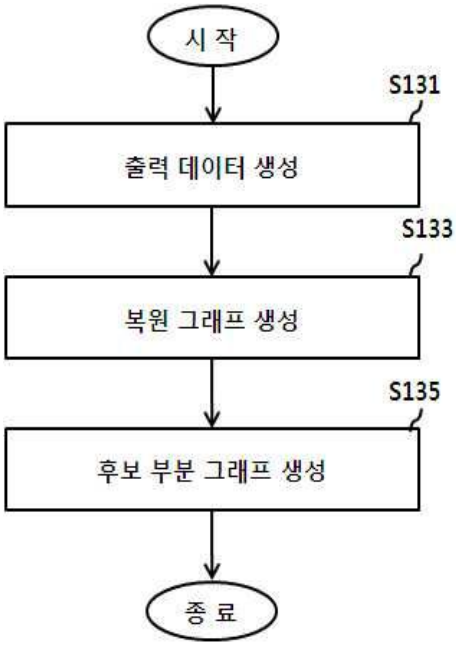
도면5


133
130
카운팅 리듀서
카운팅 맵퍼
131
111
110
에지 확장 맵퍼
에지 확장 리듀서
113

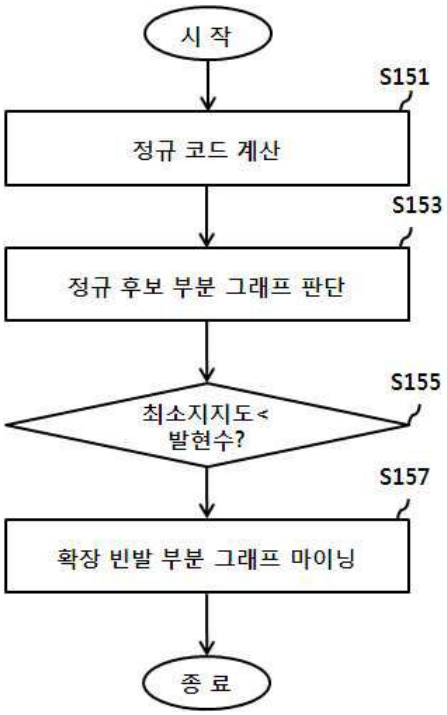
도면6


시 작
S110
빈발 부분 그래프 입력
S130
복원 그래프로부터
후보 부분 그래프 생성
S150
다른 그래프 그룹의
후보 부분 그래프 수신
S170
확장 빈발
부분 그래프 마이닝?
아니오
예
S190
확장 빈발 부분 그래프 저장
종 료

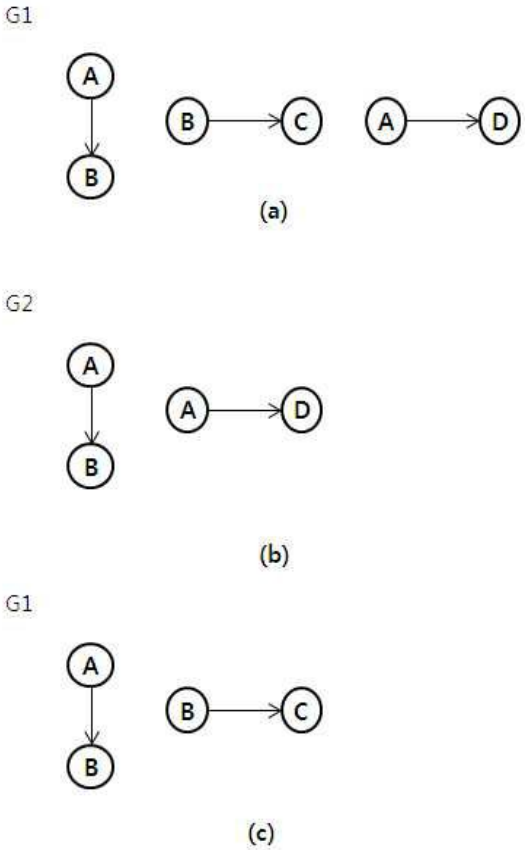
도면7


시 작
S131
출력 데이터 생성
S133
복원 그래프 생성
S135
후보 부분 그래프 생성
종 료

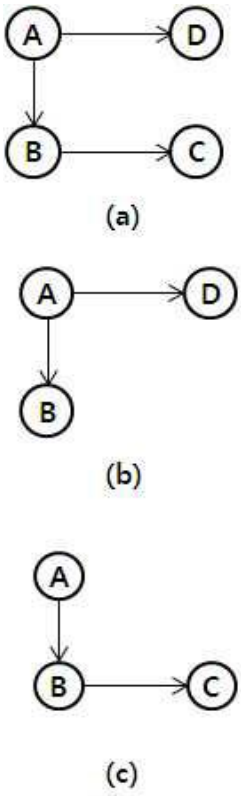
도면8


시 작
S151
정규 코드 계산
S153
정규 후보 부분 그래프 판단
S155
최소지지도 < 발현수?
S157
확장 빈발 부분 그래프 마이닝
종 료

도면9


G1
A
B
B
C
A
D
(a)
G2
A
B
A
D
(b)
G1
A
B
B
C
(c)

도면10



도면11

