(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0074806
(43) 공개일자 2014년06월18일

(51) 국제특허분류(Int. Cl.)
G06F 21/00 (2006.01) *G06F 17/30* (2006.01)
G06F 15/16 (2006.01)
(21) 출원번호 10-2013-0086873
(22) 출원일자 2013년07월23일
심사청구일자 2013년07월23일
(30) 우선권주장
1020120143010 2012년12월10일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정익래
서울 광진구 긴고랑로4길 53, 203호 (중곡동, 호동아파트)
이동훈
서울 성북구 안암로 145, 미래융합기술관 612호 (안암동5가, 고려대학교이과대학)
(뒷면에 계속)
(74) 대리인
특허법인엠에이피에스

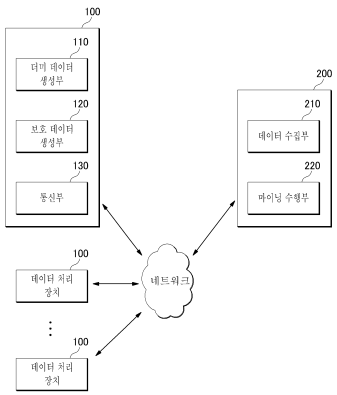
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **데이터 처리 장치, 데이터 스트림 마이닝 수행장치, 그 수행방법 및 기록매체**

(57) 요 약

본 발명은, 데이터 처리 장치, 마이닝 수행장치, 데이터 스트림 마이닝 수행방법 및 기록 매체에 대해 개시한다. 특히, 본 발명의 일 실시예에 따른 데이터 처리 장치는, 더미 데이터를 생성하는 더미 데이터 생성부; 데이터 처리 과정에 따라 생성된 원본 데이터에 상기 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성하는 보호 데이터 생성부; 및 상기 보호 데이터를 상기 마이닝 수행장치로 전송하는 통신부를 포함한다. 이때, 보호 데이터 생성부는 상기 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과가 동일하게 되도록 상기 보호 데이터를 생성한다.

대 표 도 - 도1



(72) 발명자

정재열

울산 중구 병영4길 7-8, (남외동)

김기성

경기 부천시 원미구 상일로 72, 1842동 805호 (상동, 반달마을아파트)

김동민

서울 동대문구 약령시로 154, 1동 505호 (청량리동, 미주아파트)

홍준희

서울 중구 다산로 32, 27동 201호 (신당동, 남산타운아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1415128860
부처명	미래창조과학부
연구사업명	정보통신기술인력양성
연구과제명	스마트 그리드 기술연구
기 여 율	1/1
주관기관	고려대학교 산학협력단
연구기간	2013.01.01 ~ 2013.12.31

특허청구의 범위

청구항 1

데이터 스트림 마이닝을 수행하는 마이닝 수행장치로 데이터를 전송하는 데이터 처리 장치에 있어서,

더미 데이터를 생성하는 더미 데이터 생성부;

데이터 처리 과정에 따라 생성된 원본 데이터에 상기 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성하는 보호 데이터 생성부; 및

상기 보호 데이터를 상기 마이닝 수행장치로 전송하는 통신부를 포함하고,

상기 보호 데이터 생성부는 상기 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과가 동일하게 되도록 상기 보호 데이터를 생성하는,

데이터 처리 장치.

청구항 2

제 1 항에 있어서,

상기 보호 데이터 생성부는 상기 더미 데이터를 균등 분포 또는 임의 분포로 삽입하여 상기 보호 데이터를 생성하는, 데이터 처리 장치.

청구항 3

데이터 스트림 마이닝을 수행하는 마이닝 수행장치에 있어서,

하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집하는 데이터 수집부;

상기 수집된 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행하는 마이닝 수행부를 포함하고,

상기 보호 데이터는 상기 사용자 단말의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이고,

상기 마이닝 수행부의 수행 결과는 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 동일한,

데이터 스트림 마이닝 수행장치.

청구항 4

사용자의 프라이버시를 보호할 수 있는 데이터 스트림 마이닝 수행방법에 있어서,

더미 데이터를 생성하는 단계;

데이터 처리 과정에 따라 생성된 원본 데이터에 상기 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성하는 단계; 및

상기 보호 데이터를 상기 데이터 스트림 마이닝을 수행하는 마이닝 수행장치로 전송하는 단계를 포함하고,

상기 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과가 동일하게 되도록 상기 보호 데이터를 생성하는 단계에서 상기 보호 데이터를 생성하는,

데이터 스트림 마이닝 수행방법.

청구항 5

제 4 항에 있어서,

상기 보호 데이터를 생성하는 단계는 상기 더미 데이터를 균등 분포 또는 임의 분포로 삽입하여 상기 보호 데이터를 생성하는, 데이터 스트림 마이닝 수행방법.

청구항 6

사용자의 프라이버시를 보호할 수 있는 데이터 스트림 마이닝 수행방법에 있어서,

하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집하는 단계;

상기 수집된 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행하는 단계를 포함하고,

상기 보호 데이터는 상기 사용자 단말의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이고,

상기 데이터 스트림 마이닝을 수행하는 단계의 수행 결과는 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 동일한 것인, 데이터 스트림 마이닝 수행방법.

청구항 7

제 4 항 내지 제 6 항 중 어느 한 항에 기재된 방법을 컴퓨터 상에서 수행하기 위한 프로그램을 기록한 컴퓨터 판독 가능한 기록 매체.

명 세 서

기 술 분 야

[0001] 본 발명은 데이터 처리 장치, 데이터 스트림 마이닝 수행장치 및 데이터 스트림 마이닝 수행방법에 관한 것이고, 보다 구체적으로 사용자의 프라이버시를 보호할 수 있는 데이터 처리 장치, 데이터 스트림 마이닝 수행장치 및 데이터 스트림 마이닝 수행방법에 관한 것이다.

배 경 기 술

[0002] 컴퓨터 기술의 발달과 이를 활용한 컴퓨터 응용 환경의 변화에 따라 근래 들어 많은 분야에서 데이터 스트림(Data Stream) 형태로 정보를 발생시키고 있다. 일반적으로 데이터 스트림은 구성요소가 지속적으로 생성되는 무한집합으로 정의되므로, 데이터 스트림에 대한 특성 분석을 하기 위해서는 데이터 스트림을 실시간으로 빠르게 처리할 수 있는 기술이 필요하다.

[0003] 특히, 실시간으로 수집된 데이터 스트림 중에서 어떤 데이터가 빈번하게 발생하는가에 대해 찾는 기술을 데이터 스트림 마이닝(Data Stream Mining) 기술이라고 한다. 이러한 데이터 스트림 마이닝 기술은 일반적으로 많은 데이터를 계속 수집하여 한번에 마이닝을 수행하는 것이 아니라, 실시간으로 수집된 데이터 스트림에 대해 즉시 마이닝을 수행한다. 이로 인해 사용자는 실시간으로 빈발 항목에 대한 정보를 얻을 수 있고, 얻은 정보를 기초로 패턴 분석이나 마케팅 활동을 진행할 수 있다.

[0004] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically", In Proc. of ACM PODS, 2003.의 논문은 데이터 스트림 환경에서 빈발 항목을 찾는 기술을 소개하고 있다. 구체적으로, 사용자 단말에서 마이닝 수행장치로 송신하는 데이터에 실시간으로 고유 번호가 부여된다. 마이닝 수행장치는 동일한 종류의 데이터에 대하여 동일한 고유 번호를 부여하도록 제어한다. 또한, 마

이닝 수행장치는 부여된 고유 번호를 카운팅하여 데이터 스트림 환경에서 어떤 데이터가 빈발 항목인지 분석한다.

[0005] 다만, 제 3 자가 마이닝 수행장치에 실시간으로 수집된 데이터 스트림을 체크한다면, 제 3 자는 타인의 시간대별 행동을 추측할 수 있다. 예를 들어, 마이닝 수행장치를 체크한 B는 A가 오전 10시에 DMB를 시청하고, 오후 2시에 전화를 했다는 사실을 알 수 있다. 이는 심각한 프라이버시 침해를 야기할 수 있다.

[0006] 이와 같은 문제를 해결하기 위하여 다음과 같은 선행기술이 존재한다.

[0007] Gi Sung Kim "Practical Privacy-Preserving DBSCAN Clustering Over Horizontally Partitioned Data"의 논문은 프라이버시를 보존하기 위해 DB스캔을 수행하는 기술을 제안하였다. 하지만 DB스캔 기술은 데이터 스트림 환경에서 사용하기 부적합하고, 현실적으로 구현하기 어렵다는 문제점이 있었다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 본 발명의 일부 실시예는 데이터 스트림 환경에서 사용자의 프라이버시 또는 개인정보를 보호할 수 있는 데이터 처리 장치, 데이터 스트림 마이닝 수행장치 및 데이터 스트림 마이닝 수행방법을 제공하는 데에 그 목적이 있다.

[0009] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0010] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 실시예에 따른 데이터 처리 장치는, 더미 데이터를 생성하는 더미 데이터 생성부; 데이터 처리 과정에 따라 생성된 원본 데이터에 상기 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성하는 보호 데이터 생성부; 및 상기 보호 데이터를 마이닝 수행장치로 전송하는 통신부를 포함하고, 상기 보호 데이터 생성부는 상기 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과가 동일하게 되도록 상기 보호 데이터를 생성한다.

[0011] 또한, 본 발명의 다른 실시예에 따른 마이닝 수행장치는, 하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집하는 데이터 수집부; 상기 수집된 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행하는 마이닝 수행부를 포함하고, 상기 보호 데이터는 상기 사용자 단말의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이고, 상기 마이닝 수행부의 수행 결과는 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 동일하다.

[0012] 또한, 본 발명의 또 다른 실시예에 따른 데이터 스트림 마이닝 수행방법은, 더미 데이터를 생성하는 단계; 데이터 처리 과정에 따라 생성된 원본 데이터에 상기 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성하는 단계; 및 상기 보호 데이터를 상기 데이터 스트림 마이닝을 수행하는 마이닝 수행장치로 전송하는 단계를 포함하고, 상기 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과가 동일하게 되도록 상기 보호 데이터를 생성하는 단계에서 상기 보호 데이터를 생성한다.

[0013] 아울러, 본 발명의 또 다른 실시예에 따른 데이터 스트림 마이닝 수행방법은, 하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집하는 단계; 상기 수집된 보호 데이터에 대하여 상기 데이터 스트림 마이닝을 수행하는 단계를 포함하고, 상기 보호 데이터는 상기 사용자 단말의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이고, 상기 수행하는 단계의 수행 결과는 상기 원본 데이터에 대하여 상기 데이터 스트림 마이닝을 수행한 결과와 동일하다.

발명의 효과

[0014] 전술한 본 발명의 과제 해결 수단 중 어느 데이터 처리 장치는, 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과가 실질적으로 동일하게 되도록, 보호 데이터를 생성함으로써, 제 3 자가 원본 데이터가 무엇인지 알 수 없게 처리할 수 있고, 그로 인해 사용자의 프라이버시를 보호할 수 있다는 장점이 있다.

[0015] 또한, 전술한 본 발명의 과제 해결 수단 중 어느 하나인 마이닝 수행장치는, 수집된 보호 데이터에 대하여 데이터 스트림 마이닝을 수행하는 마이닝 수행부를 포함함으로써, 사용자의 프라이버시가 보호되는 상태에서 데이터 스트림 마이닝을 수행할 수 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 일 실시예에 따른 데이터 처리 장치 및 데이터 스트림 마이닝 수행장치를 도시한 블록도,

도 2는 원본 데이터 및 더미 데이터의 일 예를 나타낸 도면,

도 3은 더미 데이터를 균등 분포로 삽입한 경우의 일 예를 나타낸 도면,

도 4는 더미 데이터를 균등 분포 또는 임의 분포로 삽입한 경우의 일 예를 나타낸 도면,

도 5는 본 발명의 일 실시예에 따른 데이터 스트림 마이닝 수행방법을 설명하기 위한 순서도,

도 6은 본 발명의 다른 실시예에 따른 데이터 스트림 마이닝 수행방법을 설명하기 위한 순서도,

도 7은 원본 데이터 및 보호 데이터의 일 예를 나타낸 도면,

도 8a 및 도 8b는 도 5에 도시된 방법을 컴퓨터 상에서 수행하기 위한 프로그램의 일부를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0017] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0018] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0019] 도 1은 본 발명의 일 실시예에 따른 데이터 처리 장치 및 데이터 스트림 마이닝 수행장치를 도시한 블록도이다.

[0020] 도 1에 도시된 데이터 처리 장치(100) 및 데이터 스트림 마이닝 수행장치(200)는 상호 간에 네트워크(network, 10)를 통해 연결된다. 네트워크(10)는 단말들 및 서버들과 같은 각각의 노드 상호 간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 이러한 네트워크의 일 예에는 이동 통신(Mobile Radio Communication) 네트워크, 위성 방송 네트워크, 아날로그 방송 네트워크, DMB(Digital Multimedia Broadcasting) 네트워크, 인터넷(Internet), LAN(Local Area Network), Wireless LAN(Wireless Local Area Network), WAN(Wide Area Network), PAN(Personal Area Network) 등이 포함되나 이에 한정되지는 않는다.

[0021] 또한, 도 1에 도시된 데이터 처리 장치(100)는 상술한 네트워크(10)를 통해 데이터 스트림 마이닝 수행장치(200)와 데이터를 송수신할 수 있는 휴대용 단말기 또는 컴퓨터로 구현될 수 있다. 여기서, 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, PCS(Personal Communication System), GSM(Global System for Mobile communications), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access)-2000, W-CDMA(W-Code Division Multiple Access), WiBro(Wireless Broadband Internet) 단말, 스마트폰(Smart Phone) 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치를 포함할 수 있다. 또한, 컴퓨터는 예를 들어, 휴대성과 이동성이 보장되는 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop), 테블릿 PC, 슬레이트 PC 등을 포함할 수 있다.

[0022] 본 발명의 일 실시예에 따른 데이터 처리 장치(100)는 후술할 마이닝 수행장치(200)로 데이터를 전송하고, 더미

데이터 생성부(110), 보호 데이터 생성부(120) 및 통신부(130)를 포함한다.

[0023] 더미 데이터 생성부(110)는 더미 데이터(dummy data)를 생성한다. 더미 데이터는 일종의 난수 생성기로부터 생성될 수 있고, 사용자의 프라이버시를 보호하기 위해 필수적인 데이터이다.

[0024] 보호 데이터 생성부(120)는 데이터 처리 과정에 따라 생성된 원본 데이터(original data)에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 보호 데이터를 생성한다. 여기서, 원본 데이터는 사용자가 데이터 처리 장치(100)를 사용함에 따라 생성될 수 있고, 데이터 처리 장치(100)의 데이터 처리 과정에서 자연스럽게 생성될 수 있다.

[0025] 또한, 보호 데이터 생성부(120)는 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과가 실질적으로 동일하게 되도록 보호 데이터를 생성한다. 이에 대하여 후술하기로 한다.

[0026] 도 2는 원본 데이터 및 더미 데이터의 일 예를 나타낸 도면이다. 도 2는 데이터 처리 장치의 일 예인 사용자 단말(100')이 후술할 마이닝 수행장치(200)로 데이터를 전송하는 것을 개략적으로 도시하고 있다. 특히, 아래 그림에서 네모 표시된 데이터(1, 2, -1, 3, 4, -2, -4, -3)가 더미 데이터일 수 있고, 원본 데이터와 더미 데이터를 합한 데이터가 보호 데이터일 수 있다.

[0027] 아울러, 보호 데이터 생성부(120)는 원본 데이터에 더미 데이터를 균등 분포 또는 임의 분포로 삽입하여 보호 데이터를 생성할 수 있다.

[0028] 도 3은 더미 데이터를 균등 분포로 삽입한 경우의 일 예를 나타낸 도면이다. 원본 데이터에 더미 데이터를 균등 분포로 삽입하는 경우, 더미 데이터는 원본 데이터의 분포에 영향을 주지 않는다. 도 3의 위에 도시된 원본 데이터와 도 3의 아래에 도시된 보호 데이터는 동일한 것을 확인할 수 있다. 즉, 보호 데이터 생성부(120)가 더미 데이터를 균등 분포로 삽입한 경우, 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과는 실질적으로 동일하다.

[0029] 원본 데이터에 더미 데이터를 임의 분포로 삽입하는 경우, 보호 데이터 생성부(120)는 더미 데이터가 원본 데이터의 분포에 영향을 최대한 주지 않도록 더미 데이터를 삽입하여 보호 데이터를 생성할 수 있다. 이 경우, 보호 데이터 생성부(120)는 더미 데이터 삽입의 최대 구간을 정한 후, 그 구간을 최대 값으로 설정한 난수생성기를 이용할 수 있다. 또한, 그 난수생성기를 통해 나온 수를 구간으로 지정하고, 그 구간에 맞게 더미 데이터를 삽입하는 방식을 사용할 수 있다.

[0030] 원본 데이터에 더미 데이터를 균등 분포로 삽입하는 경우, 제 3 자가 노출된 더미 데이터에 의해 원본 데이터를 유추해볼 수 있는 여지가 생길 수 있다. 그러나 원본 데이터에 더미 데이터를 임의 분포로 삽입하는 경우, 제 3 자가 노출된 더미 데이터로부터 원본 데이터를 유추할 수 있는 가능성을 최소화할 수 있다.

[0031] 다시 도 1을 참고하면, 통신부(130)는 유무선 경로를 통해 데이터 스트림 마이닝을 수행하는 마이닝 수행장치로 상술한 보호 데이터를 전송한다.

[0032] 한편, 본 발명의 일 실시예에 따른 마이닝 수행장치(200)는 상술한 데이터 스트림 마이닝을 수행하고, 데이터 수집부(210) 및 마이닝 수행부(220)를 포함한다. 여기서, 마이닝 수행장치(200)는 상술한 네트워크(10)를 통해 데이터 처리 장치(100) 또는 사용자 단말(100')과 데이터를 송수신할 수 있는 단말기 또는 컴퓨터로 구현될 수 있다. 단말기 또는 컴퓨터는 상술한 것과 같은 기기일 수 있다.

[0033] 데이터 수집부(210)는 하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집한다. 실시간으로 수신된 보호 데이터는 데이터 스트림이라 볼 수 있다.

[0034] 마이닝 수행부(220)는 수집된 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한다.

[0035] 이때, 보호 데이터는 사용자 단말(100')의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이고, 마이닝 수행부(220)의 수행 결과는 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 실질적으로 동일하다.

[0036] 또한, 미리 설정된 기준은 균등 분포 또는 임의 분포일 수 있다. 도 4는 더미 데이터를 균등 분포 또는 임의 분포로 삽입한 경우의 일 예를 나타낸 도면이다.

[0037] 도 4의 왼쪽에 도시된 것처럼 a, b, c로 표시된 각 사용자들로부터 실시간으로 수집된 원본 데이터에 대하여 데

이터 스트림 마이닝을 수행한 결과를 제 1 결과라고 하자. 또한, 도 4의 오른쪽에 도시된 것처럼 a, b, c로 표시된 각 사용자들로부터 실시간으로 수집된 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과를 제 2 결과라고 하자. 이때, 보호 데이터는 원본 데이터에 더미 데이터를 균등 분포 또는 임의 분포로 삽입한 것인데, 제 1 결과와 제 2 결과가 실질적으로 동일하게 되도록 본 발명에서 제안된 사용자 단말(100') 혹은 데이터 처리 장치(100)에 의해 생성될 수 있다.

[0038] 지금까지 설명한 것처럼 본 발명에서 제안되는 데이터 처리 장치(100)는, 마이닝 수행장치(200)에서 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과가 실질적으로 동일하게 되도록, 보호 데이터를 생성함으로써, 제 3 자가 원본 데이터가 무엇인지 알 수 없게 처리할 수 있고, 그로 인해 사용자의 프라이버시를 보호할 수 있다는 장점이 있다. 또한, 본 발명에서 제안되는 마이닝 수행장치(200)는 사용자의 프라이버시가 보호되는 상태에서 데이터 스트림 마이닝을 수행할 수 있다.

[0039] 이하에서는, 본 발명의 각 실시예에 따른 데이터 스트림 마이닝 수행방법에 대하여 도 5 및 도 6을 참고하여 설명하기로 한다.

[0040] 도 5는 본 발명의 일 실시예에 따른 데이터 스트림 마이닝 수행방법을 설명하기 위한 순서도이다.

[0041] 사용자의 프라이버시를 보호할 수 있는 데이터 스트림 마이닝 수행방법은 상술한 데이터 처리 장치(100)에 의해 이루어질 수 있다.

[0042] 우선 더미 데이터가 생성된다(S110). 이러한 동작은 데이터 처리 장치(100)의 더미 데이터 생성부(110)에 의해 이루어질 수 있다.

[0043] 이어서, 원본 데이터에 더미 데이터를 삽입하여 보호 데이터가 생성되는 과정이 이루어진다(S120). 특히, 원본 데이터는 데이터 처리 과정에 따라 생성된 것이고, 데이터 삽입은 미리 설정된 기준에 따라 이루어진다.

[0044] 이때, 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과가 실질적으로 동일하게 되도록, 보호 데이터를 생성하는 단계(S120)에서 보호 데이터가 생성된다.

[0045] 또한, 보호 데이터를 생성하는 단계(S120)에서 더미 데이터를 균등 분포 또는 임의 분포로 삽입하여 보호 데이터가 생성될 수 있다. 이러한 동작은 데이터 처리 장치(100)의 보호 데이터 생성부(120)에 의해 이루어질 수 있다.

[0046] 계속해서, 보호 데이터가 데이터 스트림 마이닝을 수행하는 마이닝 수행장치(200)로 전송된다(S130). 이러한 동작은 데이터 처리 장치(100)의 통신부(130)에 의해 이루어질 수 있다.

[0047] 한편, 도 6은 본 발명의 다른 실시예에 따른 데이터 스트림 마이닝 수행방법을 설명하기 위한 순서도이다.

[0048] 사용자의 프라이버시를 보호할 수 있는 데이터 스트림 마이닝 수행방법은 상술한 마이닝 수행장치(200)에 의해 이루어질 수 있다.

[0049] 우선, 하나 이상의 사용자 단말로부터 실시간으로 수신된 보호 데이터를 수집한다(S210). 이러한 동작은 마이닝 수행장치(200)의 데이터 수집부(210)에 의해 이루어질 수 있다.

[0050] 이어서, 수집하는 단계(S210)에서 수집된 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한다(S220). 이때, 보호 데이터는 상술한 사용자 단말(100')의 데이터 처리 과정에 따라 생성된 원본 데이터에 더미 데이터를 미리 설정된 기준에 따라 삽입하여 생성된 것이다. 또한, 수행하는 단계(S220)의 수행 결과는 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 실질적으로 동일하다. 이러한 동작은 마이닝 수행장치(200)의 마이닝 수행부(220)에 의해 이루어질 수 있다.

[0051] 지금까지 설명한 것처럼 본 발명에서 제안되는 데이터 스트림 마이닝 수행방법은 원본 데이터에 더미 데이터를 삽입시킨 보호 데이터를 생성할 수 있고, 이를 이용하여 사용자의 프라이버시를 보호할 수 있다는 장점이 있다.

[0052] 이하에서는, 상술한 내용을 일 예에 적용한 결과를 도 7을 참고하여 설명하기로 한다. 도 7은 원본 데이터 및 보호 데이터의 일 예를 나타낸 도면이다.

[0053] 도 7에 도시된 것과 같은 원본 데이터가 있는 경우, 사용자의 프라이버시 보호를 위해서 더미 데이터를 삽입하여 보호 데이터를 생성한다. 구체적으로 다음과 같은 과정을 거칠 수 있다.

[0054] 첫째, m 종류의 데이터가 존재한다면 $\log_2 m+1$ 개의 카운터를 가지게 된다. 그 중 첫 번째 카운터(C_0)는 데이터의 총 수를 나타낸다. 총수는 +데이터의 수에서 -데이터의 수를 뺀 값이다. 예를 들어 0, 1, 2, 3 네 종류의 데이터가 있을 때, 수집한 데이터가 1, 2, -1, 2, 0, 3 라면 카운터의 수는 3개이고, 첫 번째 카운터는 6 - 1 = 5이다.

[0055] 둘째, 수집된 데이터를 2진수로 바꾸고 2진수에서 1이 있는 위치의 카운터에 1씩 더해준다. 앞에서의 예에서 3이 수집되었다면 3은 이진수로 11이기 때문에 두 번째(C_1)와 세 번째(C_2) 카운터에 각각 1씩 더해주는 것이다.

[0056] 셋째, 데이터를 수집하면서 임의의 순간에 균등분포를 가지는 데이터들을 원본 데이터처럼 수집하여 첫째와 둘째 단계를 수행한다. 넷째, 데이터를 수집 중에 수집한 데이터의 빈발 항목이 궁금하면 첫 번째 카운터를 2로 나눈 값보다 큰 카운터들을 찾아서 그 카운터를 2진수로 나타낸다. 만약 위의 예에서 두 번째와 세 번째 카운터 모두 첫 번째 카운터를 2로 나눈 값보다 크다면 지금까지 수집한 데이터들 중에서 빈발 항목은 3이 되는 것이다.

[0057] 다시 원본 데이터만을 수집한 데이터를 분석해보면 데이터의 종류가 8개이기 때문에, 카운터의 수는 4개이다. 첫 번째 카운터는 13이고 두 번째 카운터는 5, 세 번째 카운터는 8, 네 번째 카운터는 5 값을 가진다. 더미 데이터를 포함한 보호 데이터를 분석해보면, 첫 번째 카운터는 원본 데이터를 분석한 결과와 동일한 13이고, 두 번째 카운터와 세 번째 카운터, 네 번째 카운터 역시 같은 값을 가진다. 5, 8, 5 값을 6(13/2)과 비교하면, 세 번째 카운터 값인 8만 크다는 것을 알 수 있다. 그러면 2진수 010값을 가지는 데이터 즉, 2가 가장 빈발 항목이라고 분석할 수 있다. 또한, 원본 데이터와 더미 데이터는 동일한 카운터 값을 가지고 있기 때문에, 보호 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과와 원본 데이터에 대하여 데이터 스트림 마이닝을 수행한 결과가 실질적으로 동일하게 된 것을 확인할 수 있다.

[0058] 또한, 보호 데이터를 통해 사용자의 프라이버시를 보호할 수도 있다는 것을 확인할 수 있다. 이를 위해 본 발명의 일 실시예에 따른 데이터 처리 장치(100)가 더미 데이터의 삽입 방법과 위치를 결정하여 보호 데이터를 생성한다. 이미 상술한 것처럼 더미 데이터를 균등한 위치에 삽입하면, 악의적인 공격자가 원본 데이터 및 더미 데이터를 추측해낼 수 있다. 이러한 문제를 방지하기 위해, 난수 발생기가 삽입 동작이 이루어질 때마다 더미 데이터를 생성하여, 임의의 위치에 생성된 더미 데이터를 삽입할 수 있다.

[0059] 나아가, 도 8a 및 도 8b는 도 5에 도시된 방법을 컴퓨터 상에서 수행하기 위한 프로그램의 일부를 나타낸 도면이다.

[0060] 빈발 항목 찾기 알고리즘에서 log m + 1 카운터가 유지된다.

[0061] 이때, 첫번째 카운터(c_0)는 살아있는 아이템의 전체 개수는 $n(t) = \sum_x n_x(t)$이고, $c_0,\ldots,c_{\log m}$ 카운터가 하나의 그룹을 이룬다.

[0062] bit(x , j) - x 를 이진수로 바꾸어 해당하는 j 번째 값을 넣어준다. gt(x , y) - x와 y를 비교하여, x가 크면 1이고, 아니면 0인 것으로 처리한다.

[0063] 예를 들어, 원본 데이터가 1, 3, 0, 2, 3, 3, 0, -2, -1, 3, 2, 2, 3, 1이라면, 4 item, 3 counter, c_0 = 10, c_1= 6, c_2= 7이다.

[0064] 또한, 본 발명은 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 메커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.

[0065] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가

지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0066] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0067] 100: 데이터 처리 장치 110: 더미 데이터 생성부

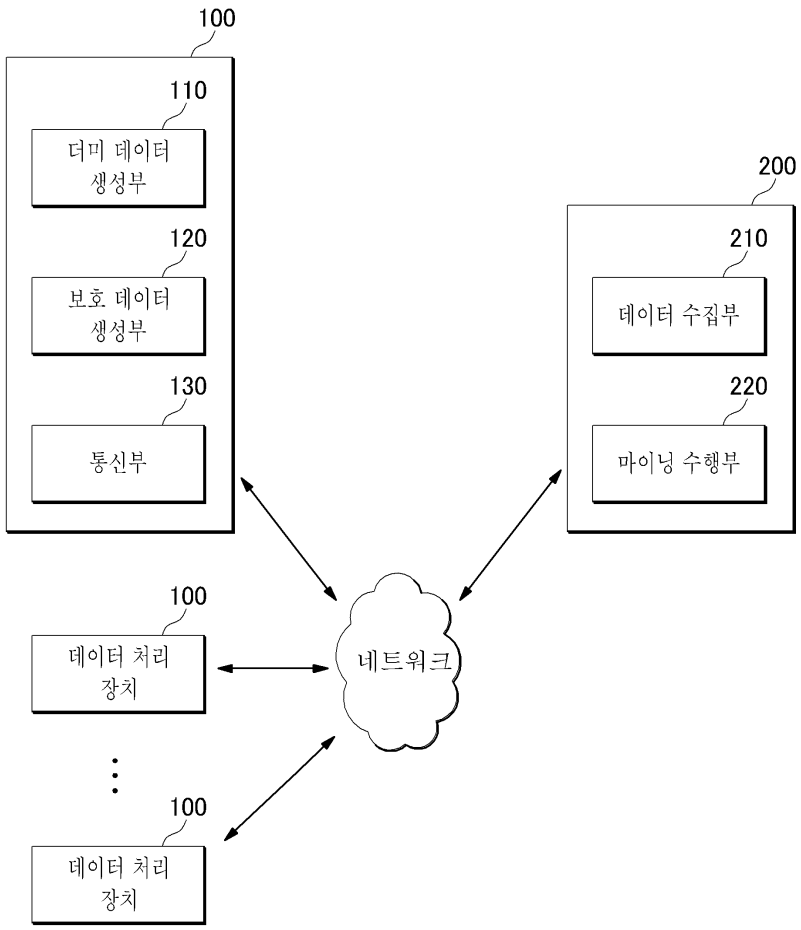
120: 보호 데이터 생성부 130: 통신부

200: 마이닝 수행장치 210: 데이터 수집부

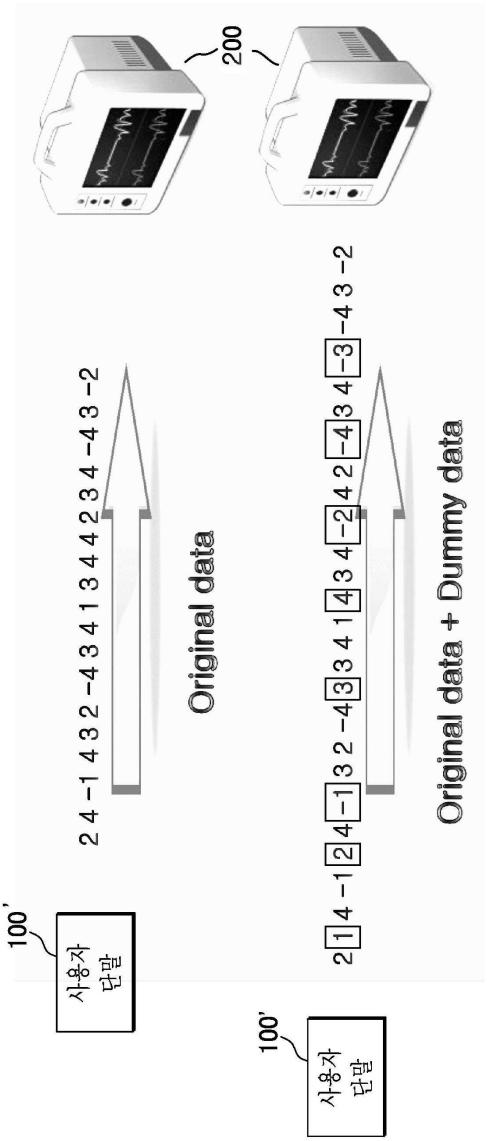
220: 마이닝 수행부

도면

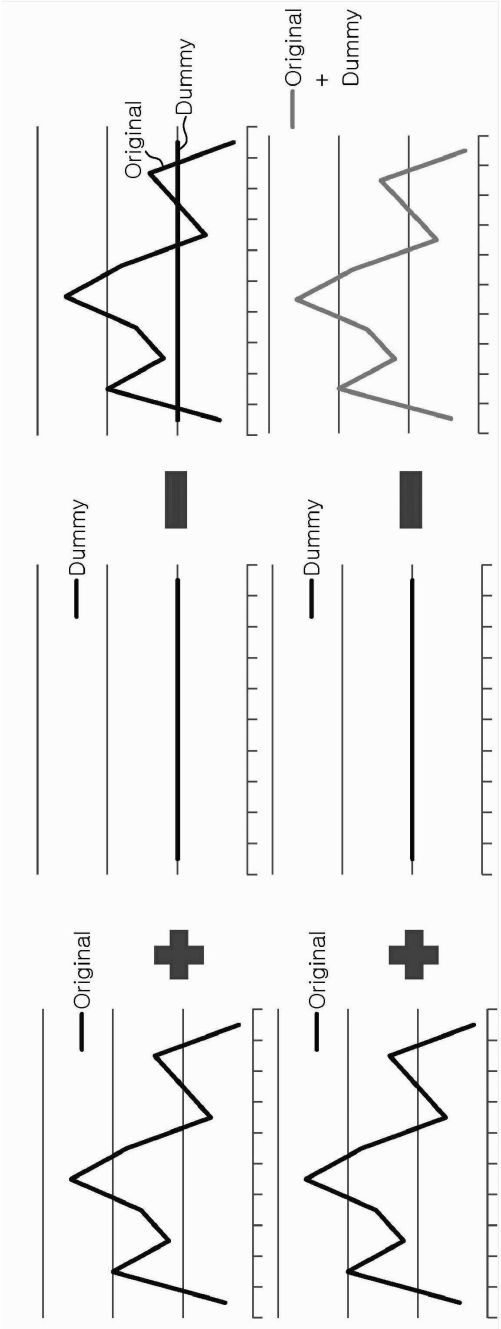
도면1



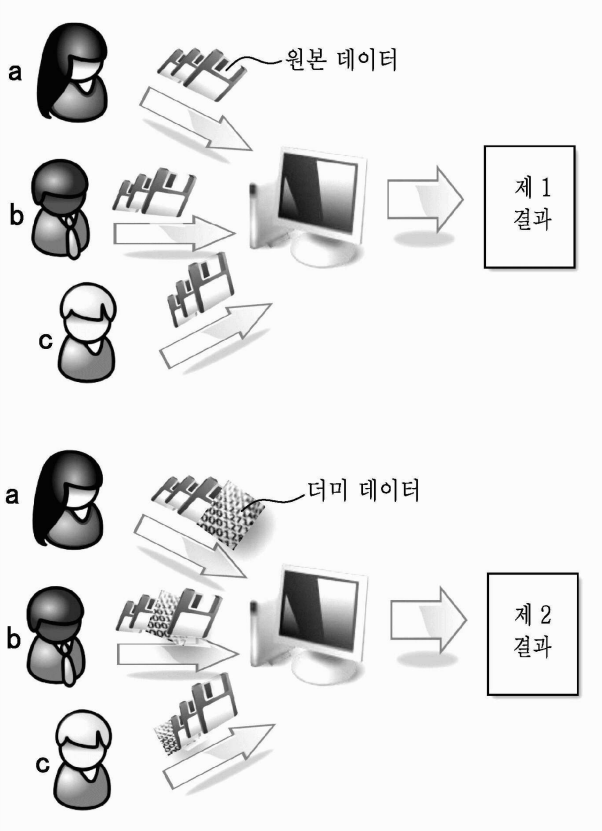
도면2


200
2 4 -1 4 3 2 -4 3 4 1 3 4 4 2 3 4 -4 3 -2
Original data
2 1 4 -1 2 4 -1 3 2 -4 3 4 1 4 3 4 -2 4 2 -4 3 4 -3 -4 3 -2
Original data + Dummy data
100'
사용자 단말
100'
사용자 단말

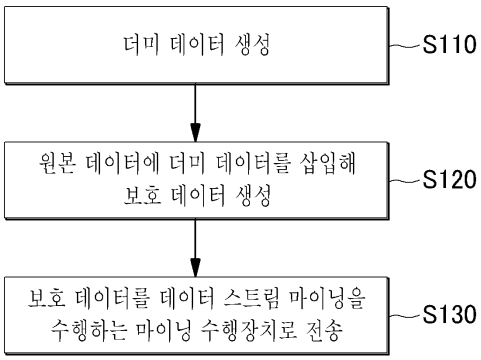
도면3


Original
Dummy
Original + Dummy
Dummy
Dummy
Original
Original

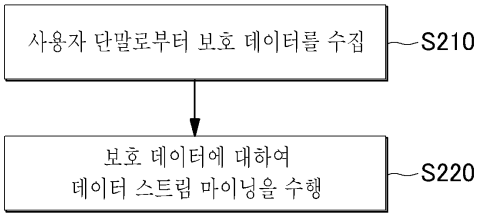
도면4


a
원본 데이터
b
제 1
결과
c
a
더미 데이터
b
제 2
결과
c

도면5


더미 데이터 생성
S110
원본 데이터에 더미 데이터를 삽입해
보호 데이터 생성
S120
보호 데이터를 데이터 스트림 마이닝을
수행하는 마이닝 수행장치로 전송
S130

도면6



도면7

	전송된 데이터
원본 데이터	2 4 −1 5 0 2 −4 3 4 1 3 5 6 2 0 7 2
보호 데이터	2 [1] 4 −1 [7] 5 [−1] 0 2 −4 3 [3] 4 1 6 3 5 6 2 [−3 −6] 0 7 2 [−7]

더미 데이터

도면8a

UpdateCounters $(x, trans, c[0 \ldots \log m])$

If $(trans = insertion)$ then

$d = 1$

else

$d = -1$

$c[0] = c[0] + d$

for $j = 1$ to $\log m$

$c[j] = c[j] + bit(x, j) * d$

Ex) data : 1 3 0 2 3 3 0 -2 -1 3 2 2 3 1

4 item → 3 counter, $c_0 = 10, c_1 = 6, c_2 = 7$

도면8b

FindMajority ($c[0 \ldots \log m]$)

$x = 0, t = 1$

for $j = 1$ to $\log m$ do

if ($c[j] > c[0]/2$) then

$x = x + t$

$t = 2 * t$

return (x)

$c_1 > 5 \,\&\, c_2 > 5$

$x = 0 \rightarrow 1 \rightarrow 3$