(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0130014
(43) 공개일자 2014년11월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/00 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2014-0011382
(22) 출원일자 2014년01월29일
심사청구일자 2014년01월29일
(30) 우선권주장
1020130048703 2013년04월30일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스 내 (서천동, 경희대학교)
(72) 발명자
이영구
경기 용인시 기흥구 덕영대로 1732, 전자정보관 314호 (서천동, 경희대학교)
박기성
경기 부천시 원미구 원미로 123-4, E동 4호 (원미동, 백송연립)
(뒷면에 계속)
(74) 대리인
서재승

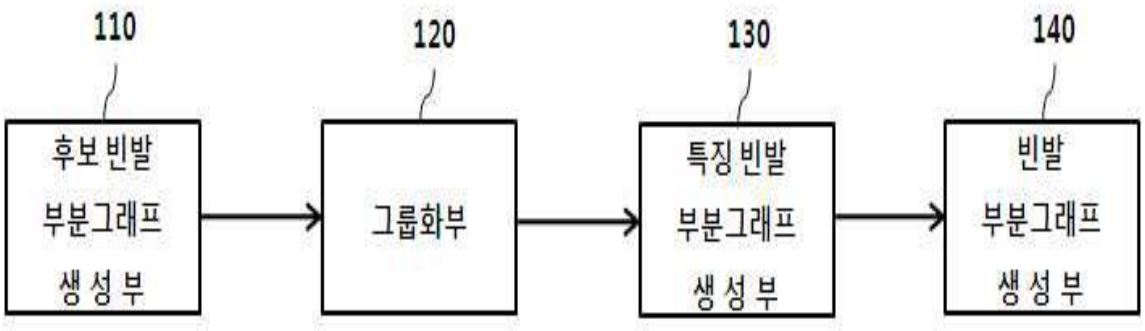
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **그래프 분류를 위한 빈발 부분그래프의 생성 방법**

(57) 요 약

본 발명은 그래프 데이터베이스에서 빈발 부분그래프를 생성하는 방법에 관한 것으로, 보다 구체적으로 그래프 데이터베이스를 구성하는 그래프의 모든 빈발 부분그래프를 생성하는 대신 구조적 유사도에 기초하여 유사한 빈발 부분그래프를 유사 그룹으로 그룹화하고 각 유사 그룹에서 분류력이 가장 높은 빈발 부분그래프로 빈발 부분그래프를 생성함으로써 그래프 마이닝시 대량의 빈발 부분그래프로 인해 그래프를 분류하는데 방대한 연산을 방지할 수 있고, 더욱이 구조적으로 유사한 빈발 부분그래프로 이루어진 유사 그룹에서 대표 빈발 부분그래프만을 생성함으로써 유사한 빈발 부분그래프가 중복 생성되는 것을 방지할 수 있으며, 한정된 수를 가지는 빈발 부분그래프를 이용하여 그래프 데이터베이스에 포함되어 있는 모든 그래프를 정확하게 마이닝할 수 있는 빈발 부분그래프를 생성할 수 있다.

대 표 도 - 도1



(72) 발명자

홍지혜

충남 천안시 동남구 충절로 42, 103동 105호 (신부동, 신부경남아너스빌아파트)

한용구

경기 용인시 기흥구 덕영대로 1732, 전자정보관 349호 (서천동, 경희대학교)

팜더안

경기 용인시 기흥구 덕영대로 1732, 전자정보대학 351호 (서천동, 경희대학교)

이 발명을 지원한 국가연구개발사업

과제고유번호	2012-0002304
부처명	교육과학기술부
연구관리전문기관	한국연구재단
연구사업명	일반연구자지원사업
연구과제명	그래프 마이닝을 위한 부분감독기반 특징선택 기술 연구
기 여 율	1/1
주관기관	경희대학교 산학협력단
연구기간	2010.05.01 ~ 2013.04.30

특허청구의 범위

청구항 1

그래프 데이터베이스에 저장되어 있는 다수 그래프 사이의 빈발 부분그래프를 생성하는 방법에 있어서,

상기 다수 그래프 사이에서 최소 지지도 이상으로 발현하는 후보 빈발 부분그래프를 생성하는 단계;

생성한 상기 후보 빈발 부분그래프 사이의 유사도로 상기 후보 빈발 부분그래프를 분류하여 상기 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화하는 단계;

상기 각 유사 그룹에서 분류력이 가장 높은 후보 빈발 부분그래프를 특징 빈발 부분그래프로 선택하는 단계; 및

상기 특징 빈발 부분그래프로 상기 다수 그래프의 빈발 부분그래프를 생성하는 단계를 포함하는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 2

제 1 항에 있어서, 상기 후보 빈발 부분그래프는

정규화 연산을 통해 최소 DFS 코드를 가지는 빈발 부분그래프인 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 3

제 2 항에 있어서, 상기 그룹화하는 단계는

초기 유사 그룹을 구성하는 후보 빈발 부분그래프의 특징 벡터의 평균값으로부터 중심 벡터를 계산하는 단계;

상기 중심 벡터와 상기 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산하는 단계; 및

상기 유사도에 기초하여 가장 높은 유사도를 가지는 초기 유사 그룹에 상기 후보 빈발 부분그래프를 포함시켜 유사 그룹을 생성하는 단계를 더 포함하며,

상기 유사 그룹이 동일하게 유지될 때까지 상기 중심 벡터를 계산하는 단계, 상기 유사도를 계산하는 단계 및 상기 유사 그룹을 생성하는 단계를 반복하는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 4

제 3 항에 있어서, 상기 초기 유사 그룹을 생성하는 단계는

생성한 상기 후보 빈발 부분그래프에서 초기 후보 빈발 부분그래프를 임의 선택하는 단계;

상기 초기 후보 빈발 부분그래프의 특징 벡터를 초기 중심 벡터로 설정하고, 상기 초기 중심 벡터와 상기 후보 빈발 부분그래프의 특징 벡터 사이의 초기 유사도를 계산하는 단계; 및

상기 초기 유사도에 기초하여 가장 높은 초기 유사도를 가지는 초기 후보 빈발 부분그래프에 상기 후보 빈발 부분그래프를 포함시켜 초기 유사 그룹을 생성하는 단계를 포함하는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 5

제 1 항에 있어서,

상기 후보 빈발 부분그래프의 분류력은 레이블 종류별 밀질 정도를 나타내는 분류값에 기초하여 계산되는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 6

제 5 항에 있어서,

상기 분류값(I)은 아래의 수학식(1)에 의해 계산되며,

[수학식 1]

$$I = H(Y) - H(Y \mid f_i)$$

여기서 H(Y)는 그래프의 레이블 종류별 밀집도를 나타내며, $H(Y \mid f_i)$는 후보 빈발 부분그래프의 레이블 종류별 밀집도를 나타내며,

여기서 H(Y)과 $H(Y \mid f_i)$ 각각 아래의 수학식(2)와 수학식(3)에 의해 계산되며,

[수학식 2]

$$H(Y) = \frac{c_a}{n} \log \frac{c_a}{n} + \frac{c_i}{n} \log \frac{c_i}{n}$$

[수학식 3]

$$H(Y \mid f_i) = P_{ci}\left(\frac{c_{ac_i}}{n_{ci}} \log \frac{c_{ac_i}}{n_{ci}} + \frac{c_{bc_i}}{n_{ci}} \log \frac{c_{bc_i}}{n_{ci}}\right) + P_{di}\left(\frac{c_{ad_i}}{n_{di}} \log \frac{c_{ad_i}}{n_{di}} + \frac{c_{bd_i}}{n_{di}} \log \frac{c_{bd_i}}{n_{di}}\right)$$

여기서 n은 그래프의 총수, c_a는 제1 레이블을 가지는 그래프의 수, c_i는 제2 레이블을 가지는 그래프의 수, P_{ci}, P_{di}, n_{ci}, c_{aci}, c_{bci}, n_{di}, c_{adi}, c_{bdi} 는 각각 i번째 후보 빈발 부분그래프의 발현 확률, 비발현 확률, 전체 발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수, 전체 비발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수인 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 7

제 1 항 내지 제 6 항 중 어느 한 항에 있어서, 상기 빈발 부분그래프의 생성 방법은

상기 특징 빈발 부분그래프로 이루어진 후보 동시 빈발 부분그래프를 생성하는 단계; 및

상기 후보 동시 빈발 부분그래프의 분류력에 기초하여 추가 동시 빈발 부분그래프를 결정하는 단계를 더 포함하며,

상기 추가 동시 빈발 부분그래프를 상기 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성하는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 8

제 7 항에 있어서,

상기 후보 동시 빈발 부분그래프의 분류력이 임계 분류력을 초과하는 경우 추가 동시 빈발 부분그래프로 결정하는 것을 특징으로 하는 빈발 부분그래프의 생성 방법.

청구항 9

그래프 데이터베이스에 저장되어 있는 다수 그래프 사이의 빈발 부분그래프를 생성하는 장치에 있어서,

상기 다수 그래프 각각에 대해 정규화 연산을 통해 최소 DFS 코드를 가지는 부분그래프를 생성하며, 생성한 부분그래프에서 최소 지지도 이상으로 발현하는 후보 빈발 부분그래프를 생성하는 후보 빈발 부분그래프 생성부;

생성한 상기 후보 빈발 부분그래프 사이의 유사도로 상기 후보 빈발 부분그래프를 분류하여 상기 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화하는 그룹화부; 및

상기 각 유사 그룹에서 분류력이 가장 높은 후보 빈발 부분그래프를 특징 빈발 부분그래프로 선택하고 상기 특징 빈발 부분그래프로 빈발 부분그래프를 생성하는 빈발 부분그래프 생성부를 포함하는 것을 특징으로 하는 빈발 부분그래프의 생성 장치.

청구항 10

제 9 항에 있어서, 상기 빈발 부분그래프의 생성 장치는

상기 특징 빈발 부분그래프로 이루어진 후보 동시 빈발 부분그래프를 생성하는 동시 빈발 부분그래프 생성부를 더 포함하며,

상기 빈발 부분그래프 생성부는 상기 후보 동시 빈발 부분그래프의 분류력에 기초하여 추가 동시 빈발 부분그래프를 결정하며, 상기 추가 동시 빈발 부분그래프를 상기 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성하는 것을 특징으로 하는 빈발 부분그래프의 생성 장치.

청구항 11

제 10 항에 있어서, 상기 그룹화부는

초기 유사 그룹을 구성하는 후보 빈발 부분그래프의 특징 벡터의 평균값으로부터 중심 벡터를 계산하는 중심 벡터 계산부;

상기 중심 벡터와 상기 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산하는 유사도 계산부;

상기 유사도에 기초하여 가장 높은 유사도를 가지는 초기 유사 그룹에 상기 후보 빈발 부분그래프를 포함시켜 유사 그룹을 생성하는 유사 그룹 생성부; 및

상기 초기 유사 그룹과 상기 유사 그룹을 비교하여 서로 동일한지 판단하는 유사 그룹 비교부를 포함하며,

상기 유사 그룹 비교부는 상기 초기 유사 그룹과 상기 유사 그룹이 동일하지 않은 경우, 상기 유사 그룹을 초기 유사 그룹으로 갱신하여 신규 유사 그룹을 생성 제어하는 것을 특징으로 하는 빈발 부분그래프의 생성 장치.

청구항 12

제 10 항에 있어서,

상기 후보 빈발 부분그래프의 분류력은 레이블 종류별 밀질 정도를 나타내는 분류값에 기초하여 계산되는 것을 특징으로 하는 빈발 부분그래프의 생성 장치.

청구항 13

제 12 항에 있어서,

상기 분류값(I)은 아래의 수학식(4)에 의해 계산되며,

[수학식 4]

$$I = H(Y) - H(Y \mid f_i)$$

여기서 H(Y)는 그래프의 레이블 종류별 밀집도를 나타내며, $H(Y \mid f_i)$는 후보 빈발 부분그래프의 레이블 종류별 밀집도를 나타내며,

여기서 H(Y)과 $H(Y \mid f_i)$ 각각 아래의 수학식(5)와 수학식(6)에 의해 계산되며,

[수학식 5]

$$H(Y) = \frac{c_a}{n} \log \frac{c_a}{n} + \frac{c_i}{n} \log \frac{c_i}{n}$$

[수학식 6]

$$H(Y \mid f_i) = P_{ci}\left(\frac{c_{ac_i}}{n_{ci}} \log \frac{c_{ac_i}}{n_{ci}} + \frac{c_{bc_i}}{n_{ci}} \log \frac{c_{bc_i}}{n_{ci}}\right) + P_{di}\left(\frac{c_{ad_i}}{n_{di}} \log \frac{c_{ad_i}}{n_{di}} + \frac{c_{bd_i}}{n_{di}} \log \frac{c_{bd_i}}{n_{di}}\right)$$

여기서 n은 그래프의 총수, c_a는 제1 레이블을 가지는 그래프의 수, c_i는 제2 레이블을 가지는 그래프의 수, P_{ci}, P_{di}, n_{ci}, c_{aci}, c_{bci}, n_{di}, c_{adi}, c_{bdi} 는 각각 i번째 후보 빈발 부분그래프의 발현 확률, 비발현 확률, 전체 발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수, 전체 비발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수인 것을 특징으로 하는 빈발 부분그래프의 생성 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 효과적인 그래프 분류를 위해 그래프 데이터베이스에서 분류력이 높은 빈발 부분그래프를 생성하는 방법에 관한 것으로, 보다 구체적으로 그래프 데이터베이스를 구성하는 그래프의 모든 빈발 부분그래프를 생성하는 대신 구조적 유사도에 기초하여 유사한 빈발 부분그래프를 유사 그룹으로 그룹화하고 각 유사 그룹에서 분류력이 가장 높은 빈발 부분그래프로 빈발 부분그래프를 생성함으로써 그래프 마이닝시 대량의 빈발 부분그래프로 인해 그래프를 분류하는데 방대한 연산을 방지할 수 있고, 더욱이 구조적으로 유사한 빈발 부분그래프로 이루어진 유사 그룹에서 대표 빈발 부분그래프만을 생성함으로써 유사한 빈발 부분그래프가 중복 생성되는 것을 방지할 수 있으며, 한정된 수를 가지는 빈발 부분그래프를 이용하여 그래프 데이터베이스에 포함되어 있는 모든 그래프를 정확하게 마이닝할 수 있는 빈발 부분그래프를 생성할 수 있다.

배 경 기 술

[0002] 그래프 분류는 화합물, XML, 웹 문서, 소셜 네트워크 등과 같은 그래프 구조를 갖는 데이터를 분류하는 기술이다. 그래프 분류를 위해 구조적 정보를 갖는 빈발 부분그래프들이 특징으로 널리 사용되고 있다. 여기서 빈발 부분그래프란 그래프 데이터베이스에서 최소 지지도 이상으로 발현되는 부분그래프를 의미하는데, 빈발 부분그래프는 그래프의 고유한 특성을 나타내어 그래프의 분류, 군집화, 인덱싱 등에서 사용된다.

[0003] 통상적으로 레이블된 기준 그래프로 이루어진 그래프 데이터베이스에서 빈발 부분그래프를 생성하고, 레이블된 그래프가 해당 빈발 부분그래프를 포함하는지에 대한 매트릭스를 생성하여 그래프의 분류, 인덱싱을 수행한다. 그러나 빈발 부분그래프는 그래프 데이터베이스에 따라 수만에서 수십만 개가 생성되기 때문에 그래프 데이터베이스의 분류에 도움이 되는 빈발 부분그래프를 제한된 수로 선택하는 것이 중요하다.

[0004] 종래 빈발 부분그래프를 생성하는 방식에 대한 연구가 수행되고 있다. 종래 빈발 부분그래프를 생성하는 방식 중 하나인 도약 탐색(leap search) 방식은 빈발 부분그래프 마이닝 단계에서 G-test 점수가 낮은 특징들에 대한 탐색공간을 줄임으로써 빠른 수행시간과 높은 분류 정확도를 가지는 빈발 부분그래프를 선택할 수 있다. 한편 다른 종래 빈발 부분그래프의 생성 방식인 모델 기반 탐색 트리(model based search tree)는 인포메이션 게인(information gain)이 높은 특징의 포함여부로 그래프 데이터베이스를 분할하는 과정을 반복하여 높은 분류 성능을 가지는 빈발 부분그래프를 선택할 수 있다.

[0005] 그러나 위에서 언급한 종래 빈발 부분그래프를 생성하는 방식은 여전히 대량의 빈발 부분그래프를 생성할 뿐만 아니라 유사한 구조적 특징을 가지는 빈발 부분그래프를 중복하여 생성하는 문제점을 가진다.

[0006] 따라서 종래 빈발 부분그래프를 생성하는 방식에서 빈발 부분그래프를 한정된 수로 생성하는 경우, 일정 특성을 분류하는 빈발 부분그래프로 치중되는 경향이 발생할 수 있으며 그래프 데이터베이스에 포함되어 있는 모든 그래프를 높은 분류력으로 분류하기 곤란하다는 문제점을 가진다.

발명의 내용

해결하려는 과제

[0007] 따라서 본 발명이 이루고자 하는 목적은 그래프 데이터베이스에서 생성한 후보 빈발 부분 그래프를 구조적 유사도에 따라 유사 그룹으로 그룹화하고 유사 그룹별로 빈발 부분그래프를 생성하여 구조적 유사한 빈발 부분그래프가 중복되어 생성되는 것을 방지하는 빈발 부분그래프의 생성 방법을 제공하는 것이다.

[0008] 한편, 본 발명이 이루고자 하는 다른 목적은 유사 그룹에서 가장 분류력이 높은 후보 빈발 부분그래프로 빈발 부분그래프를 생성하여 한정된 수를 가지며 분류력이 높은 빈발 부분 그래프의 생성 방법을 제공하는 것이다.

[0009] 또한 본 발명이 이루고자 하는 또 다른 목적은 각 유사 그룹에서 선택된 후보 빈발 부분그래프로 이루어진 동시

발생 빈발 부분그래프의 분류력을 고려하여 높은 분류력을 가지는 동시발생 빈발 부분그래프를 포함하여 빈발 부분그래프를 생성하는 방법을 제공하는 것이다.

과제의 해결 수단

[0010] 본 발명의 목적을 달성하기 위하여 본 발명에 따른 빈발 부분그래프의 생성 방법은 다수 그래프 사이에서 최소 지지도 이상으로 발현하는 후보 빈발 부분그래프를 생성하는 단계와, 생성한 후보 빈발 부분그래프 사이의 유사도로 후보 빈발 부분그래프를 분류하여 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화하는 단계와, 각 유사 그룹에서 분류력이 가장 높은 후보 빈발 부분그래프를 특징 빈발 부분그래프로 선택하는 단계와, 특징 빈발 부분그래프로 다수 그래프의 빈발 부분그래프를 생성하는 단계를 포함하는 것을 특징으로 한다.

[0011] 여기서 후보 빈발 부분그래프는 정규화 연산을 통해 최소 DFS 코드를 가지는 빈발 부분그래프인 것을 특징으로 한다.

[0012] 본 발명에 따른 그룹화 단계의 일 예는 초기 유사 그룹을 생성하는 단계와, 초기 유사 그룹을 구성하는 후보 빈발 부분그래프의 특징 벡터의 평균값으로부터 중심 벡터를 계산하는 단계와, 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산하는 단계와, 유사도에 기초하여 가장 높은 유사도를 가지는 초기 유사 그룹에 후보 빈발 부분그래프를 포함시켜 유사 그룹을 생성하는 단계를 더 포함하며, 유사 그룹이 동일하게 유지될 때까지 중심 벡터를 계산하는 단계, 유사도를 계산하는 단계 및 유사 그룹을 생성하는 단계를 반복하는 것을 특징으로 한다.

[0013] 여기서 초기 유사 그룹을 생성하는 단계는 생성한 후보 빈발 부분그래프에서 초기 빈발 부분그래프를 임의 선택하는 단계와, 초기 빈발 부분그래프의 특징 벡터를 초기 중심 벡터로 설정하고 초기 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 초기 유사도를 계산하는 단계와, 초기 유사도에 기초하여 가장 높은 초기 유사도를 가지는 초기 빈발 부분그래프에 후보 빈발 부분그래프를 포함시켜 초기 유사 그룹을 생성하는 단계를 포함하는 것을 특징으로 한다.

[0014] 바람직하게, 후보 빈발 부분그래프의 분류력은 레이블 종류별 밀질 정도를 나타내는 분류값에 기초하여 계산되는 것을 특징으로 하는데, 분류값(I)은 아래의 수학식(1)에 의해 계산되며,

[0015] [수학식 1]

[0016] $$I = H(Y) - H(Y \mid f_i)$$

[0017] 여기서 H(Y)는 그래프의 레이블 종류별 밀집도를 나타내며, $H(Y \mid f_i)$는 후보 빈발 부분그래프의 레이블 종류별 밀집도를 나타내며,

[0018] 여기서 H(Y)과 $H(Y \mid f_i)$ 각각 아래의 수학식(2)와 수학식(3)에 의해 계산되며,

[0019] [수학식 2]

[0020] $$H(Y) = \frac{c_a}{n}\log\frac{c_a}{n} + \frac{c_i}{n}\log\frac{c_i}{n}$$

[0021] [수학식 3]

[0022] $$H(Y \mid f_i) = P_{ci}\left(\frac{c_{ac_i}}{n_{ci}}\log\frac{c_{ac_i}}{n_{ci}} + \frac{c_{bc_i}}{n_{ci}}\log\frac{c_{bc_i}}{n_{ci}}\right) + P_{di}\left(\frac{c_{ad_i}}{n_{di}}\log\frac{c_{ad_i}}{n_{di}} + \frac{c_{bd_i}}{n_{di}}\log\frac{c_{bd_i}}{n_{di}}\right)$$

[0023] 여기서 n은 그래프의 총수, c_a는 제1 레이블을 가지는 그래프의 수, c_i는 제2 레이블을 가지는 그래프의 수, P_{ci}, P_{di}, n_{ci}, c_{aci}, c_{bci}, n_{di}, c_{adi}, c_{bdi} 는 각각 i번째 후보 빈발 부분그래프의 발현 확률, 비발현 확률, 전체 발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수, 전체 비발현 빈도 수, 제1 레이블의 발현 빈도 수, 제2 레이블의 발현 빈도 수인 것을 특징으로 한다.

[0024] 바람직하게, 본 발명에 따른 빈발 부분그래프의 생성 방법은 특징 빈발 부분그래프로 이루어진 후보 동시 빈발

부분그래프를 생성하는 단계와, 후보 동시 빈발 부분그래프의 분류력에 기초하여 추가 동시 빈발 부분그래프를 결정하는 단계를 더 포함하며, 추가 동시 빈발 부분그래프를 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성하는 것을 특징으로 한다.

[0025] 한편, 본 발명의 일 실시예에 따른 빈발 부분그래프의 생성 장치는 다수 그래프 각각에 대해 정규화 연산을 통해 최소 DFS 코드를 가지는 부분그래프를 생성하며, 생성한 부분그래프에서 최소 지지도 이상으로 발현하는 후보 빈발 부분그래프를 생성하는 후보 빈발 부분그래프 생성부와, 생성한 후보 빈발 부분그래프 사이의 유사도로 후보 빈발 부분그래프를 분류하여 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화하는 그룹화부와, 각 유사 그룹에서 분류력이 가장 높은 후보 빈발 부분그래프를 특징 빈발 부분그래프로 선택하고 특징 빈발 부분그래프로 빈발 부분그래프를 생성하는 빈발 부분그래프 생성부를 포함하는 것을 특징으로 한다.

[0026] 바람직하게, 본 발명의 일 실시예에 따른 빈발 부분그래프의 생성 장치는 특징 빈발 부분그래프로 이루어진 후보 동시 빈발 부분그래프를 생성하는 동시 빈발 부분그래프 생성부를 더 포함하며, 빈발 부분그래프 생성부는 후보 동시 빈발 부분그래프의 분류력에 기초하여 추가 동시 빈발 부분그래프를 결정하며, 추가 동시 빈발 부분그래프를 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성하는 것을 특징으로 한다.

발명의 효과

[0027] 본 발명에 따른 빈발 부분그래프의 생성 방법은 다음과 같은 다양한 효과를 가진다.

[0028] 첫째, 본 발명에 따른 빈발 부분그래프의 생성 방법은 그래프 데이터베이스에서 생성한 후보 빈발 부분 그래프를 구조적 유사도에 따라 유사 그룹으로 그룹화하고 유사 그룹별로 빈발 부분그래프를 생성함으로써, 구조적 유사한 빈발 부분그래프가 중복되어 생성되는 것을 방지한다.

[0029] 둘째, 본 발명에 따른 빈발 부분그래프의 생성 방법은 유사 그룹에서 가장 분류력이 높은 후보 빈발 부분그래프로 빈발 부분그래프를 생성함으로써, 한정된 수로 분류력이 높은 빈발 부분 그래프를 생성할 수 있다.

[0030] 셋째, 본 발명에 따른 빈발 부분그래프의 생성 방법은 각 유사 그룹에서 선택된 후보 빈발 부분그래프로 이루어진 동시발생 빈발 부분그래프의 분류력을 고려하여 높은 분류력을 가지는 동시발생 빈발 부분그래프를 포함하여 빈발 부분그래프를 생성할 수 있다.

[0031] 넷째, 본 발명에 따른 빈발 부분그래프의 생성 방법은 각 유사 그룹에서 선택된 후보 빈발 부분그래프만을 이용하여 동시 발생 빈발 부분그래프를 생성함으로써, 동시발생 빈발 부분그래프의 조합 생성에 소요되는 연산량을 줄일 수 있다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 일 실시예에 따른 빈발 부분그래프 생성 장치를 설명하기 위한 기능 블록도이다.

도 2는 본 발명에 따른 후보 빈발 부분그래프 생성부를 설명하기 위한 기능 블록도이다.

도 3은 그룹화부를 보다 구체적으로 설명하기 위한 기능 블록도이다.

도 4는 본 발명에 따른 특징 빈발 부분그래프 생성부를 설명하기 위한 기능 블록도이다.

도 5는 본 발명에 따른 빈발 부분그래프의 생성 방법을 설명하기 위한 흐름도이다.

도 6은 동시 발생 빈발 부분그래프의 분류력을 설명하기 위한 도면이다.

도 7은 본 발명에 따른 유사 그룹을 생성하는 단계를 보다 구체적으로 설명하기 위한 도면이다.

도 8 내지 도 10은 유사 그룹을 생성하는 일 예를 설명하기 위하 도면이다.

발명을 실시하기 위한 구체적인 내용

[0033] 이하 첨부한 도면을 참고로 본 발명에 따른 빈발 부분그래프의 생성 방법 및 그 장치에 대해 보다 구체적으로 설명한다.

[0034] 도 1은 본 발명의 일 실시예에 따른 빈발 부분그래프 생성 장치를 설명하기 위한 기능 블록도이다.

[0035] 도 1을 참고로 보다 구체적으로 살펴보면, 후보 빈발 부분그래프 생성부(110)는 그래프 데이터베이스에 저장되어 있는 그래프 각각에 대해 정규화 연산을 수행하여 최소 DFS 코드를 가지는 부분그래프를 생성하며, 생성한 부분그래프에서 최소 지지도 이상으로 발현하는 후보 빈발 부분그래프를 생성한다.

[0036] 바람직하게, 그래프 데이터베이스에 저장되어 있는 그래프는 클래스 레이블된 그래프로 클래스 레이블이란 특정 대상에 대한 그래프의 활성 또는 비활성의 성질을 나타내는 정보로, 예를 들어 암의 치료물질에서 항암 성질을 가지는 치료물질의 화학 성분 그래프는 활성 클래스로 레이블되며, 항암 성질을 가지지 않는 치료물질의 화학 성분 그래프는 비활성 클래스로 레이블된다.

[0037] 그룹화부(120)는 생성한 후보 빈발 부분그래프 사이의 유사도에 기초하여 서로 유사한 구조적 특징을 가지는 후보 빈발 부분그래프를 분류하고, 분류한 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화한다. 여기서 구조적 특징이란 후보 빈발 부분그래프가 그래프에서 발현 여부를 나타내는 특징벡터를 의미하며, 서로 특징벡터가 유사한 후보 빈발 부분그래프를 설정된 수의 유사 그룹으로 그룹화한다.

[0038] 특징 빈발 부분그래프 생성부(130)는 그룹화된 설정된 수의 각 유사 그룹에서 각 유사 그룹을 대표하는 특징 빈발 부분그래프를 선택하고, 빈발 부분그래프 생성부(140)는 선택한 특징 빈발 부분그래프로 빈발 부분그래프를 생성한다. 여기서 특징 빈발 부분그래프는 각 유사 그룹에서 분류력이 가장 높은 후보 빈발 부분그래프인 것을 특징으로 한다.

[0039] 바람직하게, 빈발 부분그래프 생성부(140)는 특징 빈발 부분그래프의 조합으로 이루어진 동시 빈발 부분그래프를 생성하며, 생성한 동시 빈발 부분그래프의 분류력이 높은, 예를 들어 임계 분류값을 초과하는 동시 빈발 부분그래프를 선택한 특징 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성한다.

[0040] 도 2는 본 발명에 따른 후보 빈발 부분그래프 생성부를 설명하기 위한 기능 블록도이다.

[0041] 도 2를 참고로 보다 구체적으로 살펴보면, 부분그래프 생성부(111)는 그래프 데이터베이스에 저장되어 있는 각 그래프별로 최소 DFS 코드를 가지는 부분그래프를 생성한다.

[0042] 최소 DFS 코드를 가지는 부분그래프를 생성하는 부분그래프 생성부(111)에 대해 자세히 살펴보면, 부분그래프의 DFS 코드는 부분그래프를 구성하는 각 에지의 확장 순서로 깊이 우선 탐색(DFS) 방식으로 생성되는데, 부분그래프를 구성하는 각 에지의 DFS 코드는 에지를 구성하는 2개 버텍스(vertex) 식별자, 2개 버텍스 레이블 및 2개 버텍스를 연결하는 에지 레이블로 구성된다. 여기서 부분그래프의 DFS 코드는 버텍스 식별자의 순서로, 즉 깊이 우선 탐색 방식으로 에지를 확장하며 확장한 에지에 대해 각각 DFS 코드를 생성한다. 부분그래프를 구성하는 각 에지의 DFS 코드는 아래의 수학식(1)와 같이 표현된다.

[0043] [수학식 1]

[0044] $\text{DFS 코드}=\{i, j, l(v_i), l(e(v_i, v_j)), l(v_j)\}$

[0045] 여기서 i, j는 각각 에지(e)를 구성하는 2개 버텍스(vertex) 식별자, $l(v_i)$와 $l(v_j)$는 각각 2개 버텍스 레이블, $l(e(v_i, v_j))$는 2개 버텍스를 연결하는 에지 레이블이다.

[0046] 생성한 부분그래프의 DFS 코드를 정규화 연산하여 생성한 부분그래프가 최소 DFS 코드를 가지는 정규 그래프인지 판단한다. 판단 결과 생성한 부분그래프가 정규 그래프인 경우 부분그래프에서 에지를 1개씩 확장하여 부분그래프로부터 생성되는 새로운 부분그래프를 생성한다(S14). 생성한 신규 부분그래프에 대해 앞서 설명한 과정을 반복 수행하여 그래프 데이터베이스에 존재하는 부분그래프를 생성한다.

[0047] 부분그래프의 DFS 코드 중 최소 DFS 코드를 부분그래프의 정규 코드 형태로 사용하여 부분그래프를 유일하게 표현한다. 부분그래프를 정규 코드 형태로 표현하는 것을 정규화 연산(canonized operation)이라 정의하고 정규 코드 형태로 표현되는 부분그래프를 정규 그래프라 언급한다. 정규화 연산을 위하여 부분그래프의 최소 DFS코드를 찾기 위한 과정은 다음과 같다. 부분그래프의 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 중에

서 가장 작게 표현되는 최소 DFS코드를 찾는다.

[0048] 생성한 부분그래프가 최소 DFS 코드를 가지는 정규 그래프가 아닌 경우, 생성한 부분그래프를 제거하여 동일한 부분그래프가 중복하여 확장 생성되는 것을 방지한다.

[0049] 최소 지지도 계산부(113)는 그래프 데이터베이스에 저장되어 있는 그래프에서 생성한 부분그래프가 발현하는 지지도 횟수를 카운트하여 각 부분그래프의 지지도를 계산하고, 후보 빈발 부분그래프 결정부(115)는 아래의 수학식(2)와 같이 부분그래프의 지지도에 기초하여 최소 지지도(Minsup) 이상의 지지도를 가지는 부분그래프를 후보 빈발 부분그래프(FG)로 결정한다.

[0050] [수학식 2]

[0051] $$FG=\frac{|g\subseteq G|}{|GD|}>MinSup$$

[0052] 여기서 $|g\subseteq G|$ 는 부분그래프(g)를 가지는 그래프(G)의 개수이며 $|GD|$ 는 그래프 데이터베이스에 존재하는 그래프의 개수이다.

[0053] 도 3은 그룹화부를 보다 구체적으로 설명하기 위한 기능 블록도이다.

[0054] 도 3을 참고로 보다 구체적으로 살펴보면, 초기 그룹 생성부(121)는 생성하고자 하는 유사 그룹의 수에 상응하는 수의 후보 빈발 부분그래프를 임의로 선택하여 초기 유사 그룹을 생성한다.

[0055] 중심 벡터 계산부(123)는 초기 유사 그룹을 구성하는 후보 빈발 부분그래프의 특징 벡터의 평균값으로부터 중심 벡터를 계산하고 유사도 계산부(125)는 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산한다. 바람직하게, 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도는 맨하튼 거리(Manhattan distance)를 이용하여 계산한다.

[0056] 유사 그룹 생성부(127)는 계산한 유사도에 기초하여 가장 높은 유사도를 가지는 초기 유사 그룹에 후보 빈발 부분그래프를 포함시켜 유사 그룹을 생성한다. 바람직하게, 제1 초기 유사 그룹과의 유사도와 제2 초기 유사 그룹과의 유사도가 서로 동일한 경우, 제1 초기 유사 그룹 또는 제2 초기 유사 그룹 중 어느 하나에 임의로 후보 빈발 부분그래프를 추가하여 유사 그룹을 생성한다.

[0057] 유사 그룹 비교부(129)는 초기 유사 그룹과 생성한 유사 그룹을 비교하여 서로 동일한지 판단한다. 유사 그룹 비교부(129)는 초기 유사 그룹과 생성한 유사 그룹이 동일한 경우, 생성한 유사 그룹을 최종 유사 그룹으로 결정한다. 그러나 초기 유사 그룹과 생성한 유사 그룹이 서로 동일하지 않은 경우, 유사 그룹을 초기 유사 그룹으로 갱신한다. 갱신한 초기 유사 그룹에 대해 다시 중심 벡터를 계산하고 갱신한 초기 유사 그룹의 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산하여 초기 유사 그룹과 생성한 유사 그룹이 동일할 때까지 유사 그룹을 반복 생성한다.

[0058] 도 4는 본 발명에 따른 특징 빈발 부분그래프 생성부를 설명하기 위한 기능 블록도이다.

[0059] 도 4를 참고로 보다 구체적으로 살펴보면, 그래프 밀집도 계산부(131)는 그래프 데이터베이스에 저장되어 있는 그래프들의 클래스 레이블 종류별 밀집도를 계산하며, 부분그래프 밀집도 계산부(133)는 후보 빈발 부분그래프의 클래스 레이블 종류별 밀집도를 계산한다. 그래프의 클래스 레이블 종류별 밀집도(H(Y))와 후보 빈발 부분그래프(fi)의 클래스 레이블 종류별 밀집도($H(Y|f_i)$)는 각각 아래의 수학식(3)과 수학식(4)에 의해 계산될 수 있다.

[0060] [수학식 3]

[0061] $$H(Y)=\frac{c_a}{n}\log\frac{c_a}{n}+\frac{c_i}{n}\log\frac{c_i}{n}$$

[0062] [수학식 4]

$$H(Y \mid f_i)=P_{ci}\left(\frac{c_{ac_i}}{n_{ci}}\log\frac{c_{ac_i}}{n_{ci}}+\frac{c_{bc_i}}{n_{ci}}\log\frac{c_{bc_i}}{n_{ci}}\right)+P_{di}\left(\frac{c_{ad_i}}{n_{di}}\log\frac{c_{ad_i}}{n_{di}}+\frac{c_{bd_i}}{n_{di}}\log\frac{c_{bd_i}}{n_{di}}\right)$$

[0063]

[0064] 여기서 n은 그래프의 총수, c_a는 제1 클래스 레이블을 가지는 그래프의 수, c_i는 제2 클래스 레이블을 가지는 그래프의 수, P_{ci}, P_{di}, n_{ci}, c_{aci}, c_{bci}, n_{di}, c_{adi}, c_{bdi} 는 각각 그래프에서 i번째 후보 빈발 부분그래프의 발현 확률, 비발현 확률, 제1 클래스 레이블과 제2 클래스 레이블의 전체 발현 빈도 수, 제1 클래스 레이블의 발현 빈도 수, 제2 클래스 레이블의 발현 빈도 수, 제1 클래스 레이블과 제2 클래스 레이블의 전체 비발현 빈도 수, 제1 클래스 레이블의 발현 빈도 수, 제2 클래스 레이블의 발현 빈도 수인 것을 특징으로 한다.

[0065] 분류값 계산부(135)는 그래프의 클래스 레이블 종류별 밀집도와 후보 빈발 부분그래프의 클래스 레이블 종류별 밀집도로부터 각 후보 빈발 부분그래프의 분류값(I)을 계산한다. 바람직하게, 분류값(I)은 아래의 수학식(5)에 의해 계산될 수 있다.

[0066] [수학식 5]

$$I=H(Y)-H(Y \mid f_i)$$

[0067]

[0068] 특징 빈발 부분그래프 선택부(137)는 계산한 분류값을 분류력으로 하여 각 유사 그룹에서 가장 높은 분류력을 가지는 후보 빈발 부분그래프를 특징 빈발 부분그래프로 선택한다.

[0069] 도 5는 본 발명에 따른 빈발 부분그래프의 생성 방법을 설명하기 위한 흐름도이다.

[0070] 도 5를 참고로 보다 구체적으로 살펴보면, 그래프 데이터베이스에 저장되어 있는 각 그래프를 정규화 연산하여 최소 DFS 코드를 가지는 부분그래프를 생성하고, 최소 지지도 이상으로 발현하는 부분그래프를 후보 빈발 부분그래프로 생성한다(S110).

[0071] 생성한 후보 빈발 부분그래프의 특징 벡터에 기초하여 서로 유사한 구조적 특징을 가지는 후보 빈발 부분그래프를 그룹화하여 유사 그룹을 생성한다(S120).

[0072] 생성한 각 유사 그룹을 구성하는 각 후보 빈발 부분그래프의 분류값을 계산하고 계산한 분류값에 기초하여 가장 높은 분류값을 가지는 후보 빈발 부분그래프를 각 유사 그룹을 대표하는 특징 빈발 부분그래프로 선택한다(S130).

[0073] 선택한 특징 빈발 부분그래프로 그래프 데이터베이스에 저장되어 있는 그래프들의 빈발 부분그래프로 생성한다(S150).

[0074] 바람직하게, 선택한 특징 빈발 부분그래프의 조합으로 이루어진 동시 빈발 부분그래프를 선택한 특징 빈발 부분그래프에 추가하여 빈발 부분그래프를 생성할 수 있다(S140).

[0075] 특징 빈발 부분그래프들의 동시 발생을 특징으로 사용하면 더 높은 분류력을 가질 수 있는데, 도 6에 도시되어 있는 바와 같이 원으로 표현된 그래프는 특징 빈발 부분그래프(A-B)만을 가지고 있는 그래프들을, 삼각형으로 표현된 그래프는 특징 빈발 부분그래프(C-D)만을 가지고 있는 그래프들을, 사각형으로 표현된 그래프는 특징 빈발 부분그래프(A-B, C-D)를 모두 가지고 있는 그래프들을 나타낸다.

[0076] 특징 빈발 부분그래프(A-B)는 포지티브 레이블 그래프와 네거티브 레이블 그래프를 특징 빈발 부분그래프(A-B)로 나누어진 영역의 동일한 영역에 분포시키고 언레이블 그래프도 고르게 분포시키지 못하므로 분류력이 낮다. 한편, 특징 빈발 부분그래프(C-D)는 포지티브 레이블 그래프와 네거티브 레이블 그래프를 특징 빈발 부분그래프(C-D)로 나누어진 영역의 서로 다른 영역에 분포시키지만 언레이블 그래프를 고르게 분포시키지 못하므로 분류력이 높지 않다. 한편, 특징 빈발 부분그래프(A-B, C-D)의 조합으로 이루어진 동시발생 빈발 부분그래프의 경우 포지티브 레이블 그래프와 네거티브 레이블 그래프를 서로 다른 영역에 분포시키며 언레이블 그래프를 고르게 분포시키므로 높은 분류력을 가짐을 알 수 있다.

[0077] 바람직하게, 동시 빈발 부분그래프에서 임계 분류값 이상을 가지는 동시 빈발 부분그래프를 빈발 부분그래프에

추가함으로써, 보다 높은 분류력을 가지는 빈발 부분그래프를 생성할 수 있다.

[0078] 바람직하게, 본 발명에 따른 빈발 부분그래프의 생성 방법에서는 모든 빈발 부분그래프의 조합으로 생성되는 동시 빈발 부분그래프에 대해 분류값을 계산하여 빈발 부분그래프에 추가할 것인지 판단하는 것이 아니라, 각 유사 그룹을 대표하는 특징 빈발 부분그래프의 조합으로 생성되는 동시 빈발 부분그래프의 분류값에 기초하여 빈발 부분그래프에 추가할 동시 빈발 부분그래프를 판단함으로써, 동시 빈발 부분그래프를 생성하거나 추가할 동시 빈발 부분그래프를 판단하는데 많은 연산량을 필요로 하지 않다. 또한 각 유사 그룹을 대표하는 특징 빈발 부분그래프의 조합으로 동시 빈발 부분그래프를 생성함으로써, 분류력이 높은 동시 빈발 부분그래프를 빈발 부분그래프에 추가할 수 있다. 동시 빈발 부분그래프의 분류력은 위에서 설명한 분류값으로 동일하게 계산될 수 있다.

[0079] 도 7은 본 발명에 따른 유사 그룹을 생성하는 단계를 보다 구체적으로 설명하기 위한 도면이며, 도 8 내지 도 10은 유사 그룹을 생성하는 일 예를 설명하기 위하 도면이다.

[0080] 먼저, 초기 유사 그룹을 생성하는 과정을 살펴보면 생성한 후보 빈발 부분그래프에서 생성하고자 하는 유사 그룹 수에 상응하는 초기 후보 빈발 부분그래프를 임의 선택한다. 예를 들어 도 8(a)와 같이 그래프 데이터베이스에 G1, G2, G3의 그래프가 포함되어 있으며 G1, G2, G3 그래프에서 생성한 후보 빈발 부분그래프가 각각 f1, f2, f3, f4, f5이고 유사 그룹을 2개 생성하고자 하는 경우, 후보 빈발 부분그래프 중에서 f1, f2를 각각 초기 후보 빈발 부분그래프로 임의 선택한다.

[0081] 초기 빈발 부분그래프 f1의 특징 벡터(1,1,0)와 초기 빈발 부분그래프 f2의 특징 벡터(0,1,1)를 각각 초기 중심 벡터(C_1, C_2)로 설정하고, 초기 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 초기 유사도를 계산한다. 유사도(S)는 초기 중심 벡터와 후보 빈발 부분그래프 특징 벡터 사이의 거리로 아래의 수학식(6)과 같이 계산된다.

[0082] [수학식 6]

$$S=\frac{|\ v_{c1}-v_{i1}\ |+|\ v_{c2}-v_{i2}\ |+|\ v_{c3}-v_{i3}\ |}{3}$$

[0083]

[0084] 여기서 v_{ci}, v_{ci}, v_{ci}는 각각 중심 벡터의 값을 의미하며, v_{1i}, v_{2i}, v_{3i}는 후보 빈발 부분그래프(fi)의 특징 벡터값을 의미한다.

[0085] 도 8(b)는 초기 중심 벡터(C_1, C_2)와 후보 빈발 부분그래프 f1, f2, f3, f4, f5의 특징벡터 사이의 유사도를 나타낸 것이며, 도 8(c)는 유사도에 기초하여 후보 빈발 부분그래프 f1, f3, f4는 초기 중심 벡터(C_1)에 속하여 초기 유사 그룹(C1)을 생성하고, 후보 빈발 부분그래프 f2, f5는 초기 중심 벡터(C_2)에 속하여 초기 유사 그룹(C2)을 생성한다.

[0086] 이와 같이 생성한 초기 유사 그룹을 이용하여 초기 유사 그룹을 구성하는 후보 빈발 부분그래프의 특징 벡터의 평균값으로부터 중심 벡터를 계산한다(S121). 예를 들어 초기 유사 그룹(C1)을 구성하는 후보 빈발 부분그래프 f1, f3, f4의 특징 벡터의 평균값은 (1+1+0)/3, (1+1+0)/3, (0+1+0)/3으로 계산된다.

[0087] 위에서 설명한 수학식(6)과 같이 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 계산하고(S123), 유사도에 기초하여 가장 높은 유사도를 가지는 초기 유사 그룹에 후보 빈발 부분그래프를 포함시켜 유사 그룹을 생성한다(S125).

[0088] 도 9(a)는 중심 벡터와 후보 빈발 부분그래프의 특징 벡터 사이의 유사도를 도시하고 있으며, 도 9(b)는 후보 빈발 부분그래프 f1, f4를 중심 벡터(C_1)에 속하여 생성한 유사 그룹(C1)과 후보 빈발 부분그래프 f2, f3, f5를 중심 벡터(C_2)에 속하여 생성한 유사 그룹(C2)의 일 예를 도시하고 있다.

[0089] 생성한 유사 그룹과 초기 유사 그룹이 동일한지 판단하고(S127), 생성한 유사 그룹과 초기 유사 그룹이 동일한 경우 생성한 유사 그룹을 최종 유사 그룹으로 결정 생성한다(S129).

[0090] 그러나 생성한 유사 그룹과 초기 유사 그룹이 서로 동일하지 않은 경우, 예를 들어 도 8(c)와 도 9(b)에 도시되어 있는 유사 그룹과 초기 유사 그룹이 서로 상이한 경우, 생성한 유사 그룹을 초기 유사 그룹으로 갱신하고 갱신한 초기 유사 그룹에 대해 앞서 설명한 S121 내지 S127 단계를 반복한다. 도 10(a)는 갱신한 초기 유사 그룹으로 생성한 중심 벡터와 후보 빈발 부분그래프 사이의 유사도를 도시하고 있으며, 도 10(b)는 계산한 유사도에 기초하여 생성한 유사 그룹(C1, C2)을 도시하고 있다. 도 9(b)와 도 10(b)에 도시되어 있는 바와 같이 갱신한 초기 유사 그룹과 생성한 유사 그룹이 서로 동일하므로, 생성한 유사 그룹을 최종 유사 그룹으로 생성한다.

[0091] 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.

[0092] 상기 컴퓨터로 읽을 수 있는 기록 매체는 마그네틱 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.

[0093] 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

[0094] 110: 후보 빈발 부분그래프 생성부 120: 그룹화부

130: 특징 후보 빈발 부분그래프 생성부 140: 빈발 부분그래프 생성부

111: 부분그래프 생성부 113: 최소 지지도 계산부

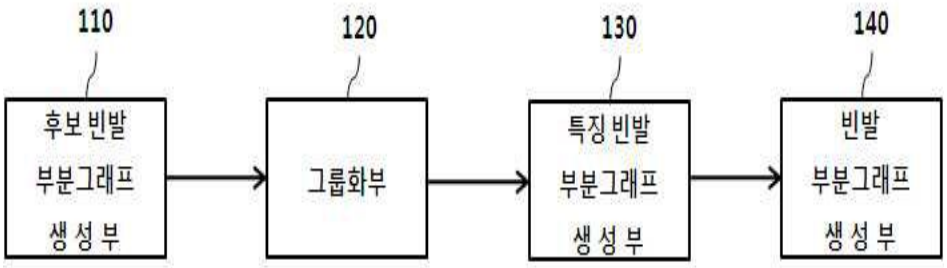
115: 후보 빈발 부분그래프 결정부 121: 초기 그룹 생성부

123: 중심 벡터 계산부 125: 유사도 계산부

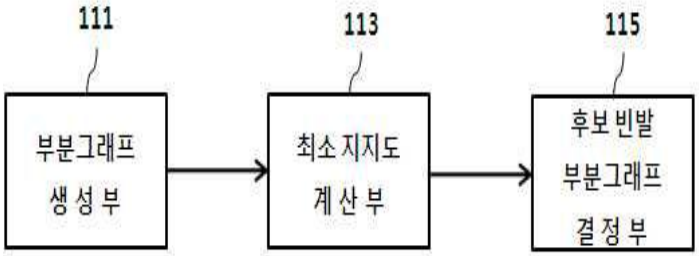
127: 유사 그룹 생성부 129: 유사 그룹 비교부

도면

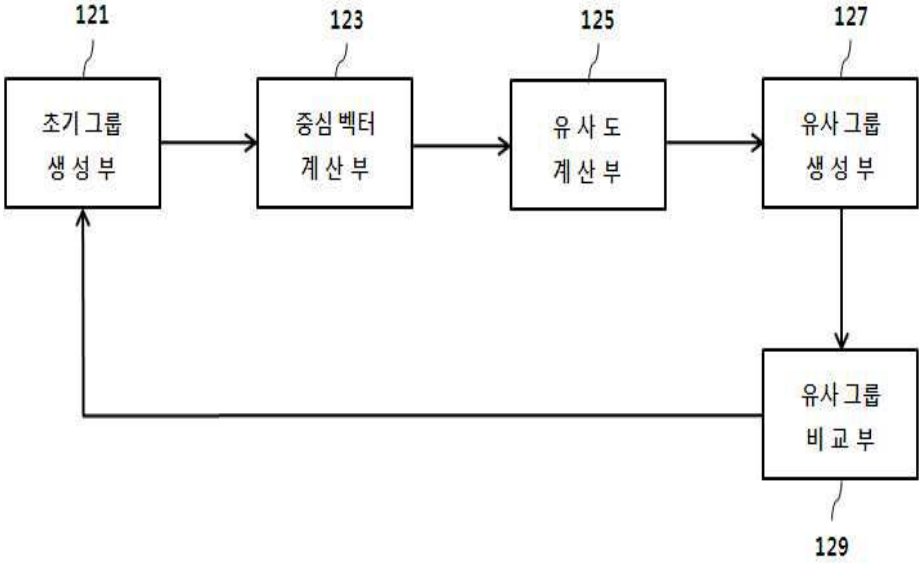
도면1



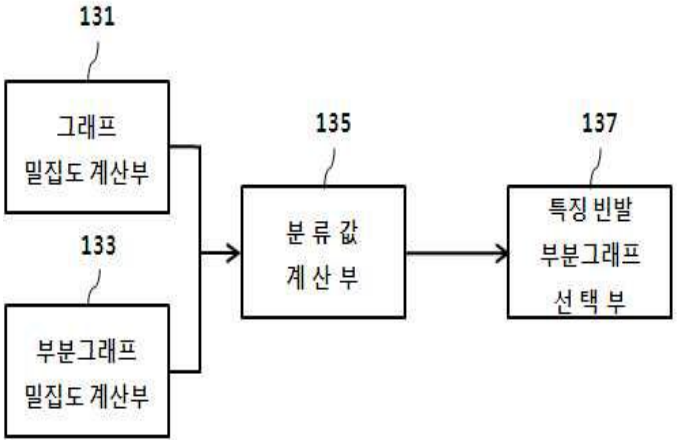
도면2


111
부분그래프
생성부
113
최소지지도
계산부
115
후보빈발
부분그래프
결정부

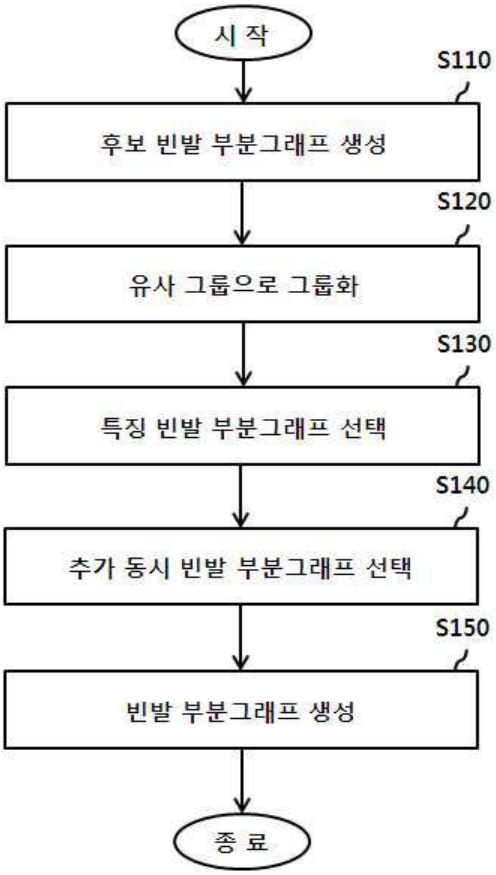
도면3


121
초기그룹
생성부
123
중심벡터
계산부
125
유사도
계산부
127
유사그룹
생성부
유사그룹
비교부
129

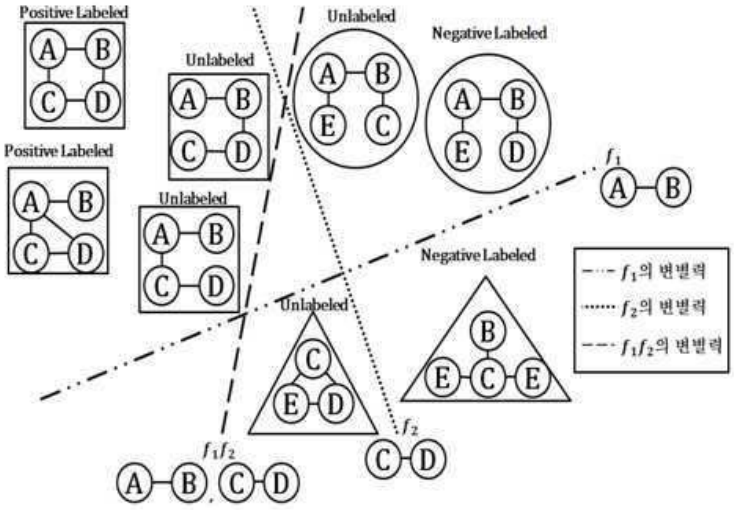
도면4


131
그래프
밀집도계산부
133
부분그래프
밀집도계산부
135
분류값
계산부
137
특징빈발
부분그래프
선택부

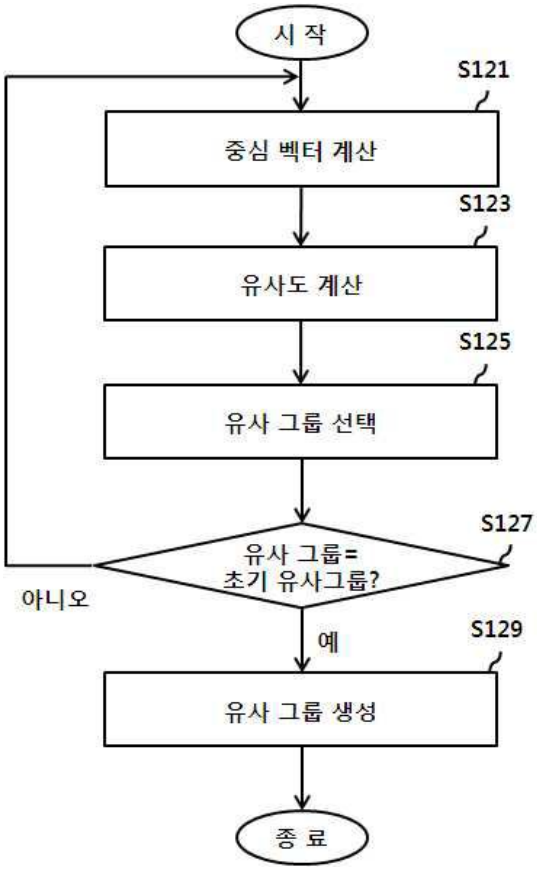
도면5


시 작
S110
후보 빈발 부분그래프 생성
S120
유사 그룹으로 그룹화
S130
특징 빈발 부분그래프 선택
S140
추가 동시 빈발 부분그래프 선택
S150
빈발 부분그래프 생성
종 료

도면6


Positive Labeled
Unlabeled
Unlabeled
Negative Labeled
Positive Labeled
Unlabeled
Negative Labeled
Unlabeled
f_1
f_2
f_1f_2
f_1의 변별력
f_2의 변별력
f_1f_2의 변별력

도면7


시 작
S121
중심 벡터 계산
S123
유사도 계산
S125
유사 그룹 선택
S127
유사 그룹=
초기 유사그룹?
아니오
예
S129
유사 그룹 생성
종 료

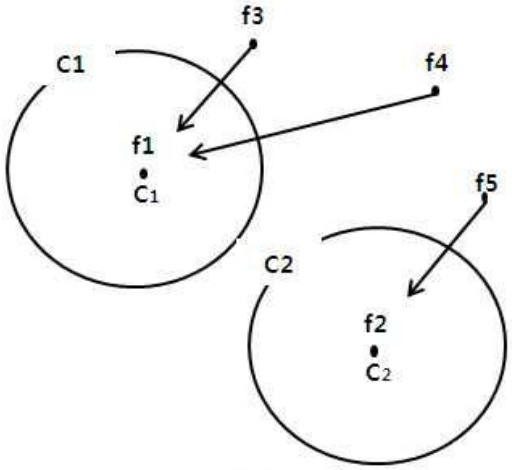
도면8

	f1	f2	f3	f4	f5
G1	1	0	1	0	0
G2	1	1	1	0	1
G3	0	1	1	0	1

(a)

	f1	f2	f3	f4	f5
C1	1	0.33	0.66	0.33	0.33
C2	0.33	1	0.66	0.33	1

(b)

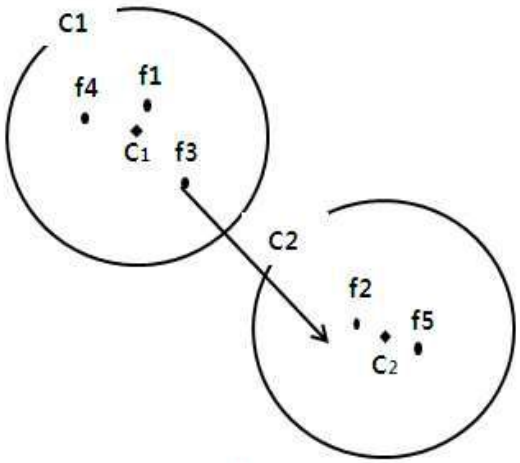


(c)

도면9

	f1	f2	f3	f4	f5
C1	0.55	0.44	0.55	0.44	0.44
C2	0.33	1	0.66	0.33	1

(a)

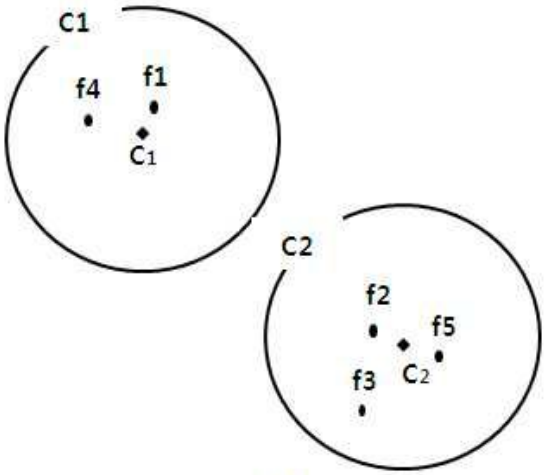


(b)

도면10

	f1	f2	f3	f4	f5
C1	0.55	0.44	0.55	0.44	0.44
C2	0.33	1	0.66	0.33	1

(a)



(b)