



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0064077
(43) 공개일자 2014년05월28일

(51) 국제특허분류(Int. Cl.)
G06F 17/40 (2006.01) G06F 17/30 (2006.01)
(21) 출원번호 10-2012-0130993
(22) 출원일자 2012년11월19일
심사청구일자 2012년11월19일

(71) 출원인
충북대학교 산학협력단
충청북도 청주시 흥덕구 내수동로 52 (개신동)
세종대학교산학협력단
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

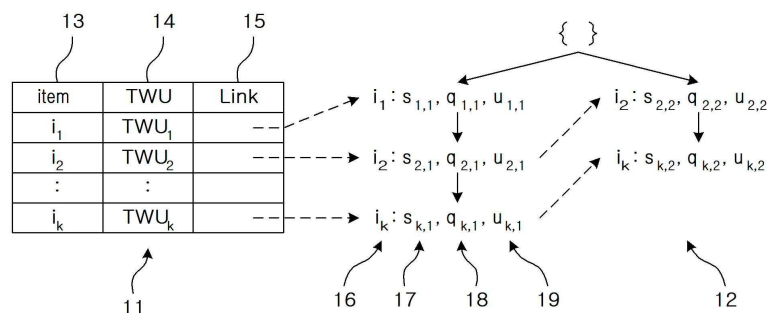
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 유용성 높은 패턴의 마이닝 방법

(57) 요약

본 발명은 유용성 높은 패턴의 마이닝 방법에 관한 것으로서, 데이터베이스로부터 관독된 트랜잭션의 유용성을 계산하여 트랜잭션 내 아이템을 정렬하고, 아이템 노드의 빈도수와 수량 설정에 따라 헤더 테이블을 갱신하여 전역 유용성 패턴 트리를 구축하는 단계와, 헤더 테이블을 트랜잭션 가중화 유용성(TWU)의 내림차순으로 재정렬하여 아이템 노드의 수량에 대응하여 감소된 유용성 추정치를 계산하고, 아이템 노드의 유용성을 결정하여 전역 유용성 패턴 트리를 재구축하는 단계와, 아이템에 대하여 패턴을 생성하고, 최소 임계치 이상의 트랜잭션 가중화 유용성을 아이템의 링크를 따라 전역 유용성 패턴 트리 내 모든 경로를 구축하고, 추출된 경로를 이용하여 지역 유용성 패턴 트리를 구축하는 단계를 포함하며, 기존의 방법들보다 더욱 빠른 속도로 높은 유용성을 가지는 패턴들을 마이닝할 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

데이터베이스로부터 판독된 트랜잭션의 유용성을 계산하여 상기 트랜잭션 내 아이템을 정렬하고, 아이템 노드의 빈도수와 수량 설정에 따라 헤더 테이블을 갱신하여 전역 유용성 패턴 트리를 구축하는 단계;

상기 헤더 테이블을 트랜잭션 가중화 유용성(TWU)의 내림차순으로 재정렬하여 상기 아이템 노드의 수량에 대응하여 감소된 유용성 추정치를 계산하고, 상기 아이템 노드의 유용성을 결정하여 상기 전역 유용성 패턴 트리를 재구축하는 단계; 및

상기 아이템에 대하여 패턴을 생성하고, 최소 임계치 이상의 트랜잭션 가중화 유용성을 상기 아이템의 링크를 따라 상기 전역 유용성 패턴 트리 내 모든 경로를 구축하고, 추출된 경로를 이용하여 지역 유용성 패턴 트리를 구축하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 2

제1항에 있어서,

상기 전역 유용성 패턴 트리는,

아이템 이름, 빈도수, 수량, 노드 링크를 포함하는 아이템 노드; 및

아이템 이름, 트랜잭션 가중화 유용성, 링크를 포함하는 전역 유용성 패턴 트리 내 헤더 테이블을 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 3

제1항에 있어서,

상기 전역 유용성 패턴 트리를 구축하는 단계는,

상기 데이터베이스에서 상기 트랜잭션을 판독하는 단계;

상기 트랜잭션의 유용성을 계산하는 단계;

상기 트랜잭션 내 아이템을 순서대로 전역 유용성 패턴 트리에 삽입하는 단계;

삽입하려는 상기 트랜잭션 내 아이템에 대한 상기 아이템 노드가 존재하지 않으면 새로운 아이템 노드를 생성하는 단계;

상기 아이템 노드의 빈도수와 수량을 설정하는 단계; 및

삽입된 상기 아이템에 대해 상기 헤더 테이블을 갱신하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 4

제3항에 있어서,

상기 아이템 노드의 빈도수와 수량을 설정하는 단계는,

빈도수 값을 1만큼 증가시키는 단계; 및

상기 아이템 노드의 수량과 삽입하려는 아이템 수량 중 큰 값으로 아이템 노드의 수량을 갱신하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 5

제1항에 있어서,
 상기 헤더 테이블의 갱신은,
 삽입된 아이টে에 대한 아이테 엔트리개 존재하는지 확인하는 단계;
 상기 삽입된 아이টে에 대한 아이테 엔트리가 없으면 새로 아이테 엔트리를 생성하는 단계; 및
 상기 삽입된 아이টে에 대한 아이테 엔트리를 갱신하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 6

제5항에 있어서,
 상기 삽입된 아이টে에 대한 아이테 엔트리를 갱신하는 단계는,
 상기 아이테 엔트리의 트랜잭션 가중화 유용성 값을 상기 삽입된 아이테이 포함된 트랜잭션의 유용성만큼 증가 시키는 단계; 및
 상기 아이테 엔트리의 링크를 새로 생성된 아이테 노드로 연결하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 7

제1항에 있어서,
 상기 트랜잭션 가중화 유용성의 내림차순으로의 재정렬은,
 상기 헤더 테이블을 상기 트랜잭션 가중화 유용성 내림차순으로 정렬하는 단계;
 상기 헤더 테이블의 상기 트랜잭션 가중화 유용성 내림차순 정렬에 따라 트리 내 경로를 재정렬하는 단계; 및
 재정렬된 상기 경로를 수량을 기반으로 감소된 유용성 추정치를 계산하여 재삽입하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 8

제1항에 있어서,
 재구축된 상기 전역 유용성 패턴 트리에서 패턴의 마이닝은,
 상기 헤더 테이블을 상향식으로 탐색하며 최소 임계치 이상의 실제 유용성을 가지는 길이가 1인 패턴을 생성하 는 단계;
 상기 최소 임계치 이상의 트랜잭션 가중화 유용성 값을 가지는 아이테의 링크를 따라 전역 유용성 패턴 트리 내 모든 경로를 추출하는 단계; 및
 추출된 경로를 사용하여 지역 유용성 패턴 트리를 구축하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 9

제1항에 있어서,
 상기 지역 유용성 패턴 트리는,
 아이테 이름, 빈도수, 노드 링크를 포함하는 아이테 노드; 및

아이템 이름, TWU, 빈도수, 링크를 포함하는 전역 유용성 패턴 트리 내 헤더 테이블을 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 10

제1항에 있어서,

상기 지역 유용성 패턴 트리를 구축하는 단계는,

상기 전역 유용성 패턴 트리에서 링크를 따라 추출된 경로에서 최소 임계치보다 작은 아이템들을 제거하는 단계; 및

상기 추출된 경로를 지역 유용성 패턴 트리 내 트랜잭션 가중화 유용성 내림차순에 따라 정렬하고 지역 유용성 패턴 트리에 삽입하여 지역 유용성 패턴 트리를 구축하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

청구항 11

제1항에 있어서,

상기 지역 유용성 패턴 트리에서 패턴의 마이닝은,

상기 지역 유용성 패턴 트리의 헤더 테이블을 상향식 탐색하며 Prefix와 헤더 테이블에서 선택된 아이템을 포함하는 패턴의 유용성 추정치를 계산하는 단계;

최소 임계치보다 작지 않은 유용성 추정치를 가지는 패턴을 생성하는 단계; 및

상기 Prefix와 선택된 아이템을 새로운 Prefix로 하는 새로운 지역 유용성 패턴 트리를 생성하는 단계를 포함하는 유용성 높은 패턴의 마이닝 방법.

명세서

기술분야

[0001] 본 발명은 데이터를 마이닝하는 방법에 관한 것으로, 더욱 상세하게는 빠른 속도로 높은 유용성을 가지는 패턴들을 마이닝하는 유용성 높은 패턴의 마이닝 방법에 관한 것이다.

배경기술

[0002] 데이터 마이닝은 거대한 데이터베이스 내의 숨겨진 유용한 지식을 발견하기 위해 사용되는 기술이다. 데이터 마이닝 기술을 통해 발견된 지식은 중요한 의사결정을 수행하는 데 활용된다. 데이터 마이닝은 크게 예측적 기법과 기술적 기법 두 가지로 나뉜다. 예측적 기법에는 분류(Classification), 회귀(Regression), 시계열 분석(Time Series Analysis), 예측(Prediction) 등이 있으며, 기술적 기법에는 군집(Clustering), 요약(Summarization), 연관 규칙(Association Rule) 등이 있다. 이 중 연관 규칙은 "A와 B 패턴이 얼마나 자주 발생하는가"를 의미하는 지지율과 "A라는 패턴이 발생했을 때 B라는 패턴이 얼마나 발생하는가"를 의미하는 신뢰도 두 가지 관점에서 중요한 연관 규칙을 찾아낸다.

[0003] 패턴 마이닝은 중요한 데이터 마이닝의 연구 분야 중에 하나로서, 지지율을 기반으로 중요한 연관 규칙, 즉 패턴을 찾아내는 기술이다. 빈발 패턴 마이닝은 데이터베이스에서 지지도 임계치(Threshold) 이상의 지지도를 가지는 모든 패턴들을 찾아낸다. 빈발 패턴 마이닝은 데이터베이스를 구성하는 모든 아이템들을 동일한 중요도로 고려하며, 데이터베이스의 각 트랜잭션 내 아이템들의 발생을 0 또는 1의 바이너리 값으로 표현한다. 그런데, 실세계 데이터베이스 내 아이템들을 실제로 서로 다른 중요도를 가지며, 빈발 패턴 마이닝은 이를 고려하지 못한다. 가중화 패턴 마이닝은 데이터베이스 내 아이템들의 서로 다른 중요도를 반영하여 중요한 패턴을 마이닝한다. 가중화 패턴 마이닝에서 중요한 아이템은 높은 가중치(Weight)를 가지며 상대적으로 중요하지 않은 아이템은 낮은 가중치를 갖는다. 또한, 가중화 패턴 마이닝은 빈발 패턴과 동일하게 데이터베이스의 각 트랜잭션 내 아이템들의 발생을 0 또는 1의 바이너리 값으로 표현한다. 그런데, 마켓 데이터베이스와 같은 실세계 데이터베이스

이스는 아이템들이 서로 다른 중요도를 가질 뿐만 아니라 하나의 트랜잭션 내에서 하나의 아이템이 비 바이너리 값으로서 발생한다. 유용성(Utility) 패턴 마이닝은 서로 다른 아이템들의 중요도와 트랜잭션 내 비 바이너리 발생 값을 고려한다.

[0004] 이와 같이, 거대한 데이터베이스에서 중요한 패턴을 찾아내기 위해 지지도를 기반으로 데이터베이스에서 지지도 임계치 이상의 지지도를 가지는 모든 패턴을 마이닝하는 빈발 패턴 마이닝이 개발된 바 있으나, 이 빈발 패턴 마이닝은 패턴 마이닝 분야에서 근본적인 역할을 수행할 뿐, 실세계 데이터베이스의 특성을 반영하여 패턴을 마이닝하지 못한다.

[0005] 또한, 실세계 데이터베이스 내 아이템들의 서로 다른 중요도를 고려하기 위해 가중화 패턴 마이닝이 개발되었으며, 지지도뿐만 아니라 패턴의 가중치, 즉 중요도를 고려하여 패턴을 마이닝한다. 하지만, 마켓 데이터베이스와 같은 실세계 데이터베이스 내 아이템들은 서로 다른 중요도를 가질 뿐만 아니라 한 번에 하나의 아이템이 여러 개 팔린 것과 같이 트랜잭션 내에서 비 바이너리 발생 값을 가지며, 가중화 패턴 마이닝은 이를 고려하지 못한다.

[0006] 최근 실세계 데이터베이스의 특성을 반영한 패턴 마이닝 기술이 요구되고 있으며, 이를 위해 유용성 패턴 마이닝 기술이 사용된다.

[0007] 그런데, 기존의 유용성 패턴 마이닝 기술들은 여러 번의 데이터베이스 스캔으로 트리를 구축하거나 과추정(Overestimated) 유용성 추정치를 안티 모노톤 속성(Anti-monotone property)을 만족시키기 위해 유용성 패턴 마이닝에 사용한다.

[0008] 즉, 마켓 데이터베이스와 같은 실세계 데이터베이스 내 아이템들은 서로 다른 가격 또는 이윤과 같은 중요도를 가지며, 각 트랜잭션 내 아이템들은 판매 수량과 같은 비 바이너리 값을 가진다. 따라서 기존의 빈발 패턴 마이닝 또는 가중화 패턴 마이닝과 같은 상기 두 가지 요소를 고려하지 않은 마이닝 기법에서는 이윤과 같은 유용성이 높은 패턴을 마이닝하지 못한다. 또한, 최근의 유용성 패턴 마이닝은 여러 번의 데이터베이스 스캔으로 트리를 구축하거나 과추정 유용성 추정치를 마이닝 과정에서 사용하여 높은 유용성을 가지는 패턴을 마이닝하는 데 많은 시간을 소비한다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 대한민국 공개특허공보 제10-2012-0049505호(공개일 2012.05.17.)

발명의 내용

해결하려는 과제

[0010] 따라서, 본 발명은 상기한 종래 기술의 문제점을 해결하기 위해 이루어진 것으로서, 본 발명의 목적은 한 번의 데이터베이스 스캔으로 아이템 수량 기반의 트리를 구축하고, 감소된 과추정 유용성 추정치를 사용하여 좀 더 빠르게 높은 유용성을 가지는 패턴을 마이닝하는 유용성 높은 패턴의 마이닝 방법을 제공하는 데 있다.

과제의 해결 수단

[0011] 상기와 같은 목적을 달성하기 위한 본 발명의 유용성 높은 패턴의 마이닝 방법은, 데이터베이스로부터 관독된 트랜잭션의 유용성을 계산하여 상기 트랜잭션 내 아이템을 정렬하고, 아이템 노드의 빈도수와 수량 설정에 따라 헤더 테이블을 갱신하여 전역 유용성 패턴 트리를 구축하는 단계; 상기 헤더 테이블을 트랜잭션 가중화 유용성(TWU)의 내림차순으로 재정렬하여 상기 아이템 노드의 수량에 대응하여 감소된 유용성 추정치를 계산하고, 상기 아이템 노드의 유용성을 결정하여 상기 전역 유용성 패턴 트리를 재구축하는 단계; 및 상기 아이템에 대하여 패턴을 생성하고, 최소 임계치 이상의 트랜잭션 가중화 유용성을 상기 아이템의 링크를 따라 상기 전역 유용성 패턴 트리 내 모든 경로를 구축하고, 추출된 경로를 이용하여 지역 유용성 패턴 트리를 구축하는 단계를 포함하는

것을 특징으로 한다.

- [0012] 상기 전역 유용성 패턴 트리는, 아이템 이름, 빈도수, 수량, 노드 링크를 포함하는 아이템 노드; 및 아이템 이름, 트랜잭션 가중화 유용성, 링크를 포함하는 전역 유용성 패턴 트리 내 헤더 테이블을 포함한다.
- [0013] 상기 전역 유용성 패턴 트리를 구축하는 단계는, 상기 데이터베이스에서 상기 트랜잭션을 판독하는 단계; 상기 트랜잭션의 유용성을 계산하는 단계; 상기 트랜잭션 내 아이템을 순서대로 전역 유용성 패턴 트리에 삽입하는 단계; 삽입하려는 상기 트랜잭션 내 아이템에 대한 상기 아이템 노드가 존재하지 않으면 새로운 아이템 노드를 생성하는 단계; 상기 아이템 노드의 빈도수와 수량을 설정하는 단계; 및 삽입된 상기 아이템에 대해 상기 헤더 테이블을 갱신하는 단계를 포함한다. 이때, 상기 아이템 노드의 빈도수와 수량을 설정하는 단계는, 빈도수 값을 1만큼 증가시키는 단계; 및 상기 아이템 노드의 수량과 삽입하려는 아이템 수량 중 큰 값으로 아이템 노드의 수량을 갱신하는 단계를 포함한다.
- [0014] 상기 헤더 테이블의 갱신은, 삽입된 아이템에 대한 아이템 엔트리가 존재하는지 확인하는 단계; 상기 삽입된 아이템에 대한 아이템 엔트리가 없으면 새로 아이템 엔트리를 생성하는 단계; 및 상기 삽입된 아이템에 대한 아이템 엔트리를 갱신하는 단계를 포함한다. 이때, 상기 삽입된 아이템에 대한 아이템 엔트리를 갱신하는 단계는, 상기 아이템 엔트리의 트랜잭션 가중화 유용성 값을 상기 삽입된 아이템이 포함된 트랜잭션의 유용성만큼 증가시키는 단계; 및 상기 아이템 엔트리의 링크를 새로 생성된 아이템 노드로 연결하는 단계를 포함한다.
- [0015] 상기 트랜잭션 가중화 유용성의 내림차순으로의 재정렬은, 상기 헤더 테이블을 상기 트랜잭션 가중화 유용성 내림차순으로 정렬하는 단계; 상기 헤더 테이블의 상기 트랜잭션 가중화 유용성 내림차순 정렬에 따라 트리 내 경로를 재정렬하는 단계; 및 재정렬된 상기 경로를 수량을 기반으로 감소된 유용성 추정치를 계산하여 재삽입하는 단계를 포함한다.
- [0016] 재구축된 상기 전역 유용성 패턴 트리에서 패턴의 마이닝은, 상기 헤더 테이블을 상향식으로 탐색하며 최소 임계치 이상의 실제 유용성을 가지는 길이가 1인 패턴을 생성하는 단계; 상기 최소 임계치 이상의 트랜잭션 가중화 유용성 값을 가지는 아이템의 링크를 따라 전역 유용성 패턴 트리 내 모든 경로를 추출하는 단계; 및 추출된 경로를 사용하여 지역 유용성 패턴 트리를 구축하는 단계를 포함한다.
- [0017] 상기 지역 유용성 패턴 트리는, 아이템 이름, 빈도수, 노드 링크를 포함하는 아이템 노드; 및 아이템 이름, TWU, 빈도수, 링크를 포함하는 전역 유용성 패턴 트리 내 헤더 테이블을 포함한다.
- [0018] 상기 지역 유용성 패턴 트리를 구축하는 단계는, 상기 전역 유용성 패턴 트리에서 링크를 따라 추출된 경로에서 최소 임계치보다 작은 아이템들을 제거하는 단계; 및 상기 추출된 경로를 지역 유용성 패턴 트리 내 트랜잭션 가중화 유용성 내림차순에 따라 정렬하고 지역 유용성 패턴 트리에 삽입하여 지역 유용성 패턴 트리를 구축하는 단계를 포함한다.
- [0019] 상기 지역 유용성 패턴 트리에서 패턴의 마이닝은, 상기 지역 유용성 패턴 트리의 헤더 테이블을 상향식 탐색하며 Prefix와 헤더 테이블에서 선택된 아이템을 포함하는 패턴의 유용성 추정치를 계산하는 단계; 최소 임계치보다 작지 않은 유용성 추정치를 가지는 패턴을 생성하는 단계; 및 상기 Prefix와 선택된 아이템을 새로운 Prefix로 하는 새로운 지역 유용성 패턴 트리를 생성하는 단계를 포함한다.

발명의 효과

- [0020] 상술한 바와 같이, 본 발명에 의한 유용성 높은 패턴의 마이닝 방법에 따르면, 트랜잭션 내 아이템들의 빈도수 수량 값을 이용하여 한 번의 데이터베이스 스캔으로 트리를 구축하고, 데이터베이스 내 아이템 유용성과 트리에 저장된 각 아이템의 수량 정보를 이용하여 유용성 값을 계산한 뒤, 이를 기반으로 사용자 정의 최소 유용성 임계치보다 큰 유용성을 가지는 모든 패턴들을 마이닝하고, 감소된 유용성 추정치를 사용하여 더욱 적은 수의 패턴을 생성하므로, 기존의 방법들보다 더욱 빠른 속도로 높은 유용성을 가지는 패턴들을 마이닝할 수 있다.
- [0021] 또한, 본 발명에 따르면, 실세계 데이터베이스 내 아이템들의 서로 다른 유용성 및 수량을 반영하여 높은 유용성을 가지는 패턴, 예를 들면 높은 판매 수익을 내는 상품 패턴을 마이닝할 수 있다. 즉, 데이터베이스 내 모든 아이템들을 동일한 유용성을 가지고 트랜잭션 내 아이템들의 발생을 빈도수 값으로 표현한 기존의 일반 빈발 패턴 마이닝과 오직 아이템들의 서로 다른 중요도만을 고려한 가중화 빈발 패턴 마이닝에서 마켓 데이터베이스

와 같은 실세계 데이터베이스의 특성을 고려하지 못한 문제를 해결할 수 있으며, 이를 통해 마켓이나 주식시장 분석 같은 실세계 데이터베이스 분석에 기여할 수 있다.

도면의 간단한 설명

[0022]

도 1은 본 발명의 일실시예에 의한 수량 기반의 전역 유용성 패턴 트리의 구조도이다.

도 2는 본 발명의 일실시예에 의한 유용성 패턴 마이닝에서 사용되는 데이터베이스의 구성도이다.

도 3은 본 발명의 일실시예에 의한 전역 유용성 패턴 트리를 구축하는 방법의 흐름도이다.

도 4는 본 발명의 일실시예에 의한 삽입된 트랜잭션 내 아이템 정보를 헤더 테이블에 반영하는 방법의 흐름도이다.

도 5는 본 발명의 일실시예에 의한 초기 전역 트리를 재구축하고 모든 유용성 패턴을 찾아내는 방법의 흐름도이다.

도 6은 본 발명의 일실시예에 의한 최소, 최대, 실제 유용성 테이블의 구조도이다.

도 7은 본 발명의 일실시예에 의한 지역 유용성 패턴 트리의 구조도이다.

도 8은 본 발명의 일실시예에 의한 지역 유용성 패턴 트리에서 패턴을 생성하는 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0023]

이하, 본 발명의 유용성 높은 패턴의 마이닝 방법에 대하여 첨부된 도면을 참조하여 상세히 설명하기로 한다.

[0024]

본 발명에서는 아이템 중요도와 트랜잭션 내 아이템 수량 정보를 포함하는 데이터베이스를 스캔하여 아이템 수량 정보를 기반으로 한 번의 데이터베이스 스캔으로 트리를 구축하고, 구축된 트리와 아이템 중요도를 이용하여 감소된 과추정 유용성 추정치를 계산하며, 이를 기반으로 높은 유용성을 가지는 패턴을 마이닝한다.

[0025]

도 1은 데이터베이스의 트랜잭션 내 아이템 수량 정보를 기반으로 한 번의 데이터베이스 스캔으로 구축된 유용성 트리의 구조이며, 도 2는 유용성 트리를 구축하는 데 사용되는 데이터베이스의 예이다. 본 발명의 수량 기반의 전역 유용성 패턴 트리(이하, "유용성 트리"라 함)(12)는 비 바이너리 값의 아이템 수량 정보를 사용하여 구축된 트리 구조로서, 구축된 유용성 트리(12)와 헤더 테이블(11)로 구성되며, 데이터베이스 스캔을 여러 번 반복하지 않고 판독된 트랜잭션을 저장 구조에 저장하여 성능을 향상시키기 위해 사용된다. 유용성 트리(12) 내의 헤더 테이블(11)은 트리에 삽입된 아이템들에 대한 엔트리로 구성되며, 각 엔트리는 아이템 이름(13), 트랜잭션 가중화 유용성(TWU ; Transaction Weighted Utilization)(14), 트리 내 생성된 동일한 아이템 이름의 트리 노드를 가리키기 위한 링크(15)를 포함한다. 또한, 유용성 트리(12)는 루트 노드와 루트 노드를 제외한 일반 아이템 노드로 구성되고 각 아이템 노드는 아이템 이름(16), 빈도수(17), 수량(18), 유용성(19), 노드 링크로 구성되며, 데이터베이스에서 판독된 트랜잭션 내 아이템들이 순서대로 루트 노드부터 시작하여 트리에 삽입되어 유용성 트리(12)가 구축된다.

[0026]

도 3은 데이터베이스를 스캔하여 상기 유용성 트리(12)를 한 번의 데이터베이스 스캔으로 구축하는 방법이다. 트리를 구축하는 방법은 데이터베이스에서 하나의 트랜잭션을 판독하는 과정(S31), 판독된 트랜잭션의 유용성을 계산하는 과정(S32), 트랜잭션 내 각 아이템을 순서대로 유용성 트리(12)에 삽입하는 과정(S33), 삽입하려는 트랜잭션 내 각 아이템에 대한 노드가 존재하지 않으면 새로운 노드를 생성하는 과정(S34), 아이템 노드의 빈도수와 수량을 설정하는 과정(S35), 헤더 테이블에 삽입된 아이템 정보를 반영하는 과정(S36)으로 이루어진다. 데이터베이스는 트랜잭션 집합이며, 트리를 구축하기 위해 데이터베이스에서 하나씩 트랜잭션을 판독한다(S31). 트랜잭션은 아이템들과 각 아이템들과 연관을 가지는 비 바이너리 값의 수량 정보로 구성되며, 데이터베이스 내 아이템들은 도 2의 데이터베이스 예처럼 서로 다른 유용성을 가진다. 트랜잭션 내 아이템의 유용성은 아이템 수량과 아이템 유용성의 곱으로 나타내며, 트랜잭션의 유용성은 트랜잭션 내 아이템들의 유용성 합이다. 예를 들어, 도 2에서, 아이템 i_1 의 트랜잭션 T_1 에서의 유용성은 $10 \times q_{1,1}$ 이다. 트랜잭션을 데이터베이스에서 판독하면 먼저 트랜잭션 유용성을 계산하고(S32), 트랜잭션 내 각 아이템을 순차적으로 트리에 삽입한다(S33). 삽입하는

아이템과 동일한 이름을 가지는 아이템 노드가 존재하지 않으면 해당 아이템 이름을 가지는 아이템 노드를 새로 생성하며(S34), 이때 새로 생성된 노드 아이템의 값을 모두 0으로 초기화한다. 새로 생성되거나 기존의 동일한 이름을 가지는 아이템 노드의 빈도수를 1 증가시키고, 아이템 수량과 아이템 노드의 수량 값 중에서 큰 값을 아이템 노드의 수량 값으로 갱신하여 새로운 빈도수와 수량을 설정한다(S35). 현재 트랜잭션 내 아이템에 대한 삽입이 완료되면 해당 아이템 정보를 헤더 테이블에 반영한다(S36). 초기 전역 트리 내의 각 노드에 저장된 수량(18) 정보는 재구축 과정에서 감소된 유용성 추정치 계산에 사용된다.

[0027] 도 4는 삽입된 트랜잭션 내 아이템 정보를 헤더 테이블에 반영하는 방법이다. 헤더 테이블을 갱신하는 방법은 아이템 노드로서 유용성 트리(12)에 삽입된 아이템에 대한 엔트리가 헤더 테이블에 존재하는지 확인하는 과정(S41), 설정하려는 아이템 엔트리가 존재하지 않으면 새로운 엔트리를 생성하는 과정(S42), 삽입된 아이템에 대한 엔트리 값을 갱신하는 과정(S43)으로 이루어진다. 먼저 유용성 트리(12)에 삽입된 트랜잭션 내 아이템에 대한 트랜잭션이 헤더 테이블(11)의 엔트리로서 존재하는지 확인한다(S41). 만약 해당 아이템에 대한 엔트리가 존재하지 않으면 새로운 아이템 엔트리를 생성한다(S42). 이때 생성된 아이템 엔트리의 아이템 이름(13)은 해당 아이템 이름으로 설정되고 TWU 값은 0으로 초기화된다. 다음, 아이템 엔트리의 TWU 값을 해당 아이템이 포함된 트랜잭션 유용성만큼 증가시키고, 만약 해당 아이템 노드가 새로 생성된 노드이면 링크를 새로 생성된 노드에 연결하고, 새로 생성된 노드를 기존의 링크로 연결되어 있던 노드와 연결하여 삽입된 아이템 정보를 엔트리 값을 갱신한다(S43).

[0028] 데이터베이스 내 모든 트랜잭션이 삽입되면 유용성 트리(12) 구축이 완료된다. 한 번의 데이터베이스 스캔으로 초기 트리가 구축되면, 구축된 트리를 재구축하고, 이를 통해 모든 유용성이 높은 패턴을 찾아낸다. 도 5는 구축된 트리를 재구축하고, 재구축된 트리에서 모든 유용성 높은 패턴을 찾아내는 방법이다. 유용성 트리(12)의 재구축 및 모든 높은 유용성 패턴을 마이닝하는 방법은 헤더 테이블과 트리 내 아이템 노드들을 TWU 내림차순으로 정렬하는 과정(S51), 아이템 노드에 저장된 수량 정보를 기반으로 감소된 유용성 추정치를 계산하여 구축된 초기 트리를 재구축하는 과정(S52), 헤더 테이블을 상향식으로 탐색하며 사용자 정의 최소 유용성 임계치 이상의 실제 유용성을 가지는 패턴을 생성하는 과정(S53), 최소 임계치 이상의 TWU 값을 가지는 아이템 엔트리의 링크를 따라 트리에서 해당 아이템의 모든 경로를 추출하여 지역 트리를 구축하는 과정(S54), 구축된 지역 트리에서 최소 임계치 이상의 유용성 추정치를 가지는 패턴을 생성하는 과정(S55), 생성된 패턴의 실제 유용성을 계산하여 실제 유용성 높은 패턴을 식별하는 과정(S56)으로 이루어진다. 모든 유용성이 높은 패턴을 마이닝하기 위해서 먼저, 헤더 테이블 내 아이템 엔트리를 TWU 내림차순으로 정렬한다(S51). 상기 헤더 테이블의 TWU 내림차순 정렬 순서에 따라 구축된 유용성 트리(12) 내 모든 아이템 노드들도 다시 정렬하고, 수량 정보를 사용하여 감소된 유용성 추정치를 계산하여 노드의 유용성 값을 갱신하며, 이를 통해 구축된 초기 트리를 재구축한다(S52). 이때 감소된 유용성 추정치 계산은, 만약 초기 트리에서 추출되고 TWU 내림차순으로 정렬된 경로가 $\{(i_1, q_1), (i_2, q_2), \dots, (i_k, q_k)\}$ 이면, 첫 번째 노드의 감소된 유용성 추정치는 i_1 의 유용성과 q_1 의 곱이다. 다음 노드의 감소된 유용성 추정치는 i_2 의 유용성과 q_2 의 곱에 첫 번째 노드의 유용성 추정치를 합한 값이고, 세 번째 노드의 감소된 유용성 추정치 역시 i_3 의 유용성과 q_3 의 곱에 두 번째 노드의 유용성 추정치를 합한 값이며, 나머지 노드들도 동일한 방법을 사용하여 유용성 추정치를 계산한다. 이는 유용성 트리를 사용한 마이닝은 상향식으로 진행되며, 헤더 테이블 내 상위 아이템을 기반으로 한 패턴 마이닝에서, 즉 상위 아이템에 대한 지역 트리에 대한 마이닝에서 하위 아이템이 포함된 패턴이 생성되지 않기 때문에 하위 아이템들의 유용성 정보를 포함시키지 않고 과추정 유용성 추정치를 감소시킬 수 있기 때문이다. 상기 초기 유용성 트리(12)에 대한 재구축을 통해 한 번의 데이터베이스 스캔만으로 전역 트리를 구축할 수 있으며, 이를 통해 추가적인 데이터베이스 스캔으로 인한 시간 소모를 감소시킬 수 있다. 또한, 감소된 유용성 추정치의 사용으로 생성되는 패턴의 수를 감소시키며, 이를 통해 생성된 패턴들의 실제 유용성을 확인하여 실제 유용성 높은 패턴을 식별해내는 과정에 소비되는 시간을 감소시킨다. TWU 내림차순에 따른 정렬 및 재구축이 끝나면 헤더 테이블을 맨 아래의 아이템부터 시작하여 상향식으로 탐색한다. 이때 선택된 아이템 엔트리의 TWU 값이 최소 임계치보다 크거나 같고, 또한 해당 아이템의 데이터베이스 내 실제 유용성도 TWU 값 이상이면 해당 아이템을 하나의 길이가 1인 패턴으로 생성한다(S53). 상기 아이템의 실제 유용성은 데이터베이스 내의 해당 아이템을 포함하는 각 트랜잭션 내에서 해당 아이템의 유용성의 합으로서 초기 트리를 구축하는 과정에서 계산할 수 있으며, 이때 데이터베이스 내 각 아이템의 최소(61), 최대(62), 실제(63) 유용성 테이블의 예이다. 예를 들어, 아이템 A가 트랜잭션 1 그리고 3에 포함되어 있고, 각 트랜잭션 내 아이템 A의 유용성이 10과 20라고 하면, 아이템 A의 TWU(14)가 트랜잭션 1과

3 각각의 트랜잭션 유용성의 합인 것과는 달리 실제 유용성(63)은 각 아이템 A의 유용성 합인 30가 된다. 또한, 아이템 A의 최대(62) 그리고 최소(61) 유용성은 각각 20과 10이 된다. 이와 같이 길이가 1인 패턴을 실제 유용성을 사용하여 프루닝 함으로써 좀 더 적은 수의 패턴이 생성되도록 하며 이를 통해 생성된 패턴의 실제 유용성 계산에 소비되는 시간을 감소시킨다. 또한, 초기 전역 유용성 트리(12)에서 선택된 아이템 엔트리가 아이템 A에 대한 것이면 해당 아이템 엔트리의 링크를 따라 구축된 지역 트리에서 생성된 모든 패턴들은 아이템 A를 포함한다. 예를 들어, 아이템 A의 TWU 값이 100, 최소 유용성 임계치가 50, 그리고 실제 유용성이 80이면, 실제 유용성이 최소 임계치보다 크므로 먼저 $\{A : 80\}$ 이라는 패턴이 생성되고 아이템 A를 기반으로 새로운 지역 트리가 생성된다. 만약 상기 지역 트리 내에 TWU 값이 50보다 큰 아이템 B가 존재하면 $\{A B\}$ 와 같이 A를 Prefix로 하는 패턴이 생성된다. 다음, TWU 값이 최소 임계치보다 크면 해당 아이템 엔트리의 링크를 따라 구축된 트리에서 해당 엔트리 아이템을 시작으로 하는 모든 경로를 추출하여 새로운 지역 유용성 트리를 구축한다(S54). 이때, 최소 유용성 임계치보다 작은 아이템들은 제거되며, 지역 트리 내의 TWU 내림차순으로 헤더 테이블과 트리가 생성된다. 또한, 추출된 경로가 TWU 내림차순으로 정렬되어 지역 트리에 삽입될 때, 맨 마지막 아이템 노드의 유용성은 해당 경로의 유용성과 같고, 해당 노드의 부모 노드의 유용성은 맨 마지막 아이템의 최소 유용성을 해당 경로의 유용성에서 뺀 것과 같다. 즉, 추출되어 TWU 내림차순으로 정렬된 경로가 $\{i_1, i_2, \dots, i_m\}$ 이고, 해당 경로의 유용성이 10이라고 하면, i_m 의 유용성은 10이고, i_{m-1} 의 유용성은 10에서 i_m 의 최소 유용성(61)을 뺀 것과 같으며, 나머지 아이템들도 하위 아이템들의 유용성을 해당 경로의 유용성에서 뺀 것과 같다. 이 역시, 전역 트리와 마찬가지로 상위 아이템들을 기준으로 패턴을 생성할 때 하위 아이템들의 유용성을 포함되지 않기 때문이다. 즉, 이렇게 트리 내 노드들의 유용성을 감소시킴으로써 생성되는 패턴의 수를 더욱 감소시키며, 이를 통해 마지막 단계에서 실제 높은 유용성을 가지는 패턴을 식별할 때 더욱 적은 시간을 소비한다. 도 7은 전역 유용성 트리로부터 구축된 지역 트리의 구조이다. 전역 트리와는 달리 지역 트리의 헤더 테이블(71)은 아이템 이름, TWU, 빈도수(73), 링크로 구성된다. 또한, 지역 유용성 패턴 트리(이하, "지역 트리"라 함)(72)의 아이템 노드는 아이템 이름, 빈도수, 유용성, 노드 링크로 구성된다. 즉, 지역 유용성 트리의 헤더 테이블(71)은 빈도수(73)를 추가로 가지며, 지역 트리(72) 내 아이템 노드들은 수량(18) 정보를 포함하지 않는다. 지역 트리의 헤더 테이블 내 아이템의 빈도수는 해당 지역 트리의 조건적 아이템, 즉 지역 트리의 Prefix와 함께 해당 아이템이 데이터베이스 내에서 몇 개의 트랜잭션에서 발생했는지를 의미하며, 지역 트리에서 패턴을 생성하는 과정은 전역 트리와 유사한 단계를 거친다(S55).

[0029]

도 8은 지역 트리에서 패턴을 생성하는 방법이다. 지역 트리에서 패턴을 생성하는 방법은 지역 트리의 헤더 테이블을 상향식으로 탐색하며 상기 Prefix와 헤더 테이블에서 선택된 아이템을 포함하는 패턴의 유용성 추정치 계산하는 과정(S81), 해당 패턴을 새로운 Prefix로 하는 새 지역 트리를 생성하는 과정(S82)으로 이루어진다. 전역 트리와 마찬가지로 지역 트리에서 패턴을 생성하기 위해서 먼저 지역 트리의 헤더 테이블(71)의 맨 마지막 아이템부터 상향식으로 탐색한다. 선택된 아이템의 빈도수는 지역 트리의 Prefix와 해당 아이템이 데이터베이스 내에서 몇 개의 트랜잭션에서 발생했는지를 의미한다. 따라서 Prefix 내의 전역 트리에서 선택된 아이템을 i_g 라고 할 때, 상기 패턴 내 i_g 의 유용성은, i_g 의 실제 유용성에서 (i_g 의 최소 유용성 X (데이터베이스 내 i_g 의 빈도수 - 선택된 아이템의 빈도수))를 뺀 값보다 작거나 같다. 예를 들어 i_g 의 실제 유용성(63)이 10, 빈도수가 3, 최소 유용성(61)이 2, 그리고 헤더 테이블에서 선택된 아이템 i_1 의 빈도수(73)가 1이면, Prefix 내 i_g 의 유용성은 $10 - (2 \times (3 - 1)) = 6$ 이다. 즉, i_g 는 데이터베이스 내의 트랜잭션에서 최소 2 이상의 유용성을 가지고, Prefix는 총 1개의 트랜잭션에서 i_1 과 같이 등장한다. 따라서 $\{i_g, i_1\}$ 패턴에서 i_g 의 유용성은 6보다 작거나 같다. 또한, i_g 를 제외한 Prefix 내 아이템들은 i_1 과 함께 총 1개의 트랜잭션에서 등장하며, 각각의 아이템은 해당 트랜잭션 안에서 도 6의 해당 아이템의 최대 유용성(62)보다 작은 유용성을 가지므로 $\{i_g, i_1\}$ 패턴의 유용성은 최대 상기 추정된 i_g 의 유용성과 i_1 의 빈도수(73)와 i_1 의 최대 유용성(62)의 곱의 합과 같다. 예를 들어, i_1 의 최대 유용성이 4라고 하면, 상기 패턴 내에서 i_1 은 최대 4×1 의 유용성을 가지며, 따라서 패턴 $\{i_g, i_1\}$ 의 유용성은 최대 $6 + 4 = 10$ 이다. 이렇게 계산한 패턴의 유용성 추정치가 최소 유용성 임계치보다 크거나 같으면 해당 패턴을 생성한다(S81). 다음, 지역 트리의 헤더 테이블(71)에서 선택된 아이템을 Prefix에 추가한 것을 새로운 Prefix로 하는 지역 트리를 생성하고 동일한 과정을 반복한다(S82). 지역 트리 구축에서는 최소 유용성 임계치보다 작은 아이템들은 제거되므로 헤더 테이블(71)에서 선택된 아이템의 TWU 값을 최소 임계치와 비교하지 않고 새로운 지역 트리를 생성한다(S82). 마지막으로 생성된 패턴들의 실제 유용성을 계산하여 실제 유용성이 높은

패턴들을 식별한다(S56).

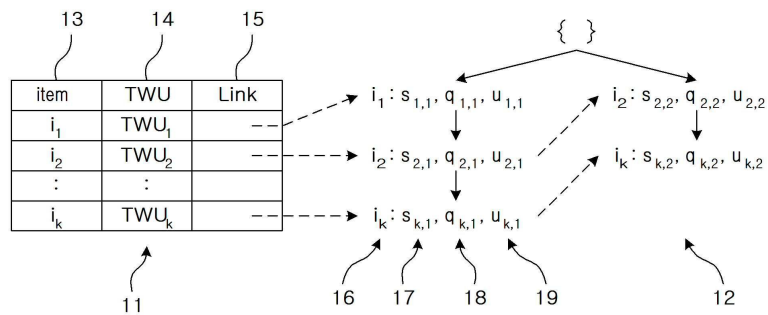
[0030] 이상에서 몇 가지 실시예를 들어 본 발명을 더욱 상세하게 설명하였으나, 본 발명은 반드시 이러한 실시예로 국한되는 것이 아니고 본 발명의 기술사상을 벗어나지 않는 범위 내에서 다양하게 변형실시될 수 있다.

부호의 설명

[0031] 11 : 헤더 테이블
12 : 전역 유용성 패턴 트리
71 : 헤더 테이블
72 : 지역 유용성 패턴 트리

도면

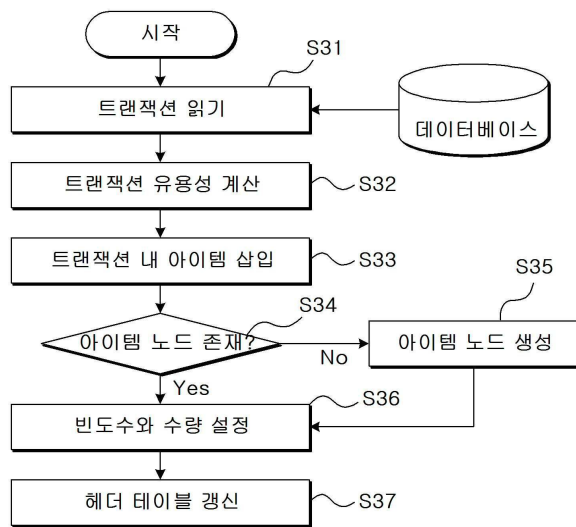
도면1



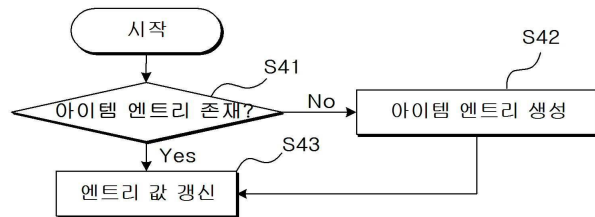
도면2

TID	Transaction	item	Utility
T ₁	(i ₁ , q _{1,1}), (i ₂ , q _{2,1})	i ₁	10
T ₂	(i ₃ , q _{3,1}), (i ₄ , q _{4,1})	i ₂	4
:	:	:	:
T _n	(i ₁ , q _{1,2}), (i ₃ , q _{3,2})	i _k	5

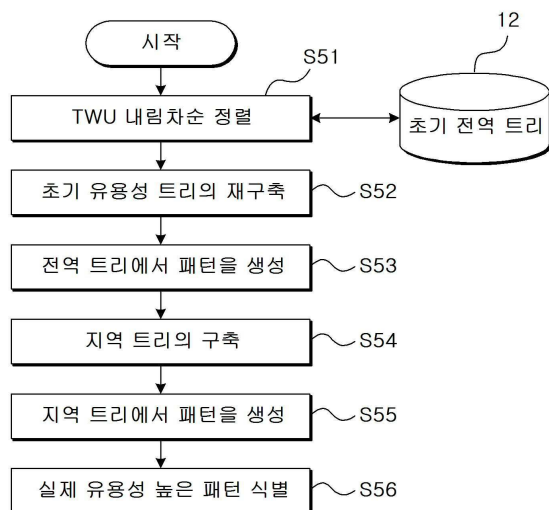
도면3



도면4



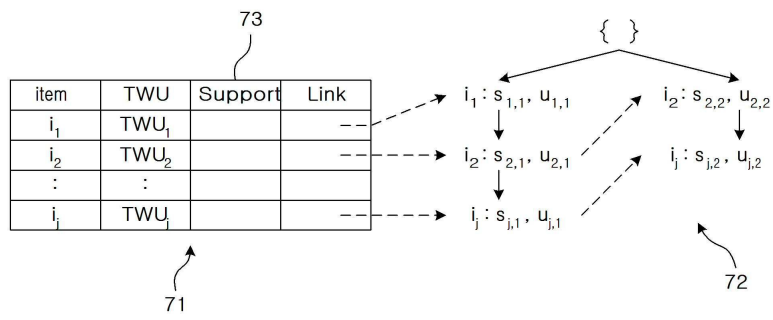
도면5



도면6

item	61 Min,Utility	62 Max,Utility	63 Real Utility
i_1	10	20	30
i_2	8	8	16
:	:	:	:
i_k	5	15	40

도면7



도면8

