



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년06월26일
 (11) 등록번호 10-1751062
 (24) 등록일자 2017년06월20일

(51) 국제특허분류(Int. Cl.)
 G06F 17/18 (2006.01)
 (21) 출원번호 10-2010-0129575
 (22) 출원일자 2010년12월17일
 심사청구일자 2015년06월23일
 (65) 공개번호 10-2011-0073287
 (43) 공개일자 2011년06월29일
 (30) 우선권주장
 09180322.1 2009년12월22일
 유럽특허청(EPO)(EP)
 (56) 선행기술조사문헌
 US20030212855 A1
 US20080077544 A1
 KR1020040101425 A
 KR1020040029157 A

(73) 특허권자
 인터내셔널 비지네스 머신즈 코포레이션
 미국 10504 뉴욕주 아몬크 뉴오차드 로드
 (72) 발명자
 버스트, 미카엘
 독일 보에블링겐 디-71032, 쇠나이처 스트리트
 220, IBM 독일 리서치 앤드 디벨롭먼트 지엠비에
 이치
 시에브, 크리스토프
 독일 보에블링겐 디-71032, 쇠나이처 스트리트
 220, IBM 독일 리서치 앤드 디벨롭먼트 지엠비에
 이치
 (뒷면에 계속)
 (74) 대리인
 허정훈

전체 청구항 수 : 총 9 항

심사관 : 경연정

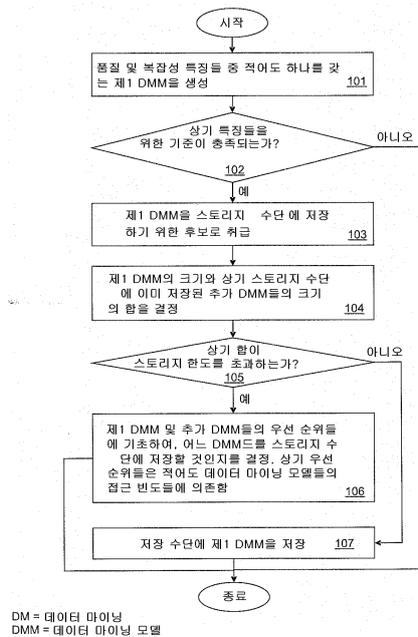
(54) 발명의 명칭 **인터랙티브 데이터 마이닝을 가능하게 하기 위한 마이닝 모델들의 선택적 저장**

(57) 요약

데이터 마이닝 모델들(data mining models, DMMs)의 저장을 위한 컴퓨터화된 방법들, 데이터 처리 시스템들, 및 컴퓨터 프로그램 제품들이 제공된다. 이 DMM은 다음의 특징들, 즉, 품질 및 복잡성 중 적어도 하나를 갖고서 생성된다. 만약 상기 특징들에 대한 미리 정의된 기준이 충족되면, 상기 DMM은 스토리지 수단에 저장하기 위한 후

(뒷면에 계속)

대표도 - 도1



보로서 핸들링된다. 상기 DMM 및 미리 저장된 DMM들의 크기들의 합이 결정된다. 만약 이 합이 스토리지 한도 아래로 떨어지면, 상기 DMM은 상기 스토리지 수단에 저장된다. 만약 이 합이 상기 스토리지 한도를 초과하면, 상기 DMM들의 우선 순위들에 기초하여 상기 스토리지 수단에 어느 DMM들이 저장될지가 결정된다. 상기 우선 순위들은 상기 DMM들의 접근 빈도들에 적어도 의존한다. 데이터 마이닝 요청시, 대응하는 DMM이 결정되고, 사용자는 만약 상기 결정된 DMM의 품질이 추가의 미리 결정된 기준을 충족시키지 않으면, 데이터 마이닝이 진행되는 것을 확인하도록 사용자는 요청받는다.

(72) 발명자

미상, 버나드

독일 보에블링겐 디-71032, 분센 스트리트 83

폴리도 데 로스 레예스, 루벤

독일 보에블링겐 디-71032, 쉐나이처 스트리트
220, IBM 독일 리서치 앤드 디벨롭먼트 지엠비에이
치

랑, 알렉산더

독일 보에블링겐 디-71032, 쉐나이처 스트리트
220, IBM 독일 리서치 앤드 디벨롭먼트 지엠비에이
치

명세서

청구범위

청구항 1

데이터 마이닝 모델들의 저장을 위한 컴퓨터화된 방법에 있어서, 상기 방법은:

품질(quality) 및 복잡성(complexity) 특징들 중 적어도 하나를 갖는 제1 데이터 마이닝 모델을 생성하는 단계;
 상기 특징들을 위한 기준이 충족되면, 상기 제1 데이터 마이닝 모델을 스토리지 디바이스에 저장하기 위한 후보(a candidate)인지를 결정하는 단계;

상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 디바이스에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도를 초과하는지를 결정하는 단계;

만약 상기 스토리지 한도가 초과되지 않으면, 상기 스토리지 디바이스에 상기 제1 데이터 마이닝 모델을 저장하는 단계;

만약 상기 스토리지 한도가 초과되면, 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 디바이스에 저장할 것인지를 결정하는 단계 - 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존함 -;

데이터 마이닝 요청들로부터 전사(surjective)(N-대-1) 매핑을 데이터 마이닝 작업들에 대하여 제공하는 단계;

데이터 마이닝 작업들로부터 전단사(bijective)(1-대-1) 매핑을 데이터 마이닝 모델들에 대하여 제공하는 단계 - 데이터 마이닝 작업은 연관된 데이터 마이닝 모델을 가짐 -;

각각의 저장된 데이터 마이닝 모델을 위해 각각의 작업 기술(task description)을 저장하는 단계; 및

새로운 데이터 마이닝 요청을 수신하자마자, 상기 전사 및 전단사 매핑들에 기초하여 상기 새로운 데이터 마이닝 요청과 연관된 데이터 마이닝 모델이 상기 스토리지 디바이스에 이미 저장되어 있는지를 결정하는 단계를 포함하는

방법.

청구항 2

삭제

청구항 3

청구항 1에 있어서, 상기 방법은

상기 새로운 데이터 마이닝 요청과 연관된 상기 데이터 마이닝 모델을 후-처리(post-processing)하는 단계를 포함하는

방법.

청구항 4

청구항 1에 있어서, 상기 방법은

상기 제1 데이터 마이닝 모델을 위해 상기 제1 데이터 마이닝 모델을 식별하는 정보 및 상기 제1 데이터 마이닝 모델 특징들을 기술하는 정보를 저장하는 단계를 더 포함하는

방법.

청구항 5

청구항 1에 있어서, 데이터 마이닝 모델은 데이터 패턴들의 세트를 포함하고, 데이터 패턴은 상기 데이터 마이닝 작업에 의해 결정되는 연관된 패턴 특성을 가지며, 상기 데이터 마이닝 모델의 상기 데이터 마이닝 품질은

상기 패턴 특성들에 기초하여 계산되는
방법.

청구항 6

청구항 1에 있어서, 상기 데이터 마이닝 모델 복잡성은 상기 데이터 마이닝 모델의 빌드 시간(a build time)에 기초하여 결정되는

방법.

청구항 7

청구항 1에 있어서, 상기 방법은

접근 빈도(access frequency), 스토리지 크기, 빌드 시간(build time), 및 품질 특징들 중 적어도 하나에 기초하여 데이터 마이닝 모델의 우선 순위를 계산하는 단계; 및

이미 저장된 데이터 마이닝 모델들의 우선순위들에 적어도 기초하여 상기 이미 저장된 데이터 마이닝 모델들을 상기 스토리지 디바이스로부터 제거하는 단계를 더 포함하는

방법.

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템에 있어서, 상기 시스템은:

데이터 마이닝 모델들을 저장하기 위한 스토리지 디바이스; 및

컴퓨터화된 방법을 수행하도록 구성된 데이터 프로세서를 포함하고, 상기 방법은:

품질(quality) 및 복잡성(complexity) 특징들 중 적어도 하나를 갖는 제1 데이터 마이닝 모델을 생성하는 단계;

상기 특징들을 위한 기준이 충족되면, 상기 제1 데이터 마이닝 모델을 스토리지 디바이스에 저장하기 위한 후보(a candidate)인지를 결정하는 단계;

상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 디바이스에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도를 초과하는지를 결정하는 단계;

만약 상기 스토리지 한도가 초과되지 않으면, 상기 스토리지 디바이스에 상기 제1 데이터 마이닝 모델을 저장하는 단계;

만약 상기 스토리지 한도가 초과되면, 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 디바이스에 저장할 것인지를 결정하는 단계 - 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존함 -;

데이터 마이닝 요청들로부터 전사(surjective)(N-대-1) 매핑을 데이터 마이닝 작업들에 대하여 제공하는 단계;

데이터 마이닝 작업들로부터 전단사(bijective)(1-대-1) 매핑을 데이터 마이닝 모델들에 대하여 제공하는 단계 - 데이터 마이닝 작업은 연관된 데이터 마이닝 모델을 가짐 -;

각각의 저장된 데이터 마이닝 모델을 위해 각각의 작업 기술(task description)을 저장하는 단계; 및

새로운 데이터 마이닝 요청을 수신하자마자, 상기 전사 및 전단사 매핑들에 기초하여 상기 새로운 데이터 마이닝 요청과 연관된 데이터 마이닝 모델이 상기 스토리지 디바이스에 이미 저장되어 있는지를 결정하는 단계를 포함하는

데이터 처리 시스템.

청구항 12

데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템에 있어서, 상기 시스템은:

데이터 마이닝 모델들을 저장하기 위한 스토리지 디바이스; 및

컴퓨터화된 방법을 수행하도록 구성된 데이터 프로세서를 포함하고, 상기 방법은:

품질(quality) 및 복잡성(complexity) 특징들 중 적어도 하나를 갖는 제1 데이터 마이닝 모델을 생성하는 단계; 상기 특징들을 위한 기준이 충족되면, 상기 제1 데이터 마이닝 모델을 스토리지 디바이스에 저장하기 위한 후보(a candidate)인지를 결정하는 단계;

상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 디바이스에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도를 초과하는지를 결정하는 단계;

만약 상기 스토리지 한도가 초과되지 않으면, 상기 스토리지 디바이스에 상기 제1 데이터 마이닝 모델을 저장하는 단계;

만약 상기 스토리지 한도가 초과되면, 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 디바이스에 저장할 것인지를 결정하는 단계 - 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존함 - 를 포함하되,

만약 상기 제1 데이터 마이닝 모델이 제2의 저장된 데이터 마이닝 모델보다 더 큰(greater) 우선순위를 갖는다면, 상기 스토리지 디바이스로부터 상기 제2 저장된 데이터 마이닝 모델을 제거하고(removing) 상기 스토리지 디바이스 내에 상기 제1 데이터 마이닝 모델을 저장하는 단계; 및

만약 상기 제1 데이터 마이닝 모델이 모든 저장된 데이터 마이닝 모델들보다 더 적은(lesser) 우선순위를 갖는다면, 상기 스토리지 디바이스 내에 상기 제1 저장된 데이터 마이닝 모델은 저장하지 않고(not storing) 상기 스토리지 디바이스 내에 모든 저장된 데이터 마이닝 모델들을 유지하는 단계(maintaining)를 포함하는

데이터 처리 시스템.

청구항 13

컴퓨터에 의해서, 데이터 마이닝 모델들의 저장을 위한 컴퓨터화된 방법을 수행하도록 구성된 컴퓨터 판독가능 프로그램 코드가 내부에 구현된 컴퓨터 판독가능 스토리지 매체로서, 상기 방법은:

품질(quality) 및 복잡성(complexity) 특징들 중 적어도 하나를 갖는 제1 데이터 마이닝 모델을 생성하는 단계; 상기 특징들을 위한 기준이 충족되면, 상기 제1 데이터 마이닝 모델을 스토리지 디바이스에 저장하기 위한 후보(a candidate)인지를 결정하는 단계;

상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 디바이스에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도를 초과하는지를 결정하는 단계;

만약 상기 스토리지 한도가 초과되지 않으면, 상기 스토리지 디바이스에 상기 제1 데이터 마이닝 모델을 저장하는 단계;

만약 상기 스토리지 한도가 초과되면, 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 디바이스에 저장할 것인지를 결정하는 단계 - 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존함 - ;

데이터 마이닝 요청들로부터 전사(surjective)(N-대-1) 매핑을 데이터 마이닝 작업들에 대하여 제공하는 단계;

데이터 마이닝 작업들로부터 전단사(bijective)(1-대-1) 매핑을 데이터 마이닝 모델들에 대하여 제공하는 단계 - 데이터 마이닝 작업은 연관된 데이터 마이닝 모델을 가짐 - ;

각각의 저장된 데이터 마이닝 모델을 위해 각각의 작업 기술(task description)을 저장하는 단계; 및

새로운 데이터 마이닝 요청을 수신하자마자, 상기 전사 및 전단사 매핑들에 기초하여 상기 새로운 데이터 마이닝 요청과 연관된 데이터 마이닝 모델이 상기 스토리지 디바이스에 이미 저장되어 있는지를 결정하는 단계를 포함하는

컴퓨터 판독가능 스토리지 매체.

발명의 설명

기술 분야

[0001] 본 발명은 데이터 마이닝 모델들(data mining models)의 저장을 위한 컴퓨터화된 방법들, 데이터 처리 시스템들, 및 컴퓨터 프로그램 제품들과 관련된다.

배경 기술

[0002] 데이터 마이닝(data mining)은 일반적으로 입력 데이터로부터 숨겨진 정보를 추출하기 위한 데이터-중심 접근법들(data-driven approaches)을 일컫는다. 추출된 정보는 데이터 마이닝의 유형에 의존하고 또한 데이터 마이닝 모델들에 합쳐진다. 나아가 이 모델 정보는 추가 프로세스 정보에 대비해 분석 또는 검증된다.

[0003] 데이터 마이닝 기술들은 일반적으로 대량의 데이터를 효과적으로 처리하는 방법을 고려할 필요가 있다. 첫 번째 예로서 제품들의 제조를 고려하자. 여기서, 입력 데이터는 컴포넌트들의 기원 및 특징들과 관련되는 여러 개의 데이터를 포함할 수 있다. 제조의 맥락에서 데이터 마이닝의 목적은 품질 분석 및 품질 보장(quality assurance)과 관련된 문제들을 해결하는 것일 수 있다. 데이터 마이닝은, 예를 들어, 루트 원인 분석을 위해, 제조 시설 내의 조기 경고 시스템들을 위해, 그리고 품질 보증 클레임들을 감소시키기 위해, 사용될 수 있다. 두 번째 예로서, 여러 가지 정보 기술 시스템들을 고려하자, 여기서, 데이터 마이닝은 침입 감지, 시스템 모니터링, 및 문제 분석들을 위해 더 사용될 수 있다. 데이터 마이닝은 또한, 예를 들어, 전형적으로 커스터머 행동이 분석될 수 있는 리테일(retail) 및 서비스들에서, 그리고 임상 연구들에서 인과 관계들을 찾아내기 위한 의료 및 생명 과학들에서, 여러 가지 다른 사용 예들을 갖는다. 데이터 마이닝의 일반적인 개요는, URL <http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=/com.ibm.dwe.welcome.doc/dwev9welcome.html> 에서의 IBM InfoSphere™ 웨어하우스 정보(Warehouse information)에서, 그리고 URL <http://www.redbooks.ibm.com/abstracts/sg247418.html>에서의 IBM 레드북들(Redbooks) "Dynamic Warehousing: Data Mining Made Easy" 의 1장(pp.1-8), 2장(pp.9-26), 4장(pp.39-50), 7장, 섹션 1-3(pp.139-175), 및 부록 A(pp.331-366)에서 제공된다. 이들 문서들은 데이터 마이닝의 세 가지 디스커버리 방법들, 즉, 클러스터링(clustering), 연관 규칙들(association rules), 및 시퀀스들(sequences)을 커버한다. 클러스터링 데이터 마이닝(clustering data mining)은 유사한 속성들을 갖는 데이터 레코드들의 별개의 클러스터들 또는 그룹들을 발견하기 위해 찾는다. 하나의 클러스터의 레코드들은 동종(homogeneous)이어야 하고, 두 개의 서로 다른 클러스터들의 레코드들은 가능한 한 이종(heterogeneous)이어야 한다. 연관 규칙들은 트랜잭션들(transactions) 내에서 어느 아이템들이 빈번하게 발생하는지를 기술(describe)하는 패턴들이다. 시퀀스들 데이터 마이닝(sequences data mining)은 주어진 입력 데이터에서 아이템들의 전형적인 시간 순으로 배열된 시퀀스들(time-order sequences)을 찾아낸다.

[0004] 전통적인 데이터 마이닝 시나리오들은 하나의 사용자가 데이터 마이닝 작업을 한 번 또는 수 번 설계하고 실행한다고 가정한다. CRISP-DM(Cross Industry Standard Process for Data Mining)과 같은 반복적인 데이터 마이닝 프로세스들은 이러한 시나리오들을 위해 설계되었다. 이들 프로세스들은 통상적으로 백그라운드 작업으로서 오프라인으로 수행된다. 더 상세한 내용을 위해서는, URL <http://www.crisp-dm.org>를 참조할 수 있다. 그러나, 많은 현재 및 미래의 데이터 마이닝 시나리오들은 오히려 다른 특징들을 보여준다. 즉, 먼저, 데이터 마이닝은 계획된 작업이 아니라, 인터랙티브 분석 프로세스(interactive analytical process) 동안에 호출되는 애드혹(ad hoc)이다. 두 번째로, 데이터 마이닝은 부분적으로 중첩되는 작업들을 갖는 동일한 데이터집합들(datasets) 상에서 동시에(in parallel) 서로 다른 많은 사용자들에 의해 호출된다.

[0005] 인터랙티브 데이터 마이닝(interactive data mining)은 오프라인 데이터 마이닝보다 훨씬 더 높은 응답 회수를 요구할 것이다. 이것은 분석되는 데이터집합들(datasets)이 보다 큰 경우들에 있어서 특히 심각한 문제이다. 만약 각각의 사용자가 서로 독립적으로 데이터 마이닝을 호출하면, 많은 리소스들이 낭비될 수 있다.

- [0006] 이 문제를 해결하기 위한 한 가지 방법은 데이터 마이닝 모델들을 위한 일반적인 목적의 캐시 알고리즘을 사용하는 것이다. 따라서, 만약 데이터 마이닝 모델이 구성되면, 데이터 마이닝 모델은 쿼리를 이슈한 사용자에게 리턴될 뿐만 아니라, 캐시에도 저장된다. 그런 다음, 이 캐시는 몇몇 사용자들이 동일한 모델을 공유할 수 있도록 한다. 만약 모델이 또 다른 사용자에게 의해 이미 구성되었으면, 그것은 나중에 다시 리빌드(rebuild)되어서는 아니된다. 그러나, 사용자들은 다양한 파라미터들을 명시할 수 있고 스토리지 공간은 빌드된(built) 모든 데이터 마이닝 모델들을 저장하기에는 제한된다.
- [0007] 데이터베이스의 분야에서, 의미론적 캐시들(semantic caches)이 개발되었는데, 이것들은 쿼리들이 서로 어떻게 관련되는지를 알고, 더 인텔리전트한 캐싱 전략(intelligent caching strategy)을 가능하게 하도록 이 정보를 이용한다. 세부 내용들은 S.Dar, M.J.Franklin, B.T.Jonsson, D.Srivastava, 및 M.Tan: "의미론적 데이터 캐싱 및 대체(Semantic data caching and replacement)", 매우 큰 데이터베이스들에 관한 국제 컨퍼런스의 행사들(Proceedings of the international conference on Very Large Data Bases)(1996, 페이지 330--341)에서 찾을 수 있다. 이들 데이터베이스 쿼리 캐시들은 데이터 마이닝 모델들의 캐싱(caching)을 지원하기에 적합하지 않다.
- [0008] 연관 규칙 데이터 마이닝의 분야에서는, 캐싱에 관한 알려진 기술들이 있다. B.Nag, P.Deshpande, 및 D.DeWitt, "다차원 데이터 마이닝 쿼리들을 위한 캐싱(Caching For Multi-Dimensional Data Mining Queries)", 시스템릭스(Systemics), 사이버네틱스(Cybernetics) 및 인포매틱스(Informatics)(SCI-01)에 관한 국제 컨퍼런스의 행사들(2001, 올란드, 플로리다)은, 의미론적 정보(semantic information)에 따른 연관 규칙 마이닝 쿼리들의 결과들을 저장하는 청크 기반 캐시(chunk-based cache)를 제안한다. 그러나 이 방법은 아이템집합들로 제한된다.
- [0009] T.Morzy, M.Wojciechowski, M.Zakrzewicz, "구현된 데이터 마이닝 뷰들을 사용하는 시퀀셜 패턴들의 빠른 디스커버리(Fast Discovery of Sequential Patterns Using Materialized Data Mining Views)", 컴퓨터 및 정보 과학들에 관한 15차 국제 심포지움(ICSIS) 컨퍼런스(2000)의 행사들은 기존의 구현된, 즉 캐시된 이전의 데이터 마이닝 쿼리들의 데이터 마이닝 뷰들에 대해 새로운 데이터 마이닝 쿼리들을 매핑하는 방법들을 시사한다. 이들 접근법들은 또한 연관 규칙 마이닝으로 제한된다. 상기 기존의 접근법들은 어느 모델들이 캐시되어야 하고 응답 횟수의 추가 감소가 어떻게 효과적으로 구현될 수 있는지에 관한 의문들을 해결하지 못한다.

발명의 내용

과제의 해결 수단

- [0010] 따라서, 본 발명의 목적은 제한된 스토리지 공간에 처한 경우에 더 사용하기 위해 데이터 마이닝 모델들의 저장을 위한 컴퓨터화된 방법, 데이터 처리 시스템, 및 컴퓨터 프로그램 제품을 제공하는 것이다.
- [0011] 본 발명의 제1 측면은 데이터 마이닝 모델들의 저장을 위한 컴퓨터화된 제1 방법을 제공한다. 상기 제1 방법은 다음의 단계들을 포함한다. 즉, 제1 데이터 마이닝 모델이 생성되되, 상기 제1 데이터 마이닝 모델은 다음의 특징들, 즉, 품질(quality) 및 복잡성(complexity) 중 적어도 하나를 갖는다. 만약 상기 특징들을 위한 기준이 충족되면 상기 제1 데이터 마이닝 모델은 스토리지 수단에 저장하기 위한 후보로서 핸들링된다. 상기 제1 데이터 마이닝 모델의 크기와 상기 스토리지 수단에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도(storage limit)를 초과하는지에 관한 결정이 수행된다. 만약 상기 스토리지 한도가 초과되지 않으면, 상기 제1 데이터 마이닝 모델은 상기 스토리지 수단에 저장된다. 만약 상기 스토리지 한도가 초과되면, 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 수단에 저장할 것인지에 관한 결정이 이뤄진다. 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존한다.
- [0012] 데이터 마이닝 작업들에 대한 데이터 마이닝 요청들로부터의 전사(surjective)(N-대-1) 매핑 및 데이터 마이닝 모델들에 대한 데이터 마이닝 작업들로부터의 전단사(bijective)(1-대-1) 매핑이 제공될 수 있다. 데이터 마이닝 작업은 연관된 데이터 마이닝 모델을 가질 수 있다. 각각의 저장된 데이터 마이닝 모델을 위해, 각각의 작업 기술(task description)이 저장될 수 있다. 새로운 데이터 마이닝 요청을 수신하자마자, 상기 전사 및 상기 전단사 매핑에 기초하여 상기 새로운 데이터 마이닝 요청이 이미 상기 스토리지 수단에 저장되어 있는지에 관한 결정이 수행될 수 있다.
- [0013] 상기 새로운 데이터 마이닝 요청과 연관된 데이터 마이닝 모델의 후-처리(post-processing)가 수행될 수 있다. 상기 제1 데이터 마이닝 모델을 위해, 비록 상기 제1 데이터 마이닝 모델이 추가 사용을 위해 저장되지 않더라도

도 상기 제1 데이터 마이닝 모델을 식별하는 정보 및 상기 제1 데이터 마이닝 모델 특징들을 기술(describe)하는 정보는 저장될 수 있다. 데이터 마이닝 모델은 데이터 패턴들의 집합(set)을 포함할 수 있다. 데이터 패턴은 상기 데이터 마이닝 작업에 의해 결정되는 연관된 패턴 특성을 가질 수 있다. 상기 데이터 마이닝 모델의 데이터 마이닝 품질은 상기 패턴 특성들에 기초하여 계산될 수 있다. 상기 데이터 마이닝 모델 복잡성은 상기 데이터 마이닝 모델의 빌드 시간(build time)에 기초하여 결정될 수 있다. 데이터 마이닝 모델의 우선 순위는 다음의 특징들, 즉, 접근 빈도(access frequency), 스토리지 크기, 빌드 시간, 및 품질 중 적어도 하나에 기초하여 계산될 수 있다. 상기 스토리지 수단으로부터, 이미 저장된 데이터 마이닝 모델들은 적어도 그것들의 특성들에 기초하여 제거될 수 있다. 상기 전사 매핑은 데이터 선택 제한들(data selection constraints) 및 데이터 마이닝 제한들(data mining constraints)에 기초하여 데이터 마이닝 요청들과 데이터 마이닝 작업들을 비교하는 것을 포함할 수 있다.

[0014] 본 발명의 제2 측면은 데이터 마이닝 모델들의 저장을 위해 컴퓨터화된 제2 방법을 제공한다. 상기 제2 방법은 다음의 단계들을 포함한다. 즉, 데이터 마이닝 모델들을 위해, 데이터 마이닝 모델을 식별하는 정보 및 데이터 마이닝 모델 품질을 기술하는 정보가 스토리지 수단에 저장된다. 제1의 미리 정의된 기준을 수행하는 데이터 마이닝 모델 품질을 갖는 데이터 마이닝 모델들은 상기 스토리지 수단에 저장된다. 새로운 데이터 마이닝 요청을 수신하자마자, 데이터 마이닝 모델은 상기 새로운 데이터 마이닝 요청을 위해 결정되고, 상기 데이터 마이닝 모델을 위한 데이터 마이닝 모델 품질을 기술하는 정보가 상기 스토리지 수단에 저장되었는지에 관한 체크가 수행된다. 만약 상기 데이터 마이닝 모델의 데이터 마이닝 품질이 제2의 미리 정의된 기준을 충족시키지 않으면, 사용자는 데이터 마이닝이 진행되는 것을 확인하도록 요청된다.

[0015] 데이터 마이닝 모델은 데이터 패턴들의 집합을 포함할 수 있다. 데이터 패턴은 상기 데이터 마이닝 작업에 의해 결정되는 연관된 패턴 특성을 가질 수 있다. 상기 데이터 마이닝 모델의 데이터 마이닝 품질은 상기 패턴 특성들에 기초하여 계산될 수 있다.

[0016] 본 발명의 제3 측면은 데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템을 제공한다. 상기 시스템은 데이터 마이닝 모델들을 저장하기 위한 스토리지 수단(storage means) 및 데이터 처리 수단(data processing means)을 포함한다. 상기 데이터 처리 수단은 제1 데이터 마이닝 모델을 생성하는데, 상기 제1 데이터 마이닝 모델은 다음의 특징들, 즉, 품질 및 복잡성 중 적어도 하나를 갖는다. 만약 상기 특징들을 위한 기준이 충족되면, 상기 데이터 처리 수단은 상기 제1 데이터 마이닝 모델을 상기 스토리지 수단에 저장하기 위한 후보로서 핸들링된다. 상기 데이터 처리 수단은 상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 수단에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도(storage limit)를 초과하는지를 결정한다. 만약 상기 스토리지 한도가 초과되지 않으면, 상기 데이터 처리 수단은 상기 스토리지 수단에 상기 제1 데이터 마이닝 모델을 저장한다. 만약 상기 스토리지 한도가 초과되면, 상기 데이터 처리 수단은 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 수단에 저장할 것 인지를 결정한다. 상기 특성들은 적어도 상기 각각의 데이터 마이닝 모델의 접근 빈도들에 의존한다.

[0017] 본 발명의 제4 측면은 데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템을 제공한다. 상기 시스템은 스토리지 수단, 입력 수단, 데이터 처리 수단, 및 출력 수단을 포함한다. 상기 스토리지 수단은 데이터 마이닝 모델들을 위해 데이터 마이닝 모델을 식별하는 정보 및 데이터 마이닝 모델 품질을 기술하는 정보를 저장하고, 상기 스토리지 수단은 데이터 마이닝 모델들을 저장한다. 상기 입력 수단은 새로운 데이터 마이닝 요청을 수신한다. 상기 데이터 처리 수단은 데이터 마이닝 모델을 식별하는 정보 및 데이터 마이닝 모델 품질을 기술하는 정보를 상기 스토리지 수단에 저장한다. 상기 데이터 처리 수단은 제1의 미리 정의된 기준을 충족시키는 데이터 마이닝 모델 품질을 갖는 데이터 마이닝 모델들을 상기 스토리지 수단에 저장한다. 상기 데이터 처리 수단은, 상기 새로운 데이터 마이닝 요청을 위한 데이터 마이닝 모델을 결정하고 또한 상기 데이터 마이닝 모델을 위해 데이터 마이닝 모델 품질을 기술하는 정보가 상기 스토리지 수단에 저장되었는지를 체크한다. 만약 상기 데이터 마이닝 모델의 데이터 마이닝 모델 품질이 제2의 미리 정의된 기준을 충족시키지 않는다면, 상기 데이터 처리 수단은 데이터 마이닝이 진행되는 것을 확인하기 위해 사용자에게 요청을 개시한다. 상기 출력 수단은 상기 사용자에게 상기 요청을 보낸다.

[0018] 본 발명의 제5 측면은 데이터 마이닝 모델들의 저장을 위한 컴퓨터 프로그램 제품(computer program product)을 제공한다. 상기 컴퓨터 프로그램 제품은 컴퓨터 사용가능 매체(computer usable medium)를 포함하는데, 상기 컴퓨터 사용가능 매체는 내부에 구현되는 컴퓨터 사용가능 프로그램 코드(computer usable program code)를 갖는다. 상기 컴퓨터 사용가능 프로그램 코드는 본 발명의 제1 및 제2 측면들 중 어느 하나에 따른 단계들을 수행하

도록 구성된다.

도면의 간단한 설명

- [0019] 본 발명에 관한 더 나은 이해를 위해 그리고 어떻게 본 발명이 효과적으로 수행될 수 있는지에 관해, 이제 첨부되는 도면들을 참조하여 단지 예시의 방법으로서 참조가 이루어질 것이다.
- 도 1은 본 발명의 일 실시예에 따른 데이터 마이닝 모델들의 저장을 위한 제1 방법의 흐름도를 도시한다.
- 도 2 및 3은 본 발명의 일 실시예에 따른 데이터 마이닝 모델들의 저장을 위한 제2 방법의 흐름도들을 도시한다.
- 도 4는 본 발명의 일 실시예에 따른 클러스터링 데이터 마이닝 모델들의 저장을 위한 일 예를 도시한다.
- 도 5 내지 7은 본 발명의 일 실시예에 따른 데이터 마이닝 모델들의 선택적 그리고 인터랙티브 저장을 설명하는 흐름도를 도시한다.
- 도 8 내지 10은 본 발명의 일 실시예에 따른 연관 규칙 데이터 마이닝 모델들의 저장을 위한 샘플 시나리오를 도시한다.
- 도 11은 본 발명의 일 실시예에 따른 데이터 마이닝 모델들의 저장을 위한 제1 데이터 처리 시스템의 블록도를 도시한다.
- 도 12는 본 발명의 일 실시예에 따른 데이터 마이닝 모델들의 저장을 위한 제2 데이터 처리 시스템의 블록도를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0020] 데이터 마이닝 모델은 데이터 마이닝 작업에 기초하여 계산된다. 데이터 마이닝 작업 기술(task description)은 데이터 마이닝 유형, 데이터 마이닝이 적용될 수 있는 데이터를 위한 데이터 선택 제한들(data selection constraints), 및 상기 데이터 마이닝 유형에 의존하는 데이터 마이닝 제한들(data mining constraints)을 명시한다. 데이터 마이닝 요청은 사용자에게 의한 데이터 마이닝 작업의 실제 호출이다. 상기 요청은 대응하는 데이터 마이닝 작업에 대한 전사(surjective) 다-대-일(many-to-one)(N-대-1) 매핑을 위한 파라미터들을 포함한다. 이 전사 매핑은 서로 다른 데이터 마이닝 요청들이 동일한 데이터 마이닝 작업과 연관될 수 있도록 한다(비록 상기 요청들 및 작업들이 서로 다른 데이터 마이닝 제한들 및 기타 다른 파라미터들을 갖더라도). 전단사(bijective) 일-대-일(one-to-one)(1-대-1) 매핑은 데이터 마이닝 작업을 데이터 마이닝 모델과 연관시킨다. 상기 매핑들은 프로그램 코드 부분들 및 구성 파라미터들(configuration parameters)에 기초하여 수행될 수 있다. 본 발명의 실시예들에 따른 의미론적 캐시는 상기 데이터 마이닝 모델들을 선택적으로 저장하기 위해 데이터 마이닝 모델 특징들을 고려한다. 캐싱(caching) 또는 구현(materialization)은 임시 메모리의 부분들에서 또는 영구적인 스토리지 디바이스(예를 들어, 하드 디스크)의 부분들에서의 추가 사용을 위한 데이터 마이닝 모델들 정보를 저장하는 것을 의미한다. 상기 캐시된 정보가 다시 요청되는 경우, 그것은 사용자에게 더 빠르게 리턴될 수 있다. 왜냐하면 상기 데이터 마이닝 모델 정보가 재계산(re-compute)될 필요가 없기 때문이다. 상기 캐시에 저장된 정보로부터 직접 응답받을 수 있는 사용자 요청들의 퍼센티지는 캐시 히트율(cache hit rate)로 일컬어진다. 데이터 마이닝 모델들이 요청되는 경우, 상기 캐시의 내용들은 재계산된 우선 순위들에 기초하여 동적으로(dynamically) 갱신되는데, 재계산된 우선 순위들은 적어도 상기 데이터 마이닝 모델들의 접근 빈도들에 기초하여 계산된다. 이것은 계산하기에 어렵고 빈번하게 사용되는 데이터 마이닝 모델들에 대해 짧은 응답 시간들을 감안하며, 상기 메모리 또는 다른 스토리지 디바이스들 내에서 불필요하게 많은 스토리지 공간을 소모하지 않는다.

[0021] 도 1은 데이터 마이닝 모델들의 저장을 위한 제1 방법의 흐름도(100)를 도시한다. 단계 101에서, 제1 데이터 마이닝 모델이 생성되는데, 이는 다음의 특징들, 즉 품질 및 복잡성 중 적어도 하나를 갖는다. 단계 102에서, 상기 특징들을 위한 기준이 충족되는지에 관한 테스트가 수행된다. 만약 아니오라면, 처리는 종료된다. 만약 예라면, 단계 103에서 상기 제1 데이터 마이닝 모델은 스토리지 수단을 저장하기 위한 후보로 핸들링된다. 단계 104에서, 상기 제1 데이터 마이닝 모델의 크기 및 상기 스토리지 수단에 이미 저장된 추가 데이터 마이닝 모델들의 크기의 합이 결정된다. 단계 105에서, 이 합이 미리 정의된 스토리지 한계를 초과하는지에 관한 테스트가 수행된다. 만약 상기 스토리지 한도가 초과되지 않으면, 단계 107에서 상기 제1 데이터 마이닝 모델은 상기 스토리지 수단에 저장된다. 만약 상기 스토리지 한도가 초과되면, 단계 106에서 상기 제1 데이터 마이닝 모델 및 상기

추가 데이터 마이닝 모델들의 특성들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 수단에 저장할 것인지가 결정된다. 상기 특성들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존한다.

[0022] 도 2는 데이터 마이닝 모델들의 저장을 위한 제2 방법의 제1 흐름도(200)를 도시한다. 제2 방법에서, 단계 201로 도시된 바와 같이, 데이터 마이닝 모델을 식별하는 정보 및 데이터 마이닝 모델 품질을 기술하는 정보는 미리 생성된 데이터 마이닝 모델들을 위해 스토리지 수단에 저장된다. 단계 202로 도시된 바와 같이, 상기 미리 생성된 데이터 마이닝 모델들은 그것들의 데이터 마이닝 모델 품질들이 제1의 미리 정의된 기준을 충족시키는지 테스트된다. 만약 예라면, 단계 203으로 도시된 바와 같이, 이들 미리 생성된 데이터 마이닝 모델들은 상기 스토리지 수단에 저장된다. 그렇지 않으면, 그것들은 저장되지 않는다.

[0023] 도 3은 데이터 마이닝 모델들의 저장을 위한 제2 방법의 제2 흐름도(300)를 도시한다. 단계 301에서 새로운 데이터 마이닝 요청이 수신된다. 단계 302에서 상기 새로운 데이터 마이닝 요청을 위한 데이터 마이닝 모델이 결정된다. 단계 303은 이 데이터 마이닝 모델을 위해 데이터 마이닝 모델 품질을 기술하는 정보가 도 2의 단계 201에서 상기 스토리지 수단에 이미 저장되었는지를 체크한다. 만약 단계303의 체크가 실패(fail)이면(아니오 분기), 상기 새로운 데이터 마이닝 요청을 위한 처리는 종료된다. 만약 단계 303의 체크가 성공이면(예 분기), 단계 304는 상기 데이터 마이닝 모델 품질이 제2의 미리 정의된 기준을 충족시키는지 테스트한다. 만약 단계 304의 테스트가 실패이면, 상기 새로운 데이터 마이닝 요청은 더 처리되지 않는다. 만약 단계 304의 테스트가 성공이면, 단계 305에서 사용자는 데이터 마이닝이 진행되는 것을 확인하도록 요청받는다. 만약 단계 306에서 상기 사용자가 확인하면, 단계 307에서 상기 새로운 데이터 마이닝 요청을 위해 데이터 마이닝이 진행된다. 그렇지 않으면, 이 데이터 마이닝 요청에 관한 처리는 계속되지 않는다.

[0024] 도 4는 클러스터링 데이터 마이닝 모델들의 저장을 위한 일 예를 도시한다. 상기 예는 네 개의 데이터 마이닝 요청들(411, 412, 413, 및 414)을 포함하며, 각각은 식별자들 R1, R2, R3, 및 R4를 갖는다. 상기 데이터 마이닝 유형은 항상 클러스터이다. 각각의 요청은 상기 테이블들 A, B, 및 C 중 하나에 관한 데이터 선택을 명시한다. 상기 전사 매핑 절차는 상기 데이터 마이닝 유형 및 상기 데이터 선택 제한들을 대응하는 데이터 마이닝 작업에 대한 상기 데이터 마이닝 요청으로부터 복사하는데, 상기 대응하는 데이터 마이닝 작업은 상기 데이터 마이닝 작업들(421, 422, 및 423) 중 하나이고, 각각은 식별자들 T1, T2, 및 T3을 갖는다. 상기 요청의 데이터 마이닝 제한들은 다음과 같이 상기 작업의 데이터 마이닝 제한들에 매핑된다. 즉, 사용자는 데이터 마이닝 모델 계산을 위한 클러스터들의 최대 수를 명시할 수 있다. 예를 들어, 데이터 마이닝 요청(412)에서 MaxNum=10 그리고 데이터 마이닝 요청(414)에서 MaxNum=2이다. 데이터 마이닝 요청은 클러스터의 최대 수 MaxNum와 같거나 더 높은 수를 갖고서 이미 존재하는 데이터 마이닝 작업에 매핑될 수 있다. 사용자가 클러스터들의 최대 수를 명시하지 않는 경우, 예를 들어, 데이터 마이닝 요청들(411 및 413)에서 MaxNum=auto인 경우, 상기 클러스터링 데이터 마이닝 시스템은 동일한 데이터 마이닝 유형 및 데이터 선택 제한들을 갖는 이미 저장된 데이터 마이닝 작업들에 대해 이러한 요청들을 매핑하려고 한다. 만약 이것이 가능하지 않으면, MaxNum을 위해 미리 정의된 값을 갖는 새로운 작업 기술(task description)이 저장된다. 그런 다음, 상기 데이터 마이닝 시스템은 상기 데이터 마이닝 작업과 연관된 데이터 마이닝 모델을 생성한다.

[0025] 제1 사용자는 클러스터들의 최대 수를 명시함이 없이, 예를 들어, MaxNum=auto로 하여 테이블 A의 클러스터 데이터에 대한 데이터 마이닝 요청(411)을 보낸다. 상기 데이터 마이닝 시스템은 새로운 데이터 마이닝 작업 T1에 대해 식별자 R1을 갖는 제1 요청(411)을 매핑하고, 데이터 마이닝 작업 T1을 위해 클러스터들의 최대 수를 MaxNum=10의 미리 정의된 값으로 설정하고, 연관된 데이터 마이닝 모델 M1을 계산한다고 가정하자. 식별자 R2를 갖는 제2 요청(412)은 동일한 테이블 A 및 값 MaxNum=10 또는 그보다 낮은 값을 명시한다. 상기 데이터 마이닝 시스템은 상기 자동으로 결정된 파라미터가 실제로 MaxNum=10으로 평가되었음을 안다. 그런 다음, 상기 제2 요청은 동일한 데이터 마이닝 작업 T1에 매핑될 것이다. 제2 요청(R2)을 위한 새로운 데이터 마이닝 모델을 계산하는 것 대신에, 캐시된 데이터 마이닝 모델 M1이 사용자에게 리턴될 수 있다.

[0026] 데이터 마이닝 작업 T1과 동일한 데이터 선택 제한들을 가지며 데이터 마이닝 작업 T1과 같거나 더 낮은 MaxNum 값을 명시하는 클러스터링 데이터 마이닝 요청은 데이터 마이닝 작업 T1에 매핑될 수 있다. 그러므로, 이 데이터 마이닝 작업 T1은 데이터 마이닝 모델 M1을 생성하는 균등하거나 더 일반적인 데이터 마이닝 작업 기술로서 고려될 수 있다. 만약 클러스터링 데이터 마이닝 요청 R5가 데이터 마이닝 모델 M1에서 클러스터들의 실제 수보다 더 낮은 MaxNum 값을 가지며 또한 데이터 마이닝 요청 R5가 데이터 마이닝 작업 T1에 매핑되었다면, 데이터 마이닝 모델 M1A를 계산하기 위해 후-처리가 요구된다. M1A가 데이터 마이닝 요청 R5에 명시된 클러스터들의 수를 정확하게 포함할 때까지, 상기 후-처리는 데이터 마이닝 모델 M1에 포함된 클러스터들의 일부를 종합하는 것을 포함한다. 만약 클러스터링 데이터 마이닝 요청 R6이 동일한 데이터 선택 제한들을 갖는 기존의 데이터 마이

닝 작업 T1보다 더 높은 MaxNum 값을 가지면, 데이터 마이닝 요청 R6의 매핑은 불가능할 수 있다. 그러면, 새로운 데이터 마이닝 작업 T6은 상기 요청된 MaxNum 값을 갖고서 생성되고 연관된 데이터 마이닝 모델 M6이 빌드된다.

- [0027] 식별자 R3를 갖는 새로운 데이터 마이닝 요청(413)에서, 클러스터들의 최대 수는 명시되지 않는다. 즉, MaxNum=auto이다. 새로운 요청 R3은 식별자 T2를 갖는 새로운 데이터 마이닝 작업(422)에 매핑된다. 이 데이터 마이닝 작업은 작업 T1으로부터 독립적이다. 왜냐하면 상기 데이터 선택은 다른 테이블 B에 기초하기 때문이다. 데이터 마이닝 작업(422)의 클러스터들의 최대 수는 미리 정의된 값으로 설정된다. 즉, MaxNum=5이다. 연관된 데이터 마이닝 모델 M2가 생성되는데, 이는 5 또는 더 적은 클러스터들을 갖는다.
- [0028] 식별자 R4를 갖는 추가의 새로운 데이터 마이닝 요청(414)은 테이블 C 상의 데이터 선택을 명시한다. 이 요청은 식별자 T3을 갖는 새로운 데이터 마이닝 작업(423)에 매핑된다. 이 작업은 작업들 T1 및 T2로부터 독립적이다. 왜냐하면 데이터 마이닝은 또 다른 테이블 C 상에서 수행되기 때문이다. 식별자 R4를 갖는 요청(414)은 클러스터들의 최대 수를 명시한다. 즉, MaxNum=2이고, 클러스터들의 동일한 최대 수를 갖고서 식별자 T3을 갖는 작업(423)에 매핑된다. 즉 MaxNum=2이다. 데이터 마이닝 작업(423)과 연관된 데이터 마이닝 모델 M3이 구성되는 경우, 그것은 데이터 마이닝 작업(T3)에 명시된, 클러스터들의 최대 수, MaxNum=2보다 더 작은 클러스터들을 가질 수 있다.
- [0029] 각각의 데이터 마이닝 작업(T1, T2, T3)의 실행은 결국 연관된 데이터 마이닝 모델(M1, M2, M3)을 야기한다. 각각의 데이터 마이닝 모델은 데이터 마이닝 모델 특징들의 집합(431, 432, 433)를 갖는데, 이는 데이터 마이닝 모델의, 품질, 복잡성, 접근 카운트(access count) 또는 접근 빈도(access frequency), 및 스토리지 크기에 관한 정보를 포함한다. 모든 특징들은 측정가능하다. 즉, 이것들은 숫자 값들로 표현된다.
- [0030] 예를 들어, 데모그래픽 클러스터링(demographic clustering) 및 코호넨 뉴럴 클러스터링(Kohonen neural clustering) 과 같은 멀티플 데이터 마이닝 방법들이 알려져 있다. 두 가지 방법들은 클러스터링 모델들을 야기 하는데, 각각의 모델은 별개의 수의 클러스터들(discrete number of clusters)을 포함한다. 상기 클러스터링 모델은 범위 [0 ... 1]에서 모델 품질 넘버 Q를 포함할 수 있다. 상기 데모그래픽 클러스터링 모델의 모델 품질 넘버 Q는 클러스터들의 전체 동질성(homogeneity)이며, 클러스터의 동질성은 상기 클러스터의 두 개의 레코드들 사이의 평균 유사성(mean similarity)이다. 상기 코호넨 클러스터링의 경우, 상기 모델 품질 넘버는 상기 클러스터의 데이터 레코드들과 상기 최상의 매칭 클러스터 센터 사이의 평균 거리에 의해 측정될 때 상기 클러스터들의 전체 동질성이다.
- [0031] 상기 데이터 마이닝 모델 복잡성은 상기 데이터 마이닝 모델의 빌드 시간(build time)에 기초하여 결정된다. 이는 상기 데이터 마이닝 모델을 즉석에서(on the fly) 계산하기 위한 시간이다. 상기 접근 카운트 또는 접근 빈도는 시간 창(window)에서 발생하는 데이터 마이닝 모델의 히트들의 일부(fraction)이다. 상기 스토리지 크기는 상기 데이터 마이닝 모델이 캐시에서 소비하는 스토리지량이다. 상기 캐시는 임시 메모리로, 또는 예를 들어, 하드 디스크와 같은 영구 스토리지로 구현될 수 있다.
- [0032] 상기 네 개의 데이터 마이닝 모델 특성들 중 어떤 것을 위해 실제 측정된 값들을 갖는 것 대신에, 대응하는 값들은 또한 다른 특성들(예를 들어, 상기 데이터 마이닝 작업으로부터의 세부 사항들(details))에 기초하여 평가될 수 있다. 이 평가는 평가된 데이터 마이닝 모델 품질, 평가된 데이터 마이닝 모델 복잡성, 평가된 접근 카운트 또는 평가된 접근 주파수, 및 평가된 데이터 마이닝 모델 크기를 제공할 수 있다. 상기 결정된 데이터 마이닝 모델 특성들은 상기 데이터 마이닝 모델의 완료된 정보를 저장하는 것에 관한 결정들에 독립적으로 상기 캐시에 저장된다.
- [0033] 상기 데이터 마이닝 모델 특성들에 기초하여, 구현 기준(materialization criterion)이 평가될 수 있다. 구현 수단은 추가 사용을 위해 완전한 데이터 마이닝 모델을 저장한다. 상기 평가는 각각의 데이터 마이닝 모델을 위한 구현 상태(materialization state)를 야기한다. 상기 구현 상태 변수(materialization state variable)는 세 개의 값들 중 하나를 가질 수 있다.
- [0034] 1. "Materialize=YES"는, 만약 데이터 마이닝 모델 품질이 너무 낮지 않고 상기 데이터 마이닝 모델의 빌드 시간이 너무 짧지 않으면, 그리고 상기 캐시의 스토리지 한도가 초과되지 않으면, 계산된 데이터 마이닝 모델이 상기 캐시에 저장되어야 함을 나타낸다.
- [0035] 2. "Materialize=NO"는, 만약 상기 데이터 마이닝 모델 품질이 너무 낮거나 상기 데이터 마이닝 모델의 빌드 시간이 너무 짧으면, 계산된 데이터 마이닝 모델이 상기 캐시에 저장되지 않아야 함을 나타낸다.

- [0036] 3. "Materialize=MAYBE"는, 상기 캐시의 스토리지 한도가 초과되기 때문에 심지어 상기 데이터 마이닝 모델 품질이 너무 낮지 않고 상기 빌드 시간이 너무 짧지 않은 경우, 제1 데이터 마이닝 요청을 위한 계산 후 데이터 마이닝 모델이 저장되지 않아야 함을 나타낸다. 계속되는 요청들에서, 상기 캐시가 충분한 자유 스토리지 공간을 갖는 경우 상기 데이터 마이닝 모델은 재계산될 수 있고, 결국 상기 캐시에 저장된다.
- [0037] 상기 구현 기준은 다음과 같은 식을 사용하여 표현될 수 있다.
- [0038] IF (ModelQuality < MinimumQuality OR ModelComplexity <
- [0039] MinimumComplexity)
- [0040] Materialize = NO
- [0041] ELSEIF (CacheStorageLimitNotExceeded)
- [0042] Materialize = YES
- [0043] ELSE
- [0044] Materialize = MAYBE
- [0045] ENDIF
- [0046] 다음의 스레시홀드 값들(즉, MinimumQuality=0.5 및 MinimumComplexity=0.3)의 경우, 구현 상태(442)의 계산은 모델 M2을 위해 Materialize=NO를 리턴한다. 왜냐하면, 그것의 품질이 상기 제1 스레시홀드보다 낮고, 그것의 복잡성이 상기 제2 스레시홀드보다 낮기 때문이다. 두 번째 이유는 상기 모델 빌드 시간이 최소 빌드 시간보다 더 작다는 것을 의미한다. 상기 데이터 마이닝 모델은 즉석에서 쉽게 계산될 수 있으므로, 작업 T2를 위한 모델 M2는 구현되지 않는다.
- [0047] 각각의 모델들(M1 및 M3)을 위한 구현 상태들(441 및 443)은 Materialize=YES 및 MAYBE이다. 이는 상기 캐시를 위한 스토리지 한도가 이미 초과했는지 그렇지 않은지에 의존한다. 그러므로, 작업 T1을 위한 모델 M1은 그것이 상기 미리 정의된 스레시홀드들보다 더 높은 품질 및 복잡성을 갖고, 상기 캐시 스토리지 한도가 아직 초과되지 않기 때문에 구현된다. 작업 T3을 위한 모델 M3은 그것의 접근 빈도가 증가하는 경우 장차 또한 구현될 수 있다. 모델의 구현 상태가 YES 또는 MAYBE로 설정되는 경우, 상기 모델은 우선순위가 주어지고, 이는 전형적으로 상기 데이터 마이닝 모델 크기의 스토리지 크기 및 그것의 접근 카운트에 의존할 수 있다.
- [0048] 도 5 내지 7은 데이터 마이닝 모델들의 선택적이고 인터랙티브 저장을 설명하는 흐름도를 도시한다. 이들 흐름도들은 연결 엘리먼트들(connector elements)(1, 2, 3, 4, 및 5)에 의해 인터링크(interlinke)된다.
- [0049] 상기 데이터 마이닝 시스템이 첫 번째로 사용되는 경우, 관리자 사용자(administrator user)는 많은 사용자들을 위해 적절하다고 고려되고 즉석에서 계산될 수 없는 몇 가지 데이터 마이닝 작업들을 명시할 수 있다. 상기 데이터 마이닝 시스템은 이들 데이터 마이닝 작업들이 상기 연관된 데이터 마이닝 모델들에 관한 미리-계산(pre-computation)을 위한 그리고 상기 캐시에 그것들을 저장하기 위한 구현 기준을 충족시키는지 체크한다.
- [0050] 도 5에 도시된 바와 같이, 단계 501에서, 상기 데이터 마이닝 시스템은 데이터 마이닝 요청을 수신한다. 단계 502에서, 상기 데이터 마이닝 시스템은 상기 수신된 데이터 마이닝 요청에 동등한(equivalent) 데이터 마이닝 작업 또는 상기 수신된 데이터 마이닝 요청보다 더 일반적인 데이터 마이닝 작업이 상기 캐시에 저장되는지를 체크한다. 두 번째의 경우, 상기 수신된 데이터 마이닝 요청은 상기 캐시된 데이터 마이닝 작업보다 더 구체적이거나 더 제한적이다. 데이터 마이닝 요청들 및 데이터 마이닝 작업들의 제한들은 파라미터들의 결합들(conjunctions)로 표현된다. 상기 파라미터들은 간단히 바이너리 관계들(binary relations)을 포함하는 문자들일 수 있다. 예를 들어, DataSource="Table A" 및 MaxNum = "10"이다. 그런 다음, 단계502는 간단한 논리적 한정(entailment)에 의해 매우 효과적으로 수행될 수 있다. 데이터 마이닝 작업들에 대한 이러한 데이터 마이닝 요청들로부터의 전사 매핑, 즉, 다-대-일 매핑은, 요청들의 집합 및 작업들의 집합 사이의 바이너리 관계이다. 바람직한 실시예는 이 관계를 실제로 결정하기 위해 두 가지의 메커니즘들을 채용한다. 첫 번째 메커니즘은 파라미터가 상기 요청 기술(request description)에 명시되지 않는 경우, 예를 들어, MaxNum = auto인 경우 적용된다. 상기 데이터 마이닝 시스템은 상기 데이터 마이닝 작업을 위해 이러한 미리 명시되지 않은 파라미터를 자동으로 결정한다. 예를 들어, MaxNum = 10이다. 두 번째 메커니즘은 서로 다른 제한들 사이의 포함 관계

(subsumption relation)를 사용한다. 요청이 더 특정된 제한들을 충족시키는 경우, 그것은 더 일반적인 제한들을 갖는 작업에 대해 매핑될 수 있다. 예를 들어, 데이터 마이닝 제한 MaxNum = 2를 갖는 데이터 마이닝 요청은 데이터 마이닝 제한 MaxNum = 10을 갖고서 데이터 마이닝 작업에 매핑될 수 있다. 왜냐하면 두 번째 제한은 첫 번째보다 더 일반적이기 때문이다. 상기 데이터 마이닝 요청은 기존의 데이터 마이닝 작업에 성공적으로 매핑된 경우, 처리는 도 7에서 연결 엘리먼트 1을 통해 계속된다.

[0051] 만약 단계 502에서 상기 수신된 데이터 마이닝 요청을 위해 어떠한 적절한 데이터 마이닝 작업이 상기 캐시에서 발견될 수 있으면, 단계 503에서 상기 데이터 마이닝 시스템은 작업 기술(task description)로서 상기 수신된 데이터 마이닝 요청의 캐시 파라미터들에 저장할 수 있다. 상기 작업 기술과 빌드(build)될 데이터 마이닝 모델 사이의 매핑은 전단사, 즉, 일-대-일 매핑이다. 그런 다음, 단계 504에서, 상기 데이터 마이닝 시스템은 대응하는 데이터 마이닝 모델 특징들을 갖는 데이터 마이닝 모델을 구성하고, 상기 캐시에 상기 데이터 마이닝 모델 특징들을 저장할 수 있다. 상기 데이터 마이닝 시스템은 상기 데이터 마이닝 모델로부터의 정보를 갖고서 상기 대응하는 작업 기술을 갱신할 수 있다. 만약 상기 대응하는 데이터 마이닝 모델 특징들이 이미 상기 캐시에 저장되어 있으면, 그것들은 갱신될 수 있다. 상기 데이터 마이닝 모델 특징들은 모델 식별자, 모델 품질에 관한 정보, 스토리지 크기, 접근 카운트 및 빌드 시간을 포함할 수 있다. 데이터 마이닝 작업 기술들(descriptions) 및 데이터 마이닝 모델 특징들은, 그들 자신들이 상기 캐시에 저장되지 않는 데이터 마이닝 모델들을 위해 상기 캐시에 저장될 수 있다.

[0052] 단계 505에서, 상기 데이터 마이닝 시스템은 상기 데이터 마이닝 모델 품질이 미리 정의된 최소 품질보다 더 낮은지 또는 상기 빌드 시간이 미리 정의된 최소 빌드 시간보다 더 낮은지를 체크한다. 만약 예라면, 상기 빌드 데이터 마이닝 모델은 구현되지 않을 것이고, 상기 프로세스는 단계 508로 계속된다. 단계 505의 결과가 "아니오"(no)인 경우, 상기 데이터 마이닝 시스템은 상기 빌드된 데이터 마이닝 모델의 크기 및 상기 캐시에 미리 저장된 추가 데이터 마이닝 모델들의 합을 결정하고, 이 합이 미리 정의된 스토리지 한도를 초과하는지를 테스트한다. 이것은 만약 상기 빌드된 데이터 마이닝 모델이 저장되었다면 상기 캐시가 스토리지 공간을 다 쓸 것이라는 것을 의미한다. 상기 스토리지 한도가 초과되지 않는 경우, 상기 데이터 마이닝 시스템은 단계 507에서 상기 캐시에 상기 빌드된 데이터 마이닝 모델을 저장한다. 단계 508에서, 상기 빌드된 데이터 마이닝 모델은 사용자에게 리턴되고, 이는 상기 캐시에 상기 데이터 마이닝 모델을 저장하는 것으로부터 독립적이거나 그렇지 않다. 만약 단계 506에서 상기 스토리지 한도가 초과되는 경우, 처리는 도 6에서 연결 엘리먼트 3을 통해 계속된다.

[0053] 단계 6에서, 처리는 도 5로부터 연결 엘리먼트(3)을 통해 시작된다. 단계 601에서, 상기 데이터 마이닝 시스템은 상기 빌드된 데이터 마이닝 모델 및 상기 이미 저장된 데이터 마이닝 모델들의 우선순위들을 계산한다. 데이터 마이닝 우선순위는 다음의 특성들, 즉, 데이터 마이닝 모델의 스토리지 크기, 그것의 접근 카운트 또는 접근 빈도, 그것의 빌드 시간 또는 그것의 계산 시간, 및 그것의 품질 중 적어도 하나에 기초하여 결정될 수 있다. 상기 데이터 마이닝 우선순위는 상기 스토리지 크기와 상기 데이터 마이닝 모델의 접근 카운트의 곱(product)일 수 있다. 상기 접근 카운트는 대응하는 데이터 마이닝 모델을 제공하기 위해 사용자 요청들의 수인데, 이는 사용자 요청들의 총 수에 의해 나뉘질 수 있다. 상기 접근 카운트 또는 접근 빈도 측정은 전형적으로 시간 창으로 한정된다. 바람직하게는, 상기 빌드된 데이터 마이닝 모델은 하나의 접근 빈도 값을 얻는다.

[0054] 단계 602에서, 상기 데이터 마이닝 시스템은, 이미 저장된 데이터 마이닝 모델들이 가장 낮은 우선 순위들을 갖는지를 테스트한다. 만약 아니라면, 상기 빌드된 데이터 마이닝 모델은 가장 낮은 우선 순위이고, 상기 데이터 마이닝 시스템은 그것을 상기 캐시에 저장하지 않도록 결정한다. 처리는 도 5에서 연결 엘리먼트(5)를 통해 계속된다. 단계 508에서 상기 빌드된 데이터 마이닝 모델은 사용자에게 리턴된다. 이미 저장된 데이터 마이닝 모델이 가장 낮은 우선 순위를 갖는 경우, 단계 603에서 이 모델은 상기 캐시로부터 제거된다. 다시, 단계 604에서 상기 데이터 마이닝 시스템은 상기 빌드된 데이터 마이닝 모델의 크기 및 상기 이미 저장된 데이터 마이닝 모델들의 크기들의 합이 상기 미리 정의된 스토리지 한도를 초과하는지를 테스트한다. 만약, 예라면, 처리는 단계 602에서 이미 저장된 데이터 마이닝 모델이 가장 낮은 우선 순위를 갖는지를 테스트하기 위해 계속된다. 만약 상기 스토리지 한도가 더 이상 초과되지 않으면, 상기 처리는 상기 캐시에 상기 빌드된 데이터 마이닝 모델을 저장하기 위해, 5에서 단계 507로 연결 엘리먼트(4)를 통해 진행된다.

[0055] 도 5의 단계 502에서, 데이터 마이닝 작업이 상기 데이터 마이닝 요청에 동등하거나 상기 데이터 마이닝 요청보다 더 일반적인 상기 캐시에서 발견되는 경우, 상기 데이터 마이닝 시스템은, 도 7에서 연결 엘리먼트 1을 통한 처리를 계속한다. 상기 데이터 마이닝 시스템은 단계 701에서 상기 연관된 데이터 마이닝 모델의 데이터 마이닝 모델 품질이 최소 품질 스레쉬홀드보다 낮은지를 테스트한다. 만약 상기 품질이 너무 낮으면, 단계 704에서 상기 데이터 마이닝 시스템은 사용자에게 경고를 이슈하고, 데이터 마이닝이 진행되는 것을 확인하도록 상기 사용

자에게 요청한다. 단계 705에서, 상기 사용자가 상기 모델을 사용하지 않을 것이라고 결정할 수 있는데, 그러면 상기 데이터 마이닝 요청의 처리는 종료된다. 단계 705에서 상기 사용자가 데이터 마이닝을 진행하도록 확인하면, 상기 데이터 마이닝 시스템은 연결 엘리먼트(2)를 통해 계속된다. 도 5에서 다시, 단계 504에서, 상기 데이터 마이닝 시스템은 단계 502에서 결정된 데이터 마이닝 태스크와 연관된 데이터 마이닝 모델을 빌드한다. 단계 701에서 상기 데이터 마이닝 품질이 상기 최소 품질 스톱워치보다 더 높은 경우, 단계 702에서 상기 데이터 마이닝 시스템은 상기 연관된 데이터 마이닝 모델이 상기 캐시에 저장되는지를 테스트한다. 만약 아니라면, 상기 데이터 마이닝 시스템은 도 5의 단계 504에서 연결 엘리먼트 2를 통해 계속되고, 상기 연관된 데이터 마이닝 모델을 빌드한다. 만약 상기 연관된 데이터 마이닝 모델이 상기 캐시에 저장되면, 단계 703에서 상기 데이터 마이닝 시스템은 상기 연관된 데이터 마이닝 모델 상에 후-처리를 수행할 수 있다. 대응하는 데이터 마이닝 작업이 상기 데이터 마이닝 요청에 동등한 경우, 어떠한 후-처리도 요구되지 않을 수 있으며, 상기 캐시된 데이터 마이닝 모델은 도 5의 단계 508에서 상기 사용자에게 직접적으로 리턴될 수 있다. 상기 대응하는 데이터 마이닝 작업이 상기 데이터 마이닝 요청보다 더 일반적인 경우, 서브-모델(sub-model)은 상기 데이터 마이닝 요청에 따라 상기 캐시된 데이터 마이닝 모델로부터 생성될 수 있다. 연결 엘리먼트(5)를 통해, 이 서브-모델은 도 5의 단계 508에서 상기 사용자에게 리턴될 수 있다.

[0056] 상기 서브-모델의 생성은 데이터 마이닝 유형에 의존할 수 있다. 클러스터링 데이터 마이닝의 경우, 클러스터링의 수는 캐시된 데이터 마이닝 모델의 기존의 클러스터들의 종합(aggregation)에 의해 감소될 수 있다. 연관 규칙 마이닝(association rule mining)을 위해, 상세한 예가 이하에서 기술된다.

[0057] 데이터 마이닝이 수행된 데이터 집합이 근본적으로(essentially) 변경되면, 상기 변경된 데이터 집합에 기초한 모든 데이터 마이닝 모델들은 상기 캐시로부터 제거된다. 그러나, 접근 카운트 및 품질에 관한 정보는 여전히 상기 캐시에 유지될 수 있다. 이들 특성들은 상기 선택된 데이터 자체에 적용된 데이터 마이닝 방법의 성질에 의존할 것으로 예상된다. 그러나, 상기 데이터 마이닝 모델의 빌드 시간 및 스토리지 크기는 고려되는 데이터의 양에 따라 증가해야 한다.

[0058] 도 8 내지 10은 연관 규칙 데이터 마이닝 모델들의 저장을 위한 샘플 시나리오를 도시한다. 연관 규칙들은 트랜잭션들 내에서 어느 아이템들이 자주 발생하는지를 기술하는 패턴들이다. 아이템들의 집합(set) $I = \{I_1, I_2, \dots, I_m\}$ 를 고려하자. D 는 트랜잭션들의 집합이라 하고, 여기서 각각의 트랜잭션 T 는 I 에 속하는 아이템들의 부분집합(subset)이다. 트랜잭션 T 는 I 에 아이템들의 부분집합 $A = \{I_1, I_2, \dots, I_p\}$ 를 포함할 수 있고, 또한 I 에 아이템들의 추가 부분집합 $B = \{I_p, I_{p+1}, \dots, I_q\}$ 를 포함할 수 있으며, 부분집합 A 및 부분집합 B 는 서로소(disjunct)이다. 즉 어떠한 공통 원소들도 갖지 않는다. 연관 규칙은 부분집합 A 와 부분집합 B 간의 형태 $A \rightarrow B$, 즉 $\{I_1, I_2, \dots, I_p\} \rightarrow \{I_p, I_{p+1}, \dots, I_q\}$ 의 함축된 의미(implication)이다. 아이템들의 부분집합 A 는 바디(body)로 일컬어지고, 아이템들의 부분집합 B 는 상기 규칙의 헤드(head)로 일컬어진다. 모든 가능한 규칙들의 집합으로부터 흥미있는 규칙들(interesting rules)을 선택하기 위해, 중요하고 흥미있는(of significance and interest) 여러 가지 조치에 관한 제한들이 사용될 수 있다. 가장 잘 알려진 제한들은 지지도(support) 및 신뢰도(confidence)에 관한 최소 스톱워치들이다. 아이템집합 A 의 지지도(support) $\text{supp}(A)$ 는 아이템집합 A 를 포함하는 데이터 집합에서 트랜잭션들의 부분들로서 정의된다. 규칙 $A \rightarrow B$ 의 신뢰도는 $\text{conf}(A \rightarrow B) = \text{supp}(S)/\text{supp}(A)$ 로 정의되는데, 여기서 S 는 트랜잭션 집합 D 에서 트랜잭션 T 에서 발생하는 부분집합 A 및 부분집합 B 의 아이템들의 합집합(union set)이다. 이것은, 만약 아이템들의 부분집합 A 를 포함하는 D 에서 트랜잭션의 $c\%$ 가 또한 아이템들의 부분집합 B 를 포함한다면, 연관 규칙 $A \rightarrow B$ 가 신뢰도(c)를 갖는 트랜잭션 집합 D 에서 참(true)을 유지한다는 것을 의미한다. 바꿔 말하면, 상기 신뢰는 조건부 확률(conditional probability), $P(B/A)$, 즉 동일한 트랜잭션 T 에서 부분집합 A 가 발생한다는 가정하에, 트랜잭션 T 에서 발생하는 부분집합 B 의 확률이다. 매우 드물거나 루스하게(loosely) 상관된 이벤트들은 일부 어플리케이션들에게는 중요하지 않을 수도 있으므로, 사용자는 상기 규칙들을 위한 최소 지지도 또는 신뢰도를 정의할 수 있다.

[0059] 다음의 예는 자동차 회사를 기술한다. 상기 회사는 큰 라이센스된 딜러 네트워크(large licensed dealer network)를 갖는다. 상기 딜러들은 또한 리페어 샵 서비스들(repair shop services)을 제공한다. 판매(sales) 및 결함들(defects)에 관한 데이터는 상기 자동차 회사에 규칙적으로(regularly) 병합되고 전달된다. 서로 다른 작업 역할들을 갖는 서로 다른 사람들은 규칙적으로 상기 데이터를 갖고서 작업하고 분석을 수행한다. 이러한 작업 역할들 중에서 사람들은 상기 딜러들을 지지하거나 품질 표준들을 보장하고 있다. 이들 종업원들은 상기 데이터를 포함하는 중앙 데이터 웨어하우스에 대해(against) 분석 요청들을 실행할 수 있다. 종종 종업원들은 동일하거나 유사한 분석을 요청한다. 그러므로, 분석 결과들의 캐싱은 응답 횟수들을 상당히 감소시킬 수 있고,

따라서 생산성을 증가시킬 수 있다.

- [0060] 구현 기준, 즉 상기 데이터 마이닝 모델이 상기 캐시에 기록되는 경우 결정 절차는 상기 캐시 스토리지 한도, 상기 데이터 마이닝 모델 품질, 및 상기 데이터 마이닝 모델의 빌드 시간에 기초하며, 이는 위에서 정의되었다. 상기 데이터 마이닝 시스템을 위해 상기 시스템 관리자는 다음의 파라미터들, 즉, CacheStorageLimit = 9MB, MinimumQuality = 0.5, 및 MinimumBuildTime = 10분을 명시했다.
- [0061] 도 8은 데이터 마이닝 요청들(R1A, R1B, R2, R3A, 및 R3B)의 사용자 요청 테이블(810)을 보여준다. 작업 기술 테이블(task description table)(820)은 데이터 마이닝 작업들(T0, T1, T2, 및 T3)을 포함하는데, 이것들은 처리 데이터 마이닝 요청들의 경우 상기 캐시에 저장된다. 상기 데이터 마이닝 작업들은 연관된 데이터 마이닝 모델들(M0, M1, M2, 및 M3)을 갖는다. 그것들의 특징들은 테이블 830에 저장되고, 상기 캐시에 저장될 데이터 마이닝 모델들을 선택하기 위한 기준이다. 사용자 요청들의 시퀀스에 응답하여, 테이블 820에서 데이터 마이닝 작업들(T0, T1, T2, 및 T3)의 기술들(descriptions)은, 테이블 830에서 연관된 데이터 마이닝 모델들(M0, M1, M2, 및 M3)의 특징들과 함께 상기 캐시에 계속해서 저장된다.
- [0062] 도 9에서, 테이블 910은 상기 샘플 시나리오의 여러 가지 단계들에서 상기 데이터 마이닝 모델들을 위한 구현 상태들을 보여준다. 각각의 데이터 마이닝 모델의 구현 상태는 상기 구현 상태가 아직 결정되지 않았는지(MATERIALIZE="-"), 상기 데이터 마이닝 모델이 이용가능한 스토리지 공간에 의존하여 상기 캐시에 저장될 수 있는지(MATERIALIZE="YES" 또는 "MAYBE"), 또는 상기 데이터 마이닝 모델이 상기 캐시에 저장되지 않을 수 있는지(MATERIALIZE="NO")를 나타낸다.
- [0063] 샘플 시나리오는 단지 상기 캐시 내의 연관된 데이터 마이닝 모델 M0 및 데이터 마이닝 작업 기술 T0로 시작한다. 상기 데이터 선택은 필터 조건 Filter=all을 갖고서 상기 커스터머들 테이블 CUST 상에서 수행된다. 상기 데이터 마이닝 제한은 상기 데이터 마이닝 모델에서 클러스터들의 최대 수는 MaxNum=5이다. 클러스터링 데이터 마이닝 모델 M0의 특징들은 다음의 특징들, 즉, 상기 데이터 마이닝 모델 품질은 0.9 또는 90%인 것을 포함한다. 상기 데이터 마이닝 모델의 빌드 시간 - 이는 상기 데이터 마이닝 모델 복잡성의 측정점 - 은 분, 예를 들어 20분으로 명시된다. 상기 접근 카운트는 상기 데이터 마이닝 모델이 사용자 요청들에 의해 몇 번이나 히트(hit)되는지(예를 들어, 5번)를 명시한다. 상기 접근 카운트는 또한 절대적 요청 빈도로 고려될 수 있다. 상기 크기는 상기 데이터 마이닝 모델을 저장하기 위해 상기 캐시에서 사용되는 스토리지를 명시한다. 예를 들어, 그것은 5MB 또는 5메가바이트이다. 상기 우선순위는 상기 데이터 마이닝 모델 크기와 상기 데이터 마이닝 모델을 위한 사용자 요청들의 접근 카운트의 곱이 "25"가 되도록 계산된다. 상기 우선 순위는 바람직하게는 통일된 측정 유닛을 갖는다. 이와는 다르게, 상기 데이터 마이닝 모델의 접근 카운트는 사용자 요청들의 총 수에 의해 분할될 수 있는데, 이는 결국 상대적 접근 빈도 및 변경된 우선순위를 초래한다. 테이블 910의 단계 911에서, 상기 데이터 마이닝 모델 M0의 구현 상태는 MATERIALIZE=YES로 설정된다.
- [0064] 전달 지지 팀 멤버는 커스터머 주문들에서 함께 발생하는 빈번한 결함 부분들을 발견하기 위해 뮌헨 지역에서 그의 딜러들에 의해 대체되는 결함 부분들을 분석하기를 원한다. 이것은 전달을 위해 함께 부분들을 그가 번들(bundle)하는 것을 돕는다. 그러므로, 그는 뮌헨에서 딜러들을 위해 상기 결함 부분 테이블 DEFECTS에 대해(against) 연관 규칙 마이닝 요청 R1A을 보내고, 10%의 최소 지지도를 명시한다. 상기 캐시에는 동일하거나 더 일반적인 작업 기술(task description)이 없으므로, 작업 기술 T1은 상기 캐시에 저장된다. 그런 다음, 데이터 마이닝 모델 M1은 Filter=Munich를 갖고서 상기 테이블 DEFECTS로부터의 데이터에 기초하여 구성된다.
- [0065] 도 10은 가능한 데이터 마이닝 모델 M1을 보여주는데, 10%보다 더 높은 지지도를 갖는 테이블 1010에서 세 개의 규칙들이 발견되는데, 예를 들어, 라인 1011에서 규칙 1은 차(car)의 두 개의 결함 부분들, 실린더 헤드 가스켓 및 이그니션 플러스 사이의 연관성을 기술한다. 이 규칙은 24%의 지지도 및 68%의 신뢰도를 갖는다. 라인들 1012 및 1013에서 나머지 규칙들은 결함 부분들의 다른 조합들을 상관시킨다. 상기 데이터 마이닝 모델 품질은 발견되는 규칙들의 신뢰도 평균으로서, 라인 1014에 도시된 바와 같이, 즉 76%로 계산된다. 상기 데이터 마이닝 모델 M1을 위해, 이러한 품질 및 15분의 빌드 시간은 상기 시스템 관리자에 의해 정의된 스톱워치들보다 더 크다. 그러므로, 상기 데이터 마이닝 모델은 상기 캐시에 임시로 저장된다. 모델 M0의 5 메가바이트 및 모델 M1의 2 메가바이트는 9 메가바이트의 캐시 스토리지 한도를 초과하지 않으므로, 어떠한 데이터 마이닝 모델들도 상기 캐시로부터 삭제될 필요가 없다. 단계 912에서, 모델 M1의 구현 상태는 MATERIALIZE=YES로 설정된다. 데이터 마이닝 모델 M1을 위한 접근 카운트는 "1"로 설정되고, 상기 우선 순위는 "2"로 계산된다. 데이터 마이닝 모델 M1은 상기 지지 팀 멤버에게 리턴된다. 상기 규칙들로부터의 이해(insight)는 그가 상기 딜러들에게 연관된 여분의(spare) 부분들의 유용한 조합들을 제공하도록 돕는다.

- [0066] 품질 보장 팀 멤버는, 동일한 결합 부분들 데이터에 기초하지만, 다른 목적들을 위해 유사한 분석을 수행할 필요가 있다. 특정 부분에서 결합이 또 다른 부분들이 고장나도록 하는지를 조사하기 위해, 그는 여분의 부분 오더들(orders)에서 함께 발생하는 결합 부분들에 흥미를 갖는다. 그는 대부분의 결합들을 초래하는 루트 부분들(root parts)을 인식할 것을 예상하고, 그래서 그것들은 조사될 수 있고 가능하다면 재설계(redesign)될 수 있다. 이러한 목적을 위해, 그는 또한 뮌헨 딜러들을 위해 상기 결합 부분들 테이블 DEFECTS에 대해(against) 연관성 규칙 데이터 마이닝 요청 R1B를 보내고 최소 지지도를 MinSupp = 20%로 명시한다. 이 데이터 마이닝 요청 R1B는 데이터 마이닝 작업 기술 T1에 매핑되는데, 이는 최소 지지도 MinSupp = 10%를 갖고서 상기 캐시에 이미 존재하고, 그래서 상기 데이터 마이닝 요청보다 더 일반적이다. 상기 캐시에 저장된 데이터 마이닝 모델 M1은 데이터 마이닝 작업 T1과 연관된다. 요청 R1B를 위해, 상기 데이터 마이닝 시스템은 데이터 마이닝 모델 M1로부터 서브-모델을 추출할 수 있다. 상기 서브-모델은 최소 지지도가 20%인 모든 규칙들을 포함한다. 즉, 라인들 1011 및 1012에서 단지 규칙 1 및 규칙 2가 그것이다. 라인 1013에서 규칙 3의 지지도는 20% 아래이고 서브-모델로부터 배제된다. 데이터 마이닝 모델 M1의 접근 카운트는 "2"로 증가되고, 상기 우선 순위 값은 "4"가 되도록 재계산된다. 상기 결정된 서브-모델은 상기 요청하는 사용자에게 리턴되는데, 상기 요청하는 사용자는 더 짧은 응답 시간으로부터 혜택을 보며, 상기 결과들의 조사로 당장 시작할 수 있다. 그는 동료들에게 실린더 헤드 가스켓 및 스티어링 샤프트에 관한 문제들을 조사하도록 제안하여, 상기 이그니션 플러그 및 스티어링 스쿠기어에서 각각의 문제들을 방지하도록 할 수 있다.
- [0067] 추가 지지 팀 멤버는 슈투트가르트 지역의 딜러들을 책임진다. 그는 결합 부분들에 관해 유사한 분석을 수행하기를 원하지만, 필터 조건은 Filter=Stuttgart를 가지며, 대응하는 데이터 마이닝 요청 R2를 보낸다. 비록 테이블 DEFECTS이 데이터 마이닝 작업 T1에 대해서는 동일하지만, 상기 필터 조건은 다르며 저장된 데이터 마이닝 모델 M1은 이 경우에 재-사용될 수 없다.
- [0068] 그러므로, 상기 캐시에 새로운 데이터 마이닝 작업 기술 T2이 저장될 수 있고, 연관된 데이터 마이닝 모델 M2이 빌드된다. M2의 데이터 마이닝 모델 품질은 40%이며, 상기 시스템 관리자에 의해 정의된 50%의 최소 품질보다 더 낮으며, 모델 M2은 캐시되지 않는다. 모델 M2의 구현 상태는 단계 913에서 MATERIALIZE=NO로 설정된다. 결정된 데이터 마이닝 모델 M2은 추가 사용을 위해 저장되지 않으며, 단지 상기 요청하는 사용자에게 리턴된다. 상기 데이터 마이닝 모델이 저장되지 않는 경우, 데이터 마이닝 작업 기술 T2은 상기 캐시에 전혀 저장되지 않거나 상기 캐시로부터 제거되는 것이 가능하다.
- [0069] 추가 품질 보증 팀 멤버는 뮌헨 지역에서 상기 딜러들을 위해 이미 했던 그의 동료와 같이 베를린 지역 딜러들을 위한 분석 R3A를 수행하기를 원한다. 새로운 데이터 마이닝 작업 정의 T3은 상기 캐시에 저장되고 데이터 마이닝 모델 M3이 구성된다. 상기 품질 및 상기 응답 시간 둘 다는 상기 시스템 관리자에 의해 저장되는 스투츨름드들보다 더 크므로, 상기 데이터 마이닝 모델 M3은 상기 캐시에 저장될 것으로 계획된다. 그러나, 데이터 마이닝 모델들(M1, M2, 및 M3)의 크기들의 합(즉, 10MB)은 상기 최대 스토리지 한도(즉, 9MB)를 초과한다. 이 경우, 상기 캐시에서 모든 모델들의 우선 순위들이 비교된다(즉, M0를 위해 25, M1을 위해 4, M3을 위해 3). 상기 크기들의 합이 상기 최대 스토리지 한도 아래에 들어올 때까지 가장 낮은 우선순위들을 갖는 모델들은 상기 캐시로부터 제거된다. 이 경우, 데이터 마이닝 모델 M3은 상기 캐시에 저장되지 않고, 그것의 구현 상태는 단계 914에서 MATERIALIZE=MAYBE로 갱신된다. 대응하는 데이터 마이닝 작업 기술 T3 및 데이터 마이닝 모델 M3의 특성들은 그러나 상기 캐시에 남아 있을 수 있다. 결국, 상기 결정된 연관성 규칙들 데이터 마이닝 모델(M3)은 상기 요청하는 사용자에게 리턴된다.
- [0070] 또 다른 추가 품질 보장 팀 멤버는 또한 상기 베를린 지역에서 딜러들을 위해 결합 부분들에 대해 유사한 분석 R3B를 수행하기를 원할 수 있지만, 더 높은 최소 지지도 MinSupp=20%를 가질 수 있다. 이 요청을 위해, 더 일반적인 데이터 작업 기술 T3은 MinSupp=10%를 갖고서 상기 캐시에서 발견될 수 있다. 이미 존재하는 데이터 마이닝 작업 기술 T3은 요청 R3B를 위해 사용된다. 상기 저장된 데이터 마이닝 품질 및 빌드 시간은 상기 캐시에서 데이터 마이닝 모델 M3을 저장하는 것을 가능하게 하기 위해 충분히 높다. 그러나 연관된 데이터 마이닝 모델 M3은 상기 캐시에 존재하지 않으며 재구성될 필요가 있다. 데이터 마이닝 모델 M3의 품질 및 빌드 시간은 통상적으로 변경되지 않고 남는다. 그것의 접근 카운트는 "2"로 증가되고 그것의 우선 순위는 "6"으로 재계산된다. 10MB는 9MB보다 커서 상기 스토리지 한도는 다시 초과될 것이므로, 가장 낮은 우선 순위들을 갖는 데이터 마이닝 모델들은 상기 캐시로부터 제거된다. 이 경우, 우선 순위 "4"를 갖는 모델 M1은 상기 캐시로부터 제거되는 반면, 우선 순위 "6"을 갖는 모델 M3은 상기 캐시에 저장된다. 단계 915에서, 데이터 마이닝 모델 M1을 위한 상기 구현 상태는 MATERIALIZE=MAYBE로 설정되고, 데이터 마이닝 모델 M3을 위해 상기 구현 상태는 MATERIALIZE=YES로 설정된다. 결국, 모델 M3은 상기 요청하는 사용자에게 리턴된다.

- [0071] 도 11은 데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템(1100)의 블록도를 도시한다. 상기 데이터 처리 시스템은 데이터 처리 수단(1110), 및 데이터 마이닝 모델들(1121, 1126)의 저장을 위한 스토리지 수단(1120)을 포함한다. 상기 데이터 처리 수단은 프로세서(1111) 및 메모리(1112)를 갖는다. 상기 프로세서는 상기 메모리의 각각의 엘리먼트들에 프로그램 코드 부분들(1113, ..., 1117)을 저장하고 다음과 같이 이들 프로그램 코드부분들을 실행한다. 즉, 제1 프로그램 코드 부분들(1113)에 따라, 상기 프로세서는 다음의 데이터 마이닝 모델 특징들, 즉 품질 및 복잡성 중 적어도 하나를 갖는 제1 데이터 마이닝 모델(1118)을 생성한다. 제2 프로그램 코드 부분들(1114)에 따라, 상기 프로세서는 만약 상기 특징들을 위한 기준이 충족되면, 상기 스토리지 수단에 저장하기 위한 후보로서 상기 제1 데이터 마이닝 모델을 핸들링한다. 제3 프로그램 코드 부분들(1115)에 따라, 상기 프로세서는 제1 데이터 마이닝 모델(1118)의 크기 및 스토리지 수단(1122, ..., 1123)에 이미 저장된 추가 데이터 마이닝 모델들의 크기들의 합이 스토리지 한도를 초과하는지를 결정한다. 제4 프로그램 코드 부분들(1116)에 따라, 만약 상기 스토리지 한도가 초과되지 않으면, 상기 데이터 처리 수단은 상기 스토리지 수단에 제1 데이터 마이닝 모델(1121)을 저장한다. 제5 프로그램 코드 부분들(1117)에 따라, 만약 상기 스토리지 한도가 초과되면, 상기 데이터 처리 수단은 상기 제1 데이터 마이닝 모델 및 상기 추가 데이터 마이닝 모델들의 우선 순위들에 기초하여 어느 데이터 마이닝 모델들을 상기 스토리지 수단에 저장할 것인지를 결정한다. 상기 우선 순위들은 적어도 상기 각각의 데이터 마이닝 모델들의 접근 빈도들에 의존한다.
- [0072] 도 12는 데이터 마이닝 모델들의 저장을 위한 데이터 처리 시스템(1200)의 블록도를 도시한다. 상기 데이터 처리 시스템은 입력 수단(1210), 스토리지 수단(1220), 데이터 처리 수단(1230), 및 출력 수단(1240)을 포함한다. 상기 입력 수단은 새로운 데이터 마이닝 요청을 수신한다. 상기 스토리지 수단은 데이터 마이닝 모델들 정보(1221, ..., 1224)를 저장하며, 데이터 마이닝 모델 및 정보(1225, ..., 1228)를 식별하며, 데이터 마이닝 모델 품질을 기술한다. 상기 데이터 처리 수단은 프로세서(1231) 및 메모리(1232)를 갖는다. 상기 프로세서는 상기 메모리의 각각의 엘리먼트들에 프로그램 코드 부분들(1233, ..., 1236)을 저장하며, 이들 프로그램 코드 부분들을 다음과 같이 실행한다. 즉, 제1 프로그램 코드 부분들(1233)에 따라, 상기 프로세서는 상기 스토리지 수단에 데이터 마이닝 모델을 식별하는 데이터 마이닝 모델들 정보 및 데이터 마이닝 모델 품질을 기술하는 정보를 저장한다. 제2 프로그램 코드 부분들(1234)에 따라, 상기 프로세서는 상기 스토리지 수단에 데이터 마이닝 모델들(1229, ..., 1230)을 저장하는데, 이는 데이터 마이닝 모델 품질(1227, ..., 1228)을 가지며, 제1의 미리 정의된 기준을 충족시킨다. 제3 프로그램 코드 부분들(1235)에 따라, 상기 프로세서는 상기 새로운 데이터 마이닝 요청을 위해 데이터 마이닝 모델을 결정하고, 상기 데이터 마이닝 모델을 위해 데이터 마이닝 모델 품질을 기술하는 정보(1225, ..., 1228)가 상기 스토리지 수단에 저장되었는지를 체크한다. 제4 프로그램 코드 부분들(1236)에 따라, 상기 프로세서는 만약 상기 데이터 마이닝 모델의 데이터 마이닝 모델 품질이 제2의 미리 정의된 기준을 충족시키지 않으면 데이터 마이닝이 진행되는 것을 확인하도록 사용자에게 요청을 개시한다. 출력 수단은 상기 사용자에게 상기 요청을 보낸다.
- [0073] 본 발명의 다른 실시예들에서, 상기 데이터 마이닝 모델들을 위한 다른 품질 기준들이 명시될 수 있다. 연관성 규칙 데이터 마이닝에서, 흥미있는 다른 수단은, 예를 들어, "확신(conviction)", "리프트(lift)", 및 다른 것들과 같이 데이터 마이닝 모델들의 품질들을 정의하기 위해 사용될 수 있다. 더 상세한 설명을 위해, URL http://en.wikipedia.org/wiki/Association_rule_learning을 참조할 수 있다. 규칙 $A \rightarrow B$ 의 "확신"은 부분집합 A 가 부분집합 B 없이 발생하는 예상 빈도의 비율로 해석될 수 있다. 규칙의 "리프트"는 우연히 예상되는 관찰되는 신뢰도의 비율로서 정의된다. 추가적인 제한들이 상기 규칙 선택에 적용될 수 있는데, 이는 품질 정의, 예를 들어, 규칙들의 가중치들(weights)(비용, 가격), 규칙 길이의 제한들, 및 양 또는 음의 아이템 제한들에 포함될 수 있다. 규칙 $A \rightarrow B$ 의 "length"는 서브세트들 A 및 B 에서 아이템들의 수에 의존한다.
- [0074] 데이터 마이닝 모델들의 품질은 통상적으로 비교적인 수 값들, 예를 들어, 낮은 품질에 대해서는 0.1 그리고 높은 품질을 위해서는 0.9로서 표현된다. 이와는 다르게, 분류된 리스트의 스트링 엘리먼트들이 대신에 사용될 수 있는데(예를 들어, ["bad", "medium", "good"]), 여기서 상기 리스트 위치는 품질의 측정을 나타낸다.
- [0075] 데이터 마이닝 모델 복잡성은 모델 빌드 시간 이외의 다른 파라미터들, 예를 들어, 데이터 마이닝 모델 아이템들의 수에 기초할 수 있다.
- [0076] 상기 구현 기준은 데이터 마이닝 모델 품질 및 데이터 마이닝 모델 복잡성에 한정되지 않고, 데이터 마이닝 모델들의 추가 특성들을 고려할 수 있다.
- [0077] 상기 우선 순위 계산은 상기 데이터 마이닝 모델의 접근 카운트와 스토리지 크기의 곱으로 제한되지 않는다. 상기 네 개의 데이터 마이닝 모델 특징들(크기, 품질, 접근 카운트, 빌드 시간)의 다른 조합들이 또한 고려될 수

있으며, 또는 심지어 다른 모델 특성들이 고려될 수도 있다. 이는 스토리지 소비 이외의 다른 데이터 마이닝 모델 특성들에 기초하여 계산되는 데이터 마이닝 모델 크기일 수 있다.

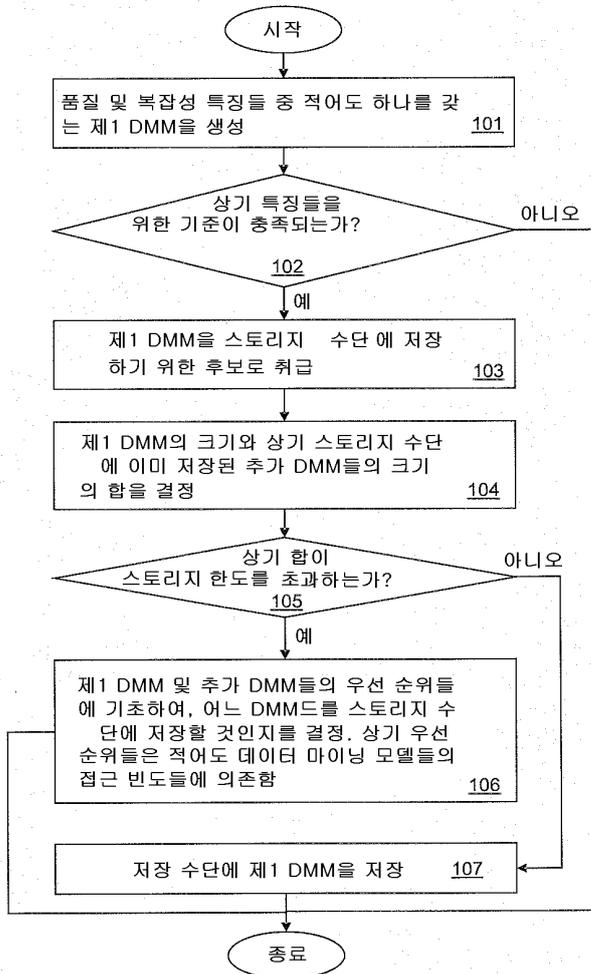
- [0078] 상기 구현 기준을 위한 식들(formulas) 및 상기 우선 순위 계산은 또한 서로 다른 데이터 마이닝 요청 패턴들에 대해 동적으로 변경될 수 있다. 일부 시나리오들에서, 모델 응답 시간은 모델 품질보다 더 중요할 수 있다. 다른 시나리오들은 다른 사용자 요구들에 중점을 둘 수 있다.
- [0079] 시퀀스들 데이터 마이닝은 아이템들의 규칙들을 위한 지지도 및 신뢰도 측정들을 사용하는데, 이는 연관성 규칙 마이닝과 유사한 방법으로 품질 값들을 정의하도록 허용한다.
- [0080] 상기 데이터 마이닝 요청들의 데이터 마이닝 작업들에 대한 전사 매핑, 데이터 마이닝 작업들의 데이터 마이닝 모델들에 대한 전단사 매핑, 상기 사용자 다이얼로그들, 및 상기 연관된 데이터 마이닝 모델들의 구성은 동일한 데이터 처리 시스템에서 수행될 수 있다. 이와는 다르게, 상기 단계들의 일부는, 예를 들어, 서로 다른 데이터 처리 시스템들에서 수행될 수 있는데, 이는 예를 들어, 네트워크를 통한 전송에 의해 발생된 정보의 전달을 요구한다.
- [0081] 여기에 사용되는 용어는 단지 특정 실시예들을 기술하려는 목적으로 사용된 것이지 발명의 범위를 한정하려고 사용된 것은 아니다. 여기에 사용되는 바와 같이, 단수 형태들인 "하나의", "한" 및 "일" 등은, 그 맥락에서 분명하게 그렇지 않은 것으로 나타내지 않는다면 복수 형태들도 포함하도록 의도된다. 나아가 "포함하다" 및/또는 "포함하는" 이라는 용어들은 본 명세서에서 사용되는 경우, 언급된 특징들, 정수들, 단계들, 동작들, 엘리먼트들, 및/또는 컴포넌트들을 특정하지만, 하나 또는 그 이상의 특징들, 정수들, 단계들, 동작들, 엘리먼트들, 컴포넌트들, 및/또는 이것들의 그룹들을 배제하려는 의도가 아님을 이해해야 할 것이다.
- [0082] 청구항들에서 대응하는 구조들, 재료들, 작용들, 및 모든 수단 또는 단계 플러스 기능 엘리먼트들(step plus function elements)의 균등물들은 이하에서, 구체적으로 청구되는 다른 청구된 엘리먼트들과 함께 그 기능을 수행하기 위한, 어떤 구조, 재료, 또는 작용이든 포함하도록 의도된다. 본 발명에 관한 기술은 예시 및 설명의 목적으로 제시되지만, 개시되는 그 형태대로 발명을 한정하거나 실시예들을 총 망라하려는 의도는 아니다. 발명의 범위 및 사상을 벗어남이 없이 많은 변경 및 변형 예들이 있을 수 있다는 것은 당해 기술 분야에서 통상의 기술을 가진 자들에게는 자명할 것이다. 상기 실시예는 발명의 원리들 및 실제 어플리케이션을 가장 잘 설명하기 위해, 그리고 고려된 특정 사용에 적합하게 당해 기술 분야에서 통상의 기술을 가진 자들이 여러 가지 변형 예들을 갖는 다양한 실시예들에 대해 발명을 이해할 수 있도록 하기 위해, 선택되고 기술되었다.
- [0083] 당해 기술 분야에서 숙련된 자라면 알 수 있는 바와 같이, 본 발명의 측면들은 시스템, 방법, 또는 컴퓨터 프로그램 제품으로 구현될 수 있다. 따라서, 본 발명의 측면들은 전적으로 하드웨어 실시예의 형태를 취할 수도 있고, 전적으로 소프트웨어 실시예(펌웨어(firmware), 상주 소프트웨어(resident software), 마이크로-코드(micro-code) 등을 포함함)의 형태를 취할 수도 있고, 또는 소프트웨어 및 하드웨어 측면들을 조합하는 실시예의 형태 - 이것들 모두는 본 명세서에서는 일반적으로 "회로", "모듈" 또는 "시스템"으로 일컬어질 수 있음 - 를 취할 수 있다. 더 나아가, 본 발명의 측면들은 하나 또는 그 이상의 컴퓨터 판독가능 매체(들) - 이것들은 그 내부에 구현된 컴퓨터 판독가능 프로그램 코드를 가짐 - 에 구현된 컴퓨터 프로그램 제품의 형태를 취할 수도 있다.
- [0084] 하나 또는 그 이상의 컴퓨터 판독가능 매체(들)의 조합이 이용될 수 있다. 상기 컴퓨터 판독가능 매체는 컴퓨터 판독가능 신호 매체 또는 컴퓨터 판독가능 스토리지 매체일 수 있다. 상기 컴퓨터 판독가능 스토리지 매체의 더 구체적인 예들에는, 휴대용 컴퓨터 디스켓, 하드 디스크, RAM(random access memory), ROM(read-only memory), EPROM(erasable programmable read-only memory) 또는 플래쉬 메모리, 휴대용 콤팩트 디스크 읽기 전용 메모리(휴대용 CD-ROM), 광 스토리지 디바이스, 자기 스토리지 디바이스, 또는 이것들의 적절한 조합 등이 포함될 수 있으나, 이러한 예들로 한정되는 것은 아니다. 본 문서의 맥락에서, 컴퓨터 판독가능 스토리지 매체는 명령 실행 시스템, 장치, 또는 디바이스에 의해 사용하기 위한 또는 명령 실행 시스템, 장치, 또는 디바이스와 함께 사용하기 위한 프로그램을 포함, 또는 저장할 수 있는 실제적인 매체일 수 있다.
- [0085] 컴퓨터 판독가능 신호 매체는, 예를 들어, 기저대역에서 또는 반송파의 일부로서 전파되는 데이터 신호 - 상기 전파되는 데이터 신호는 그 안에 구현된 컴퓨터 판독가능 프로그램 코드를 가짐 - 를 포함할 수 있다. 이러한 전파되는 신호는 다양한 형태들을 취할 수 있는데, 그 형태의 예들에는, 전자기, 광학, 또는 이것들의 적절한 조합이 포함될 수 있으나, 이러한 예들로 한정되는 것은 아니다. 컴퓨터 판독가능 신호 매체는 명령 실행 시스템, 장치, 또는 디바이스에 의해 사용하기 위한 또는 명령 실행 시스템, 장치, 또는 디바이스와 함께 사용하기

위한 프로그램을 전달, 전파, 전송할 수 있는 컴퓨터 판독가능 스토리지 매체가 아닌 어떤 컴퓨터 판독가능 매체일 수도 있다.

- [0086] 컴퓨터 판독가능 매체 상에 구현된 프로그램 코드는, 무선, 유선, 광섬유 케이블, RF 등, 또는 이것들의 어떤 적절한 조합을 포함하는, 어떤 적절한 매체를 사용해서든 전송될 수 있다.
- [0087] 본 발명의 측면들을 위한 동작들을 수행하기 위한 컴퓨터 프로그램 코드는 하나 또는 그 이상의 프로그래밍 언어들의 조합으로 기록될 수 있는데, 이러한 예들에는, 자바, 스몰토크, C++ 등과 같은 객체 지향 프로그래밍 언어 및 "C" 프로그래밍 언어 또는 유사 프로그래밍 언어들과 같은 전통적인 절차 프로그래밍 언어들도 포함된다. 상기 프로그램 코드는, 독립형 소프트웨어 패키지에 따라 전적으로 사용자의 컴퓨터 상에서, 부분적으로 상기 사용자의 컴퓨터 상에서, 상기 사용자의 컴퓨터 상에서 부분적으로 그리고 원격 컴퓨터 상에서 부분적으로, 또는 상기 원격 컴퓨터 또는 서버 상에서 전적으로 실행될 수 있다. 후자의 시나리오에서, 상기 원격 컴퓨터는 LAN(local area network) 또는 WAN(wide area network)를 포함하는 어떤 유형의 네트워크를 통해 상기 사용자의 컴퓨터에 연결될 수 있고, 또는 상기 연결은 외부 컴퓨터에 대해 이뤄질 수 있다(예를 들어, 인터넷 서비스 제공자를 사용하여 인터넷을 통해).
- [0088] 본 발명의 측면들은 위에서 발명의 실시예들에 따른 방법들, 장치(시스템들) 및 컴퓨터 프로그램 제품들의 흐름도들 및/또는 블록도들을 참조하여 기술되었다. 흐름도들 및/또는 블록도들에서의 각각의 블록, 및 흐름도들 및/또는 블록도들에서 블록들의 조합들은 컴퓨터 프로그램 명령들에 의해 구현될 수 있다. 이들 컴퓨터 프로그램 명령들은 일반적인 목적의 컴퓨터, 특정 목적의 컴퓨터, 또는 기타 프로그램가능한 데이터 처리 장치의 프로세서에 제공되어 머신을 생성할 수 있다. 그리하여, 상기 명령들은 상기 컴퓨터 또는 기타 프로그램가능 데이터 처리 장치의 프로세서를 통해 실행될 경우, 흐름도 및/또는 블록도의 블록 또는 블록들에 명시된 기능들/작용들을 구현하기 위한 수단을 생성하도록 한다.
- [0089] 이들 컴퓨터 프로그램 명령들은 또한 컴퓨터, 기타 프로그램가능한 데이터 처리 장치, 또는 기타 디바이스들에게 특정 방법으로 기능하도록 지시하는 컴퓨터 판독가능 매체에 저장될 수 있다. 그리하여, 상기 컴퓨터 판독가능 매체에 저장된 명령들은 흐름도 및/또는 블록도의 블록 또는 블록들에 명시된 기능/작용을 구현하는 명령들을 포함하는 제조 물품을 생성하도록 한다.
- [0090] 상기 컴퓨터 프로그램 명령들은 또한 컴퓨터, 기타 프로그램가능한 데이터 처리 장치, 또는 기타 디바이스들에 로드되어, 일련의 동작 단계들이 상기 컴퓨터, 기타 프로그램가능한 장치 또는 기타 디바이스들 상에서 수행되도록 하여, 컴퓨터로 구현된 프로세스를 생성하도록 한다. 그리하여, 상기 컴퓨터 또는 기타 프로그램가능한 장치 상에서 실행되는 명령들은 흐름도 및/또는 블록도의 블록 또는 블록들에 명시된 기능들/작용들을 구현하기 위한 프로세스들을 제공한다.
- [0091] 도면들에서의 흐름도 및 블록도들은 본 발명의 여러 가지 실시예들에 따른 시스템들, 방법들 및 컴퓨터 프로그램 제품들의 가능한 구현들의 아키텍처, 기능, 및 동작을 도시한다. 이 점에서, 상기 흐름도 또는 블록도들에서의 각각의 블록은 상기 특정된 논리적 기능(들)을 구현하기 위한 하나 또는 그 이상의 실행가능한 명령들을 포함하는 모듈, 세그먼트, 또는 코드 부분을 표현할 수 있다. 또한 일부 다른 구현들에서, 상기 블록에 표시된 기능들은 도면들에서 표시된 순서를 벗어나서 발생할 수 있음을 주목하자. 예를 들어, 연속해서 보여지는 두 개의 블록들은, 관련되는 기능에 의존하여, 사실상, 실질적으로 동시에 실행될 수도 있고, 또는 상기 블록들은 때로는 역순으로 실행될 수도 있다. 또한 상기 블록도들 및/또는 흐름도의 각각의 블록, 및 상기 블록도들 및/또는 흐름도에서의 블록들의 조합들은 특정된 기능들 또는 작용들, 또는 특정 목적 하드웨어 및 컴퓨터 명령들의 조합들을 수행하는 특정 목적 하드웨어 기반의 시스템들에 의해 구현될 수 있음을 주목해야 할 것이다.
- [0092] 전술한 내용은 발명의 특정 실시예들을 참조하지만, 당해 기술 분야에서 숙련된 자들이라면 첨부된 청구항들에 의해 정의되는 발명의 범위, 발명의 사상 및 원리들을 벗어나지 않고 이들 실시예들에서의 변경들이 이뤄질 수 있다는 것을 이해할 것이다.

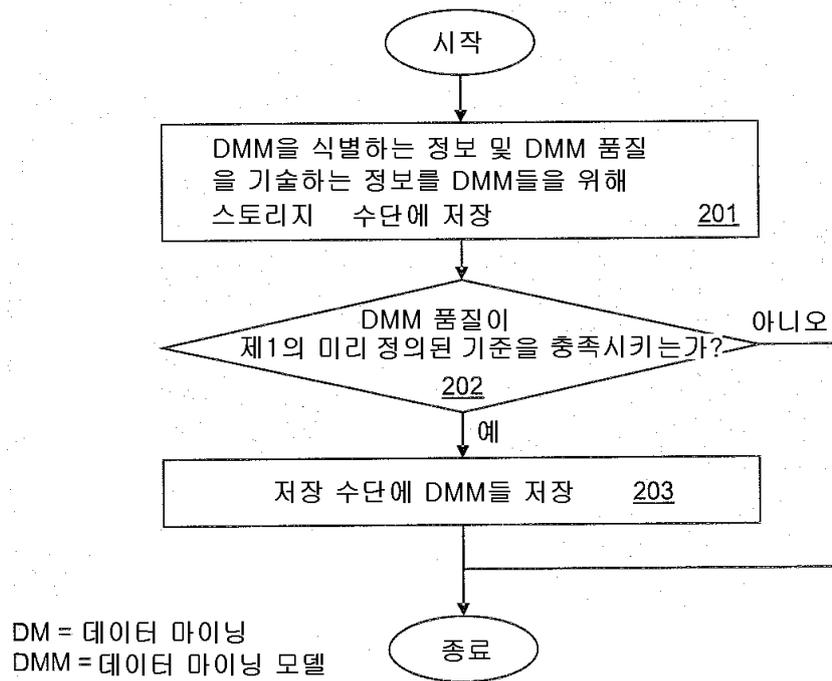
도면

도면1

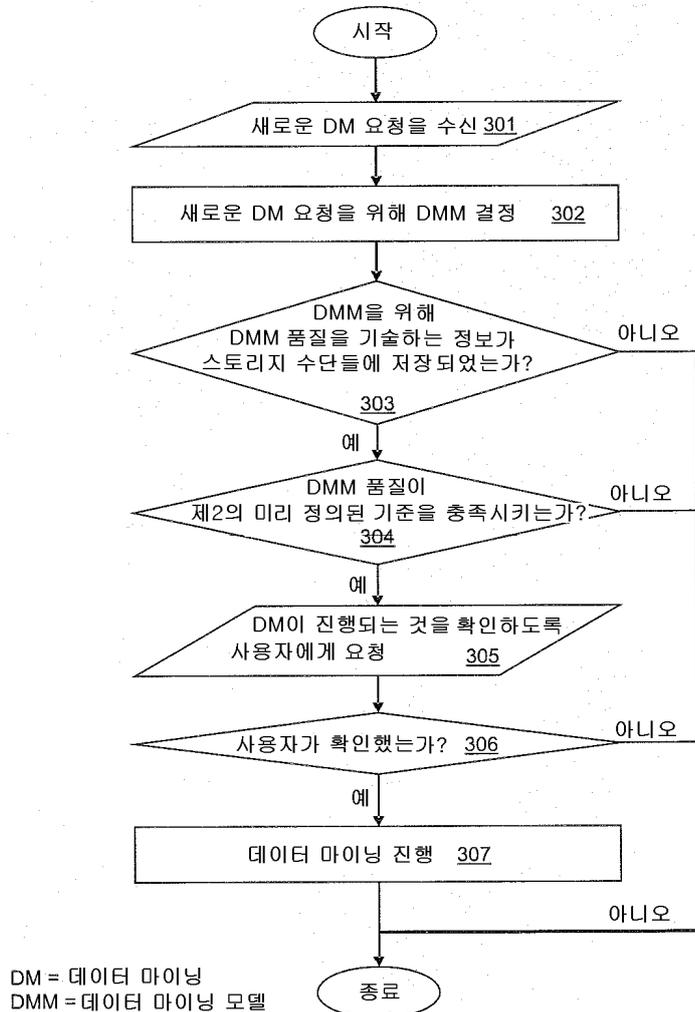


DM = 데이터 마이닝
DMM = 데이터 마이닝 모델

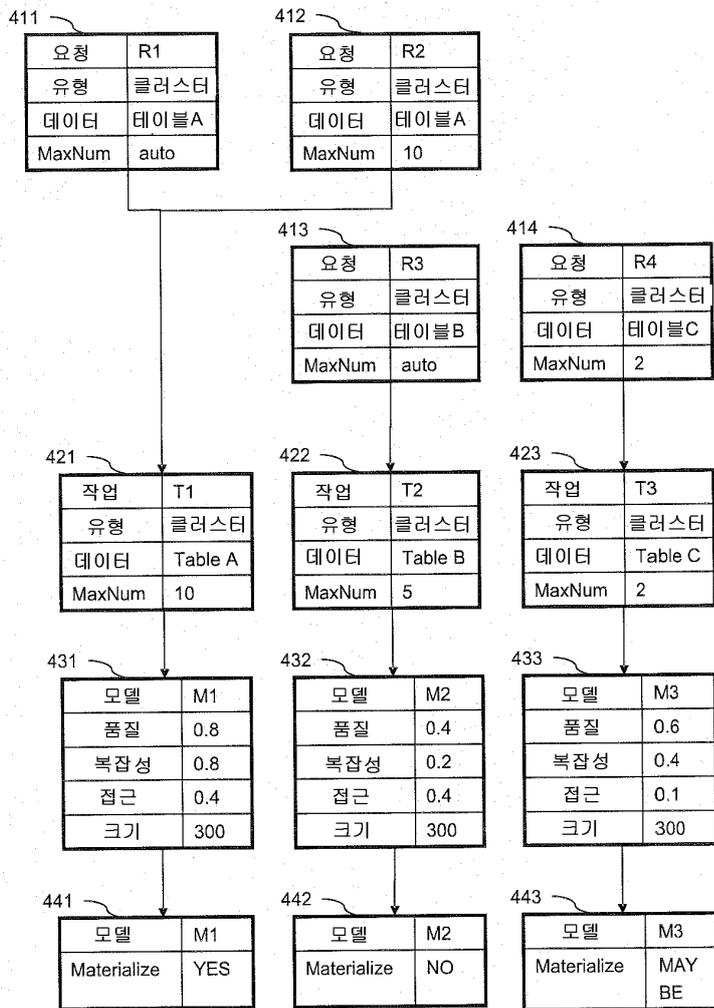
도면2



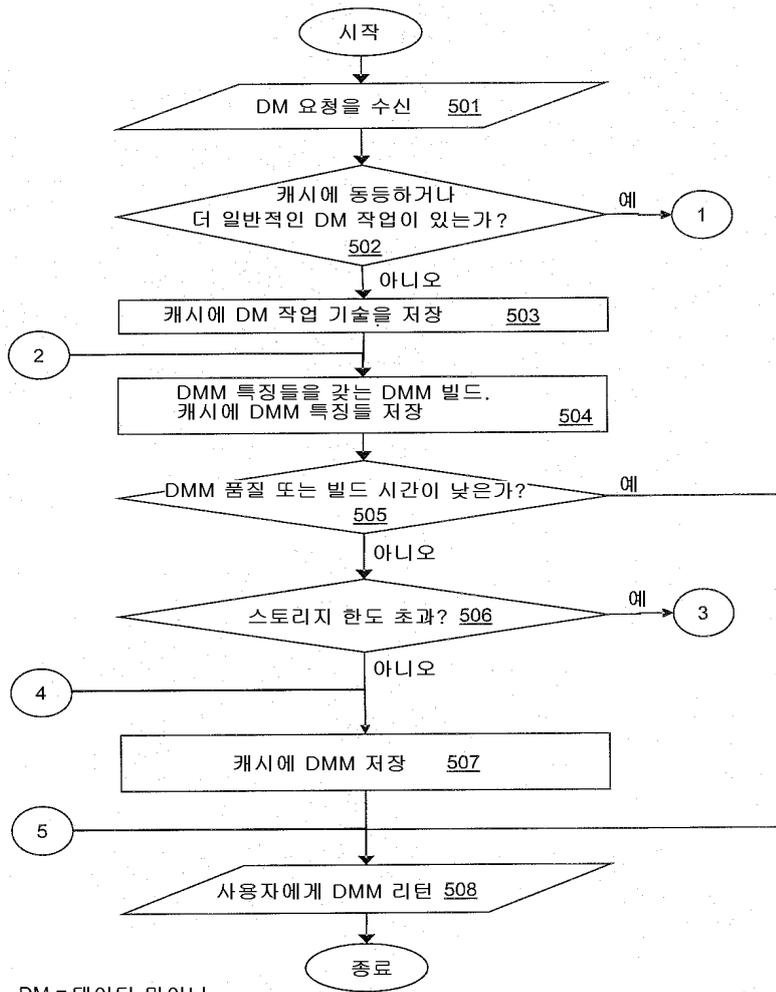
도면3



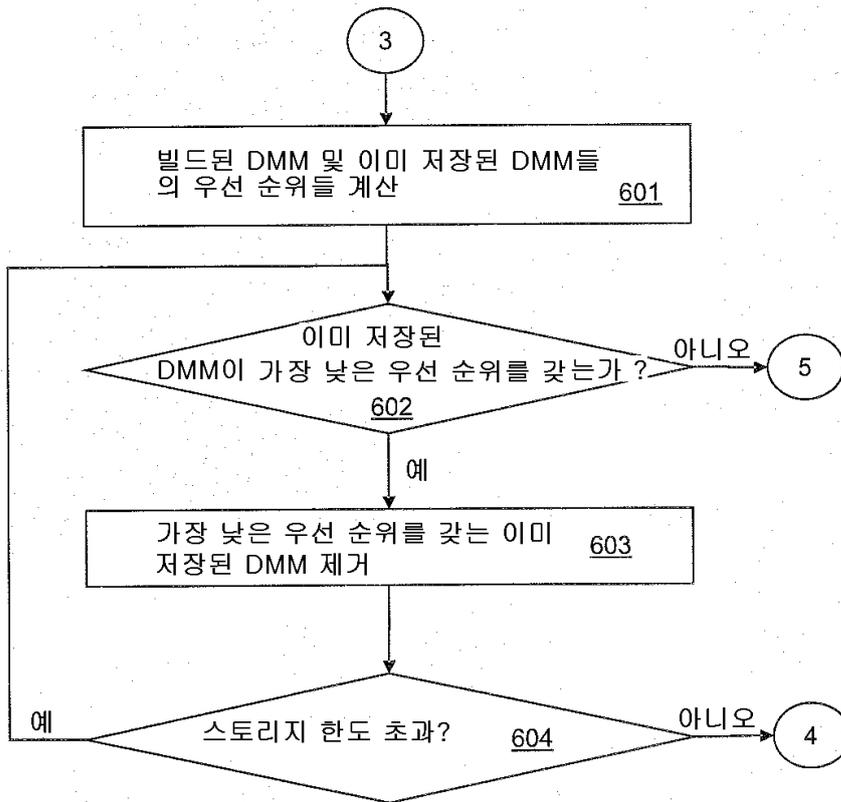
도면4



도면5

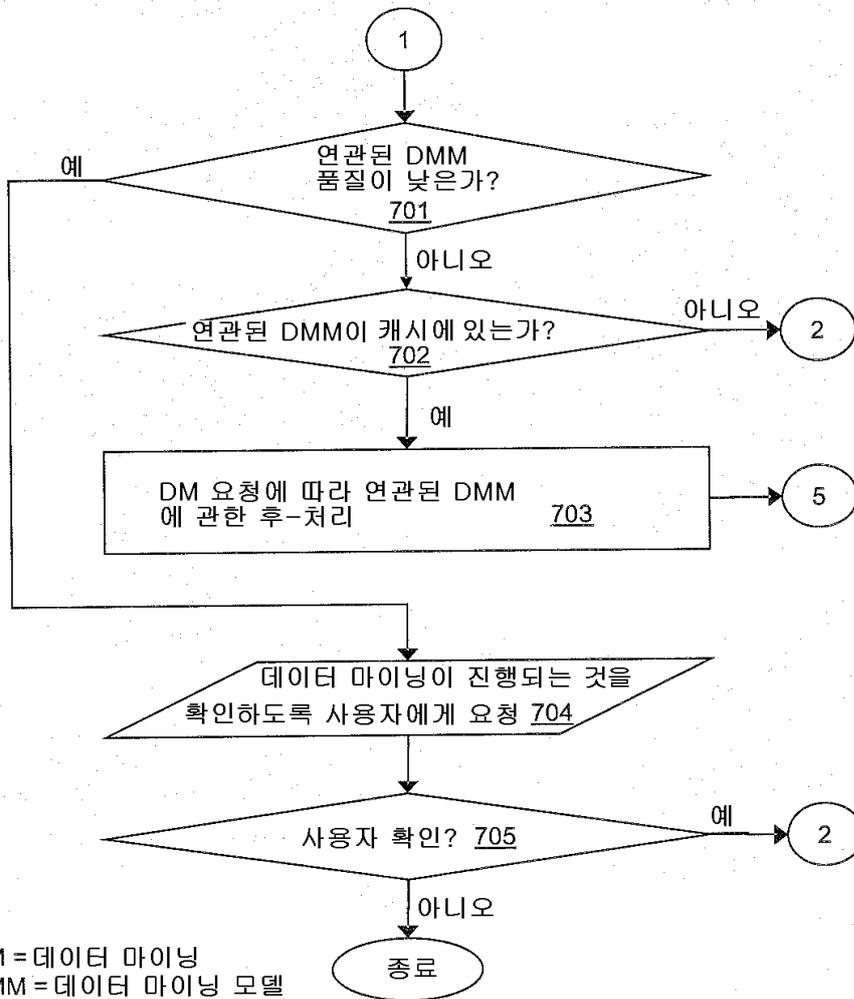


도면6



DM = 데이터 마이닝
 DMM = 데이터 마이닝 모델

도면7



도면8

810

요청 Id	R1A	R1B	R2	R3A	R3B
마이닝 유형	연관 규칙들	연관 규칙들	연관 규칙들	연관 규칙들	연관 규칙들
데이터 선택 제한	Table= DEFECTS Filter= Munich	Table= DEFECTS Filter= Munich	Table= DEFECTS Filter= Stuttgart	Table= DEFECTS Filter= Berlin	Table= DEFECTS Filter= Berlin
마이닝 제한	MinSupp= 10%	MinSupp= 20%	MinSupp= 10%	MinSupp= 10%	MinSupp= 20%

820

작업 Id	T0	T1	T2	T3
마이닝 유형	클러스터링	연관 규칙들	연관 규칙들	연관 규칙들
데이터 선택 제한	Table= CUST Filter= All	Table= DEFECTS Filter= Munich	Table= DEFECTS Filter= Stuttgart	Table= DEFECTS Filter= Berlin
마이닝 제한	MaxNum=5	MinSupp= 10%	MinSupp= 10%	MinSupp= 10%

830

모델 Id	M0	M1	M2	M3
품질	90%	76%	40%	61%
빌드 시간	20 min	15 min	12 min	18 min
접근	5	1, 2	1	1, 2
크기	5 MB	2 MB	5 MB	3 MB
우선 순위	5 * 5 MB = 25	1 * 2 MB = 2 2 * 2 MB = 4	1 * 5 MB = 5	1 * 3 MB = 3 2 * 3 MB = 6

도면9

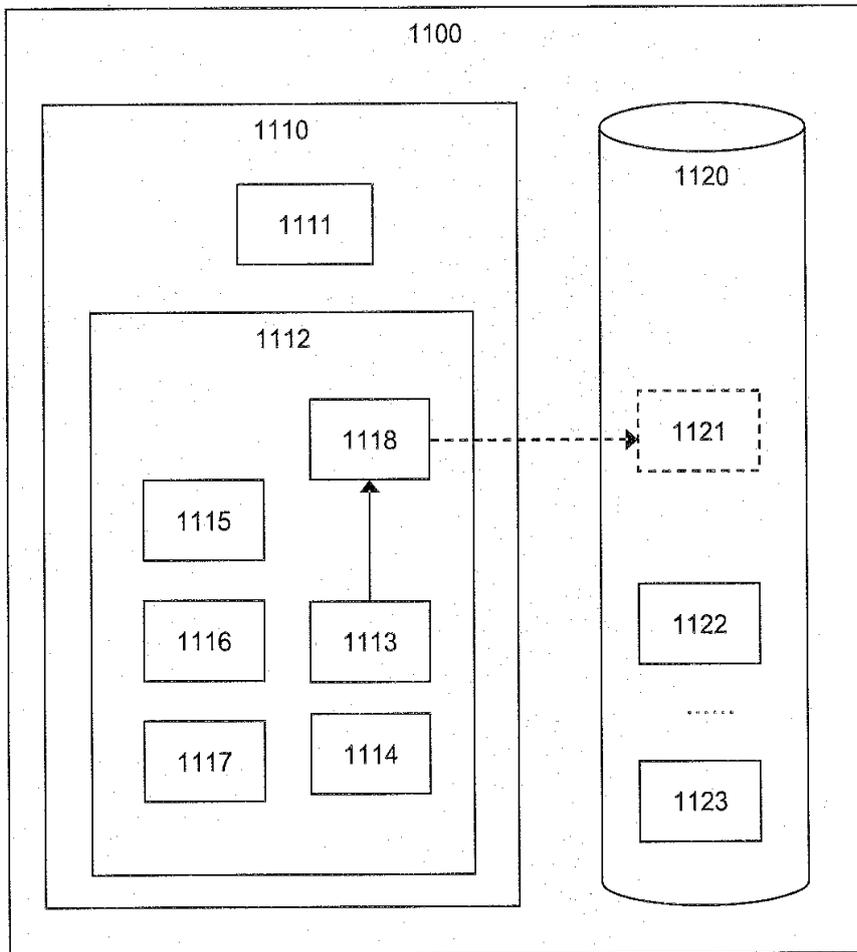
910

단계	모델 Id를 위한 구현			
	M0	M1	M2	M3
911	YES	-	-	-
912	YES	YES	-	-
913	YES	YES	NO	-
914	YES	YES	NO	MAYBE
915	YES	MAYBE	NO	YES

도면10

모델	결함 부분들을 위한 규칙	지지도	신뢰도
규칙1	실린더 헤드 가스켓-> 이그니션 플러그	24%	68%
규칙 2	스티어링 샤프트 -> 스티어링 스쿠 기어	23%	83%
규칙 3	휠 서스펜션-> 악셀 박스	11%	77%
품질			평균 : 76%

도면11



도면12

