



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0064081
(43) 공개일자 2014년05월28일

(51) 국제특허분류(Int. Cl.)

G06F 17/00 (2006.01)

(21) 출원번호 10-2012-0130999

(22) 출원일자 2012년11월19일

심사청구일자 2012년11월19일

(71) 출원인

충북대학교 산학협력단

충청북도 청주시 흥덕구 내수동로 52 (개신동)

세종대학교 산학협력단

서울특별시 광진구 능동로 209 (군자동, 세종대학교)

(72) 발명자

윤은일

충청북도 청주시 흥덕구 내수동로 52 충북대학교 S1-4 219호

양홍모

충청북도 청주시 흥덕구 내수동로 52 충북대학교 S1-4 219호

(뒷면에 계속)

(74) 대리인

원성수, 박종경, 강세창

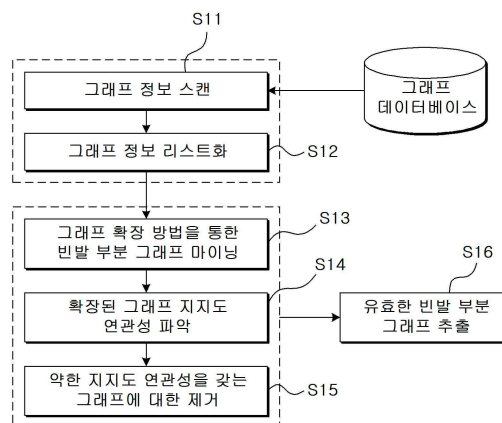
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법

(57) 요약

본 발명은 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법에 관한 것으로서, 어느 한 정점을 시작으로 부분 그래프가 점진적인 확장 과정을 거치면서, 확장의 각 단계마다 해당하는 부분 그래프가 설정된 최소 지지도 임계값 이상의 지지도 수치일 경우에, 추출되는 부분 그래프의 지지도 연관성에 대한 정도를 파악하는 단계와, 상기 지지도 연관성의 정도에 따라 의미 없는 부분 그래프를 선별하는 단계와, 의미 없는 패턴으로 선별된 부분 그래프와, 상기 부분 그래프로부터 파생되는 상위 그래프 정보들을 마이닝 과정에서 제거하는 단계를 포함함으로써, 사용자에게 보다 더 의미 있는 그래프 마이닝 결과를 제공할 수 있고, 보다 더 즉각적인 수행 시간을 보장할 수 있는 물론 마이닝에 필요한 메모리 사용량을 효과적으로 줄일 수 있다.

대표도 - 도1



(72) 발명자

편광범

충청북도 청주시 흥덕구 내수동로 52 충북대학교
S1-4 219호

이강인

충청북도 청주시 흥덕구 내수동로 52 충북대학교
S1-4 219호

김지원

충청북도 청주시 흥덕구 내수동로 52 충북대학교
S1-4 219호

이 발명을 지원한 국가연구개발사업

과제고유번호 2013-005682

부처명 미래창조과학부

연구사업명 일반연구자지원사업(신진연구)

연구과제명 효율적인 스트림 패턴 처리를 위한 마이닝 핵심기술 개발

기 여 율 1/2

주관기관 충북대학교

연구기간 2013.05.01 ~ 2014.04.30이 발명을 지원한 국가연구개발사업

과제고유번호 2008-0062611

부처명 미래창조과학부

연구사업명 선도연구센터지원사업 이공학분야

연구과제명 개인특화 의료를 위한 자기조직적 지식진화 기술 개발

기 여 율 1/2

주관기관 충북대학교

연구기간 2013.03.01 ~ 2014.02.28

특허청구의 범위

청구항 1

어느 한 정점을 시작으로 부분 그래프가 점진적인 확장 과정을 거치면서, 확장의 각 단계마다 해당하는 부분 그래프가 설정된 최소 지지도 임계값 이상의 지지도 수치일 경우에,

- (a) 추출되는 부분 그래프의 지지도 연관성에 대한 정도를 파악하는 단계;
- (b) 상기 지지도 연관성의 정도에 따라 의미 없는 부분 그래프를 선별하는 단계; 및
- (c) 의미 없는 패턴으로 선별된 부분 그래프와, 상기 부분 그래프로부터 파생되는 상위 그래프 정보들을 마이닝 과정에서 제거하는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 2

제1항에 있어서,

상기 단계 (a)는,

- (a-1) 그래프를 구성하는 간선들 각각의 지지도 정보를 파악하는 단계;
- (a-2) 파악된 상기 지지도 정보를 바탕으로 최소 지지도 값을 갖는 간선과 최대 지지도 값을 갖는 간선을 파악하는 단계; 및
- (a-3) 상기 최소 지지도 값의 간선과 최대 지지도 값의 간선의 비율을 구하여, 그래프의 지지도 연관성을 수치화하는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 3

제2항에 있어서,

상기 단계 (a-3)은,

패턴 내의 간선 지지도 수치 중 상기 최소 지지도 값을 분자로, 최대 지지도 값을 분모로 하여 0과 1 사이에 실수 형태로 존재하는 비율 값을 계산하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 4

제1항에 있어서,

상기 단계 (b)는,

- (b-1) 상기 지지도 연관성에 대한 유효 범위를 한정짓기 위한 최소 지지도 연관성 임계값을 설정하는 단계; 및
- (b-2) 상기 단계 (a)를 통해 도출된 값과, 상기 단계 (b-1)을 통해 얻어진 임계값을 비교하는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 5

제4항에 있어서,

상기 단계 (b-2)는,

그래프 패턴의 지지도 값이 설정된 최소 지지도 연관성 임계값보다 크거나 같을 경우 해당하는 부분 그래프를 유효한 결과물로 인지하고, 지지도 값이 임계값보다 작을 경우 의미 없는 패턴으로 인지하는 지지도 연관성을

갖는 빈발 부분 그래프의 마이닝 방법.

청구항 6

제1항에 있어서,

상기 단계 (c)는,

(c-1) 선별된 부분 그래프 자체를 마이닝 과정에서 제거하는 단계; 및

(c-2) 제거된 상기 부분 그래프를 통해서 확장될 수 있는 모든 상위 부분 그래프 정보를 제거하는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 7

제6항에 있어서,

상기 단계 (c-2)는,

(c-2-1) 안티모노톤 법칙을 보존하면서 상위 부분 그래프들을 제거하는 단계; 및

(c-2-2) 상기 상위 부분 그래프 정보의 제거에 따라 마이닝의 탐색 공간을 줄이는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 8

제7항에 있어서,

상기 단계 (c-2-1)은,

(c-2-1-1) 상기 추출되는 부분 그래프의 지지도 연관성에 대한 정도를 파악하는 단계를 통해 얻어지는 지지도 연관성 값이 상기 안티모노톤 법칙을 만족함을 증명하는 단계; 및

(c-2-1-2) 상기 증명에 대응하여 의미 없는 부분 그래프의 모든 상위 부분 그래프들을 제거하는 단계를 포함하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 9

제8항에 있어서,

상기 단계 (c-2-1-1)은,

상기 단계 (a-3)에 의거하여, 현재 부분 그래프의 지지도 연관성 값에 대해, 해당 부분 그래프가 확장된 어떠한 상위 그래프의 지지도 연관성 값도 현재 부분 그래프의 지지도 연관성 값을 넘을 수 없게 되고, 따라서 상기 현재 부분 그래프의 지지도 연관성 값이 설정된 최소 지지도 연관성 임계값보다 낮다면, 그에 따른 상위 부분 그래프의 지지도 연관성 값들도 낮다는 특성을 판단하여 안티모노톤 법칙을 만족함을 확인하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법.

청구항 10

제7항에 있어서,

상기 단계 (c-2-2)는,

상기 단계 (c)를 통해 의미 없는 부분 그래프의 추출을 위해 할당되는 탐색 공간과 그에 따른 컴퓨터 자원의 할당을 미연에 방지하고, 그래프 마이닝 작업에 요구되는 시간적, 공간적 자원량을 감소시키는 지지도 연관성을

갖는 빈발 부분 그래프의 마이닝 방법.

명세서

기술분야

- [0001] 본 발명은 그래프를 마이닝하는 방법에 관한 것으로, 더욱 상세하게는 그래프로 구성된 데이터베이스로부터 지지도 연관성을 갖는 빈발 부분 그래프를 효율적으로 마이닝하는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법에 관한 것이다.

배경기술

- [0002] 데이터 마이닝은 거대한 규모의 데이터베이스로부터 숨겨진 유용한 정보를 찾기 위한 일련의 작업을 의미하며, 데이터 마이닝의 여러 분야들 가운데, 빈발 패턴 마이닝 분야는 가장 광범위하고 활발하게 연구가 진행되어 온 분야들 중 하나이다. 이 빈발 패턴 마이닝은 막대한 크기의 데이터베이스로부터 빈발하게 발생하는 특정 패턴들을 추출하게 되고, 따라서 이 패턴들은 의미 있는 정보를 포함하게 된다. 이 분야는 Apriori와 같은 BFS(Broad First Search) 방법을 기반으로 하여 많은 기법들이 제안되었으며, 그 이후에는 대표적으로 FP-growth와 같은 DFS(Depth First Search) 기반의 다양한 방법들이 제안되고 있으며, 다채로운 변형 기법이나 알고리즘들이 개발되고 있는 추세이다.
- [0003] 그러나 근래에 들어서는, 실세계에서 얻어지는 데이터들이 점점 더 복잡해짐에 따라, 과거의 단순한 구조를 가지는 데이터베이스에 대한 빈발 패턴 마이닝 기법들은 곧 한계에 직면하게 되었다. 이러한 이유로, 한층 더 고차원적으로 패턴 마이닝 작업을 수행 할 수 있는 빈발 그래프 마이닝이 등장하게 되었다. 기존의 데이터베이스를 구성하는 각 트랜잭션들은 단순한 형태인 항목집단으로 이루어져 있기 때문에 다차원적인 정보는 표현하지 못하는 단점이 존재한다. 반면에 그래프 데이터베이스를 구성하는 트랜잭션들은 각각이 하나의 그래프 형태로 구성되어 있으며, 이 그래프 형태는 실세계에 존재하는 대부분의 객체들에 대한 정보를 담을 수 있는 복잡한 자료구조이다. 그러나 빈발 그래프 마이닝은 복잡한 자료구조인 그래프를 상대로 마이닝을 수행하기 때문에 일반적인 빈발 패턴 마이닝에 비해 막대한 연산의 오버헤드를 발생시키는 물론 추출되는 그래프들의 중복성을 파악하기 위해서는 그래프의 동형관단을 수행해야 하며 이것은 NP-hard 문제로서 역시 큰 연산 오버헤드를 초래한다. 따라서 이에 대한 문제들을 해결하고 더욱 효율적으로 마이닝 연산을 수행하기 위한 방법이 필요성이 증대되어 왔다.
- [0004] 이는, 기존의 그래프 마이닝 접근 방법들은 패턴 내 요소들의 연관 정도를 고려하지 않고, 오로지 그 패턴이 가지는 지지도 자체만을 고려해 빈발한 부분 그래프를 추출하기 때문에 생성되는 부분 그래프들 중 많은 부분이 실제로는 유효하지 않은 정보임에도 불구하고 유효한 결과물로 도출되는 문제를 가짐은 물론 오로지 최소 지지도 임계값만을 기준으로 부분 그래프들을 추출하기 때문에 막대한 수의 결과물들을 초래하는 단점이 존재하게 된 것이다.
- [0005] 이와 같이, 빈발 패턴 마이닝은 설정된 최소 지지도 임계값이 감소함에 따라, 생성되는 빈발 패턴의 수가 증가하는 특성을 가지며, 그 증가치는 임계값이 낮을수록 기하급수적으로 커진다. 하물며 빈발 그래프 마이닝에서는 그래프가 가지는 특성 상 훨씬 더 많은 결과물들이 생성되며, 그에 따른 연산의 오버헤드 또한 상당하다. 그러나 이렇게 얻어진 빈발 부분 그래프들 모두가 사용자에게 실제로 유용한 정보가 되는 것은 아니며, 심지어는 이들의 대부분이 실제로는 의미 없는 정보로 전락할 수 있으며, 이 경우는 결국 쓸모없는 부분 그래프를 마이닝하기 위해 막대한 시간 및 메모리 자원을 소모한 격이 된다. 따라서, 이러한 비효율적인 면을 극복하고 마이닝의 성능을 높이기 위해서는 최적의 자원을 가지고 실제로 유용한 그래프 정보들만을 마이닝하기 위한 방법이 요구된다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 공개특허공보 제10-2010-0099378호(공개일 2010.09.13.)

발명의 내용

해결하려는 과제

[0007] 따라서, 본 발명은 상기한 종래 기술의 문제점을 해결하기 위해 이루어진 것으로서, 본 발명의 목적은 그래프 데이터베이스로부터 유용한 부분 그래프들을 선별(약한 연관성을 갖는 부분 그래프들을 사전에 제거)할 수 있는 물론 그래프 마이닝 과정에서 발생하는 연산의 오버헤드를 효과적으로 감소시킬 수 있는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법을 제공하는데 있다.

[0008] 즉, 본 발명에서는 복잡한 구조의 그래프 데이터베이스로부터 패턴 내 요소들 사이에 강력한 연관성을 가지는 빈발 서브 그래프들만을 선별적으로 마이닝하기 위한 방법을 제안한다. 항목 집단 형태의 데이터베이스와 달리 그래프 데이터베이스에서는 상대적으로 많은 빈발 패턴이 등장 할 수 있으며, 그래프의 특성 상 모양은 다르지만 결국 동일한 패턴이 되는 부분 그래프들에 대해서는 일일이 그래프 동형판단 작업을 수행함으로써 동일한 부분 그래프의 중복 추출을 방지해야 한다. 하지만 그래프의 동형판단은 NP-hard문제로서 이 자체로는 효율적인 방법을 도출 할 수 없기 때문에 전체적인 마이닝 성능을 향상시키기 위해서는 생성되는 부분 그래프의 수 자체를 줄이는 방법이 필요하다. 그리고 빈발 그래프 마이닝을 통해 발생하는 부분 그래프들 모두가 실질적으로 유용한 결과물들이 되는 것은 아니기 때문에 이들 가운데 핵심적인 그래프들을 선택적으로 마이닝하기 위한 기준과 관련 기법이 필요하다. 따라서 본 발명은 제안되는 기준과 기법들을 통해 그래프를 구성하는 요소들 사이에 얼마나 강한 연관성이 존재하는지를 파악하여 빈발 부분 그래프들 중에서도, 강력한 연관 특성을 가지는 핵심 그래프 정보만을 추출하며, 이를 통해 감소된 패턴의 수에 따라 마이닝의 효율성을 향상시키는 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법을 제공하는데 있다.

과제의 해결 수단

[0009] 상기와 같은 목적을 달성하기 위한 본 발명의 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법은, 어느 한 정점을 시작으로 부분 그래프가 점진적인 확장 과정을 거치면서, 확장의 각 단계마다 해당하는 부분 그래프가 설정된 최소 지지도 임계값 이상의 지지도 수치일 경우에, (a) 추출되는 부분 그래프의 지지도 연관성에 대한 정도를 파악하는 단계; (b) 상기 지지도 연관성의 정도에 따라 의미 없는 부분 그래프를 선별하는 단계; 및 (c) 의미 없는 패턴으로 선별된 부분 그래프와, 상기 부분 그래프로부터 파생되는 상위 그래프 정보들을 마이닝 과정에서 제거하는 단계를 포함하는 것을 특징으로 한다.

[0010] 상기 단계 (a)는, (a-1) 그래프를 구성하는 간선들 각각의 지지도 정보를 파악하는 단계; (a-2) 파악된 상기 지지도 정보를 바탕으로 최소 지지도 값을 갖는 간선과 최대 지지도 값을 갖는 간선을 파악하는 단계; 및 (a-3) 상기 최소 지지도 값의 간선과 최대 지지도 값의 간선의 비율을 구하여, 그래프의 지지도 연관성을 수치화하는 단계를 포함한다. 이때, 상기 단계 (a-3)은, 패턴 내의 간선 지지도 수치 중 상기 최소 지지도 값을 분자로, 최대 지지도 값을 분모로 하여 0과 1 사이에 실수 형태로 존재하는 비율 값을 계산한다.

[0011] 상기 단계 (b)는, (b-1) 상기 지지도 연관성에 대한 유효 범위를 한정짓기 위한 최소 지지도 연관성 임계값을 설정하는 단계; 및 (b-2) 상기 단계 (a)를 통해 도출된 값과, 상기 단계 (b-1)을 통해 얻어진 임계값을 비교하는 단계를 포함한다. 이때, 상기 단계 (b-2)는, 그래프 패턴의 지지도 값이 설정된 최소 지지도 연관성 임계값보다 크거나 같을 경우 해당하는 부분 그래프를 유효한 결과물로 인지하고, 지지도 값이 임계값보다 작을 경우 의미 없는 패턴으로 인지한다.

[0012] 상기 단계 (c)는, (c-1) 선별된 부분 그래프 자체를 마이닝 과정에서 제거하는 단계; 및 (c-2) 제거된 상기 부분 그래프를 통해서 확장될 수 있는 모든 상위 부분 그래프 정보를 제거하는 단계를 포함한다. 이때, 상기 단계 (c-2)는, (c-2-1) 안티모노톤 법칙을 보존하면서 상위 부분 그래프들을 제거하는 단계; 및 (c-2-2) 상기 상위 부분 그래프 정보의 제거에 따라 마이닝의 탐색 공간을 줄이는 단계를 포함한다. 또한, 상기 단계 (c-2-1)은, (c-2-1-1) 상기 추출되는 부분 그래프의 지지도 연관성에 대한 정도를 파악하는 단계를 통해 얻어지는 지지도 연관성 값이 상기 안티모노톤 법칙을 만족함을 증명하는 단계; 및 (c-2-1-2) 상기 증명에 대응하여 의미 없는 부분 그래프의 모든 상위 부분 그래프들을 제거하는 단계를 포함한다. 그리고, 상기 단계 (c-2-1-1)은, 상기 단

계 (a-3)에 의거하여, 현재 부분 그래프의 지지도 연관성 값에 대해, 해당 부분 그래프가 확장된 어떠한 상위 그래프의 지지도 연관성 값도 현재 부분 그래프의 지지도 연관성 값을 넘을 수 없게 되고, 따라서 상기 현재 부분 그래프의 지지도 연관성 값이 설정된 최소 지지도 연관성 임계값보다 낮다면, 그에 따른 상위 부분 그래프의 지지도 연관성 값들도 낮다는 특성을 판단하여 안티모노톤 법칙을 만족함을 확인한다. 한편, 상기 단계 (c-2-2)는, 상기 단계 (c)를 통해 의미 없는 부분 그래프의 추출을 위해 할당되는 탐색 공간과 그에 따른 컴퓨터 자원의 할당을 미연에 방지하고, 그래프 마이닝 작업에 요구되는 시간적, 공간적 자원량을 감소시킨다.

발명의 효과

- [0013] 상술한 바와 같이, 본 발명에 의한 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법에 따르면, 생성되는 그래프 패턴들이 실제로 얼마나 유용한가를 평가할 수 있는 기준인 지지도 연관성을 새로운 기준으로 삼아 빈발하면서 동시에 강력한 연관성을 가지는 부분 그래프들만을 마이닝할 수 있도록 하는 기법을 제공함으로써, 사용자에게 보다 더 의미 있는 그래프 마이닝 결과를 제공할 수 있다.
- [0014] 또한, 본 발명에 따르면, 불필요한 그래프 패턴에 대한 정보와 해당하는 연산 과정들을 사전에 제거함으로써, 보다 더 즉각적인 수행 시간을 보장함은 물론 마이닝에 필요한 메모리 사용량을 효과적으로 줄일 수 있다.

도면의 간단한 설명

- [0015] 도 1은 본 발명의 일실시예에 의한 지지도 연관성을 갖는 빈발 부분 그래프를 마이닝하기 위한 방법의 흐름도이다.
- 도 2는 본 발명의 일실시예에 의한 그래프 데이터베이스로부터 각각의 그래프 트랜잭션을 읽어 들이고 리스트화하기 위한 전처리 방법의 흐름도이다.
- 도 3은 본 발명의 일실시예에 의한 전처리 후 구축된 정보를 바탕으로 지지도 연관성을 갖는 빈발 부분 그래프를 마이닝하기 위해 그래프 패턴을 확장하기 위한 방법의 흐름도이다.
- 도 4는 본 발명의 일실시예에 의한 그래프 동형판단의 예시도이다.
- 도 5는 본 발명의 일실시예에 의한 부분 그래프 내부의 간선들이 가지는 지지도 연관성을 파악하기 위한 방법의 흐름도이다.
- 도 6은 본 발명의 일실시예에 의한 그래프의 지지도 관련성을 고려한 경우와 그렇지 않은 경우에 대한 빈발 부분 그래프 마이닝 결과의 차이점 예시도이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 이하, 본 발명의 지지도 연관성을 갖는 빈발 부분 그래프의 마이닝 방법에 대하여 첨부된 도면을 참조하여 상세히 설명하기로 한다.
- [0017] 본 발명의 빈발 그래프 마이닝 방법은 도 1의 과정에 따라 수행되며, 각각은 막대한 수의 그래프 트랜잭션들로 구성된 데이터베이스로부터 그래프를 스캔하기 위한 과정(S11), 스캔한 정보를 마이닝에 적합한 리스트 형태로 메인 메모리에 적재하기 위한 전처리 과정(S12), 빈발 부분 그래프를 추출하기 위해 그래프를 확장하기 위한 과정(S13), 확장된 그래프 내의 각요소들 사이에 존재하는 지지도 연관성의 정도를 파악하기 위한 과정(S14), 파악된 연관 정도를 고려해 유효하지 않은, 즉 약한 지지도 연관성을 갖는 그래프를 제거하기 위한 과정(S15), 마지막으로 밀접한 지지도 연관성을 가지는 빈발 부분 패턴들을 추출하는 과정(S16)으로 구성되어 있다. 그래프 데이터베이스의 스캔 과정에서는, 읽어 들일 그래프 트랜잭션이 있을 경우에(S21), 그래프 데이터베이스 내의 각 트랜잭션들에 대해 정점과 간선 정보를 읽어 들이며, 각각의 정점과 간선에 대한 지지도를 계산하고 트랜잭션에 일련번호를 할당한다(S22). 전처리 과정은 데이터베이스의 스캔 과정과 병행하여 진행되며, 하나의 트랜잭션을 읽어 들여 관련 정보를 파악한 후(S23), 간선을 기준으로 정보를 리스트화 하여 저장한다(S24). 이 과정에 대한 단계는 도 2에 나타내고 있다.
- [0018] 그래프 데이터베이스 내 모든 트랜잭션들에 대해 이러한 작업을 반복하게 되면, 모든 그래프 트랜잭션들의 정보

를 담고 있는 리스트가 구축이 되며, 이 리스트는 임베딩 리스트로 명명된다. 임베딩 리스트는 추후 그래프 확장 과정에서 그래프의 실제 위치정보를 참조하는데 쓰이고, 이 리스트를 사용함으로써 반복적인 데이터베이스의 스캔을 회피할 수 있게 된다. 임베딩 리스트는 마이닝 과정동안 메인메모리에 적재된다. 모든 트랜잭션에 대한 처리가 끝난 후에는 정점과 간선들의 지지도 정보를 확인하면서 설정된 최소 지지도 임계값보다 작은 지지도 값을 갖는 요소들이 있다면, 이들에 대한 정보는 안티모노톤 법칙(Anti-monotone property)에 의해 사전에 제거되며 추후 마이닝 과정에서 고려되지 않는다. 전처리 과정이 끝난 후에는 이 과정을 통해 얻어진 정점과 간선의 정보와 임베딩 리스트를 가지고 도 3과 같이 그래프 확장 과정을 수행함으로써 본격적인 마이닝 작업을 진행한다.

[0019] 처리되지 않은 정점이 있을 경우에(S31), 처음에는 하나의 정점만이 부분 그래프로 선택되며(S32), 확장 가능한 간선이 있을 경우에(S33), 단계별로 간선을 하나씩 확장해 나가면서 부분 빈발 그래프들을 찾게 된다(S34). 그래프가 가질 수 있는 형태는 크게 경로(Path), 자유트리(Free-tree), 순환 그래프(Cyclic graph) 총 3가지로 분류될 수 있다. 그래프 확장 초기에는 자연히 경로의 형태만을 가지게 되고, 이후에는 자유롭게 확장이 되지만, 자유트리에서 경로의 형태로, 순환 그래프에서 경로나 자유트리의 형태로 회귀할 수는 없다. 또한 본 발명에서는 경로는 자유롭게 확장되고, 자유트리는 그 자신이나 순환그래프의 형태로, 순환그래프는 오로지 순환 간선의 추가를 통해서만 확장되도록 그 범위를 제한함으로써 부분 그래프 패턴의 중복 생성과 그에 따른 동형 판단의 NP-hard 연산 문제를 상당 수 줄인다.

[0020] 그래프의 동형 판단 문제는 도 4와 같이 각기 다른 형태를 가지는 것처럼 보이는 부분 그래프들이 비교 결과 동일한 형태로 밝혀지는 것을 말하며, 그래프의 규모가 커질수록 해당하는 연산의 양 역시 막대하게 증가하기 때문에 본 발명의 효과중 하나인 의미 없는 패턴의 사전 제거를 통한 탐색 공간의 축소가 성능 향상의 중요한 요소로 작용하는 것이다. 처음에 선택된 정점으로부터 간선 하나와 그에 따른 정점 하나가 추가되어 확장이 진행되면 경로의 형태를 가지는 부분 그래프가 된다. 추가되는 간선과 정점의 정보는 임베딩 리스트를 참고하고, 이 리스트를 통해 실제로 빈발한 요소들만이 확장 과정에 쓰이기 때문에 확장이 완료된 그래프는 최소 지지도 임계값을 만족하는 값을 가진다. 확장이 진행된 후에는 그래프 내 각 요소들 사이에 존재하는 연관성 정도를 파악하는 과정을 진행하며(S35), 그 세부 단계는 도 5에 나타내고 있다.

[0021] 본 발명에서 연관성을 파악할 때는 그래프의 정점 정보가 아닌 간선 정보를 이용한다. 그 이유는 그래프의 특성 상 서로 다른 그래프 패턴이 같은 정점 정보를 가질 수는 있어도 같은 간선 정보를 가질 수는 없기 때문이며 따라서 간선 정보는 그래프들을 유일하게 구별하기 위한 식별자로 사용된다. 지지도 연관성은 그래프 내 요소들이 얼마나 밀접한가를 지지도 정보를 가지고 판별하기 때문에 패턴을 구성하는 요소들 가운데 가장 큰 지지도 값을 갖는 간선과 가장 작은 지지도 값을 갖는 간선 사이의 비율을 구하면 연관성의 정도를 파악할 수 있다(S51). 이 비율이 높다는 것은 곧 그래프를 구성하는 간선들 가운데 가장 작은 지지도와 가장 큰 지지도 값을 갖는 간선 사이의 격차가 작다는 것을 나타내며, 따라서 이때의 그래프는 내부의 요소들이 전체적으로 비슷한 지지도 정보를 갖는다는 것을 알 수 있다. 반대의 경우는 최소 지지도 값을 갖는 간선과 최대 지지도 값을 갖는 간선 사이의 차이가 크다는 것을 의미하게 되고 이는 지지도 면에서 그래프 내 요소들 사이에 연관성이 떨어진다는 것을 나타낸다(S52).

$$SA(G_E) = \frac{\text{support}((e_1, e_2, e_3, \dots, e_n))}{\max_{1 \leq i \leq n}(\text{support}(e_i))}$$

지지도 연관성은 와 같은 수식으로 정의할 수 있다. 여기에서 SA는 지지도 연관성을 뜻하고, GE는 대상 그래프의 간선 집합을 의미한다. 분모는 간선 집합에 속하는 간선들 가운데 가장 큰 지지도 값을 얻기 위한 계산식을 의미하고, 분자는 그래프를 구성하는 모든 간선들에 대한 지지도, 즉 현재 부분 그래프 자체의 지지도 값을 나타내며 그래프의 지지도는 곧 그 그래프 내부의 요소들 가운데 최소값을 가지는 간선의 지지도와 동일하기 때문에 이에 같음한다. 지지도 연관성은 0부터 1사이의 실수에 해당하는 수치로 값으로 가지며 0에 가까울수록 연관성의 정도가 약해지고, 반대로 1에 가까워질수록 그 정도가 강해진다.

[0022] 첫 번째 확장으로 생성된 서브 그래프는 간선을 하나만 가지기 때문에, 이때의 지지도 연관성은 자연히 1이 되고 이 단계에서는 연관 정도를 고려할 필요가 없다. 다음 확장 단계에서는 간선과 해당하는 정점(순환그래프로 확장될 때는 간선만)이 확장되고, 이때부터는 위의 식에 따라 지지도 연관성을 구하게 된다. 그 후에는 계산된 값을 가지고 현재까지 확장된 부분 그래프를 프루닝, 즉 제거 할지 계속 그래프 확장을 진행할지에 대한 판단을 한다. 본 발명에서는 판단을 위한 기준으로써, 최소 지지도 연관성 임계값을 사용하며 현재 부분 그래프의 연관성 값이 이 임계값보다 크거나 같다면 이 그래프는 높은 지지도 연관성을 갖는 부분 그래프로 판별되고(S54), 반대로 이 임계값에 미치지 못한다면 이는 낮은 지지도 연관성을 갖는 부분 그래프로 판별되어(S53) 그 그래프

와 그로부터 파생되는 모든 상위 부분 그래프들 모두를 사전에 제거하게 된다(S36). 이 임계값은 사용자에 의해 자유롭게 설정될 수 있기 때문에 임계값을 어떻게 설정하냐에 따라 추출되는 빈발 부분 그래프의 연관성 정도를 조절할 수 있다. 지지도 연관성은 안티모노톤 법칙(Anti-monotone property)을 만족하기 때문에 위와 같이 해당하는 그래프들을 미리 제거해도 최종 결과물에는 아무런 영향을 미치지 않을뿐더러 제거를 통해 마이닝의 탐색공간을 효율적으로 줄일 수 있기 때문에 성능에 긍정적인 영향을 미치게 된다. 최소 지지도 연관성 임계값을 만족하는 빈발한 부분 그래프는 최종적으로 유효한 정보가 되므로, 패턴 추출과정을 통해 부분 그래프가 추출되게 된다.

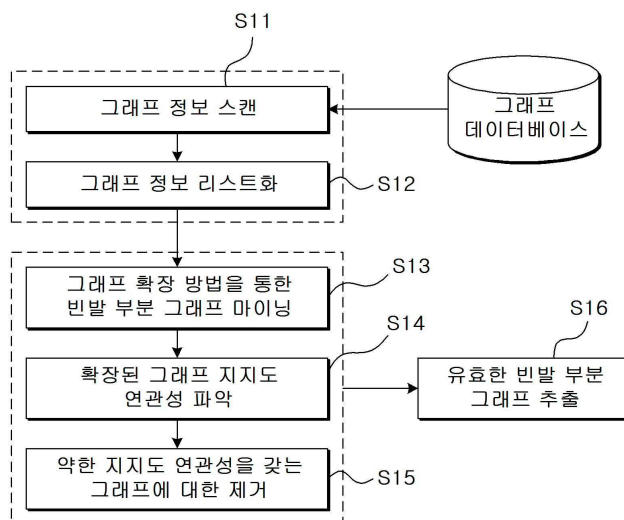
[0023] 추출 과정이 끝난 후에는 현재 확장된 부분 그래프에서 다시 그래프가 확장될 여지가 있는지를 임베딩 리스트를 통해 확인하게 되고, 확장이 가능하다면 다시 확장 과정으로 돌아가 해당하는 작업을 수행하고 이어서 확장된 그래프의 지지도 연관성 파악 과정, 프루닝 여부를 결정하는 과정, 패턴 추출 과정을 거치게 된다. 반면에 그래프가 더 이상 확장될 여지가 없다면 현재 단계를 마무리 짓고 다음 정점을 기준으로 놓고 다시 위의 작업들을 반복하게 된다. 그래프 데이터베이스 내에 존재하는 모든 정점에 대해 이러한 작업을 진행한 후에는 결과적으로 강한 지지도 연관성을 가지는 빈발 부분 그래프들만을 얻을 수 있다.

[0024] 도 6a는 최소 지지도 임계값이 2이고, 최소 지지도 연관성 임계값이 0.7인 그래프 데이터베이스 예제에 대한 빈발 부분 그래프 마이닝 결과를 나타낸 것이며, 도 6b는 그래프 데이터베이스 예제에 대해 지지도 관련성을 고려하지 않은 경우(a)와 고려한 경우(b)에 따른 마이닝 결과를 나타낸 것이고, 패턴의 지지도 연관성을 고려한 결과, 강한 연관성을 가지는 부분 그래프를 제외한 나머지 부분은 제거가 되기 때문에 전체적인 패턴의 수가 감소(14개 → 8개)한 것을 알 수 있다.

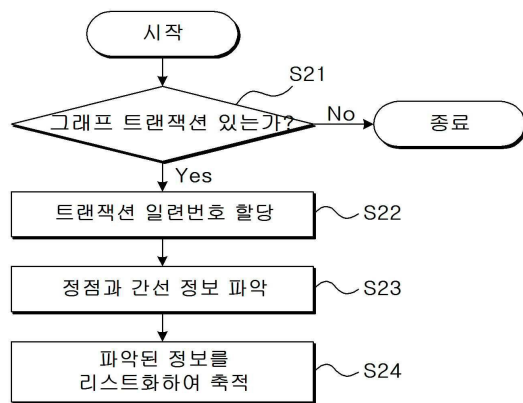
[0025] 이상에서 몇 가지 실시예를 들어 본 발명을 더욱 상세하게 설명하였으나, 본 발명은 반드시 이러한 실시예로 국한되는 것이 아니고 본 발명의 기술사상을 벗어나지 않는 범위 내에서 다양하게 변형 실시될 수 있다.

도면

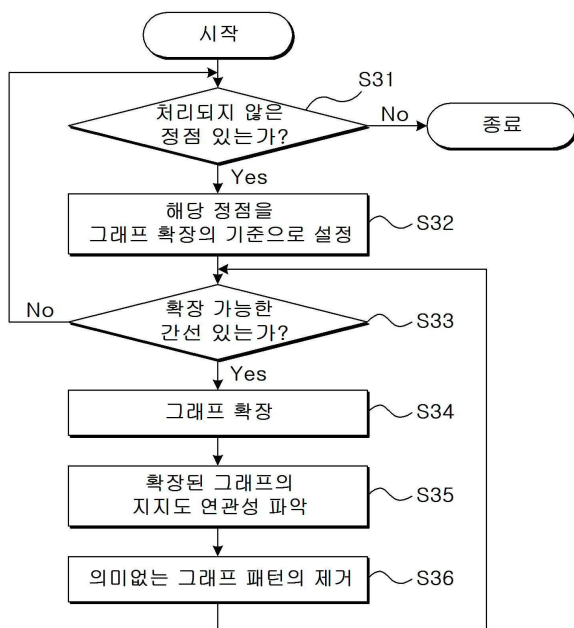
도면1



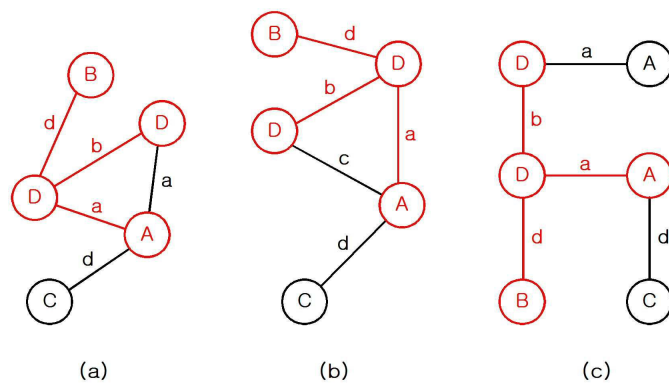
도면2



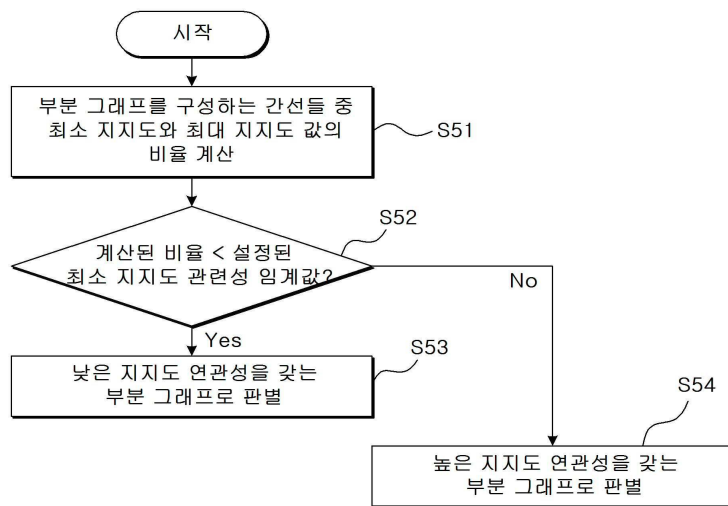
도면3



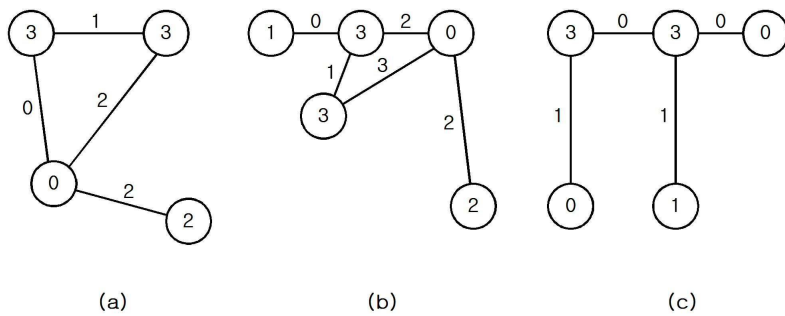
도면4



도면5



도면6a



도면6b

