(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2022 년 4 월 21 일 (21.04.2022)

WIPO | PCT

(10) 국제공개번호
WO 2022/080583 A1

(51) 국제특허분류:
G06Q 10/06 (2012.01) *G06Q 20/06* (2012.01)
G06Q 10/04 (2012.01) *G06F 40/30* (2020.01)

(21) 국제출원번호: PCT/KR2020/017919

(22) 국제출원일: 2020 년 12 월 9 일 (09.12.2020)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2020-0132680 2020 년 10 월 14 일 (14.10.2020) KR

(71) 출원인: 고려대학교 세종산학협력단 (**KOREA UNIVERSITY RESEARCH AND BUSINESS FOUNDATION, SEJONG CAMPUS**) [KR/KR]; 30019 세종시 조치원읍 세종로 2511, Sejong-si (KR).

(72) 발명자: 김명섭 (**KIM, Myung-Sup**); 30019 세종시 조치원읍 세종로 2511, Sejong-si (KR). 백의준 (**BAEK, EuiJun**); 30019 세종시 조치원읍 세종로 2511, Sejong-si (KR). 박지태 (**PARK, JeeTae**); 30019 세종시 조치원읍 세종로 2511, Sejong-si (KR). 김보선 (**KIM, Boseon**); 30019 세종시 조치원읍 세종로 2511, Sejong-si (KR).

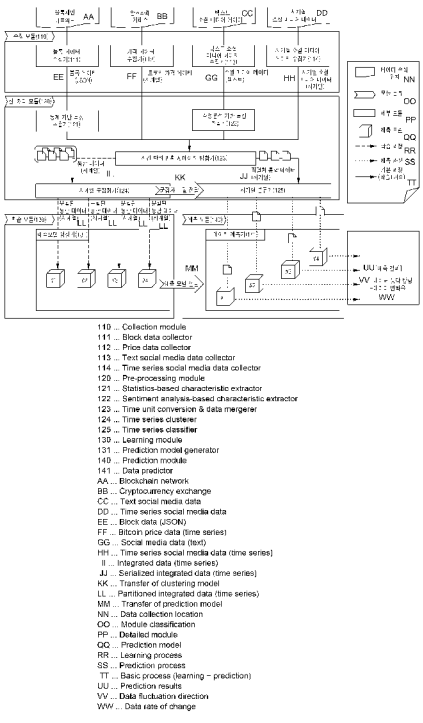
(74) 대리인: 양성보 (**YANG, Sungbo**); 06099 서울시 강남구 선릉로125길 14 삼성빌딩 2층 피앤티특허법률사무소, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

WO 2022/080583 A1

(54) **Title:** DEEP LEARNING-BASED BITCOIN BLOCK DATA PREDICTION SYSTEM TAKING INTO ACCOUNT TIME SERIES DISTRIBUTION CHARACTERISTICS

(54) **발명의 명칭:** 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템



(57) **Abstract:** Disclosed are a deep learning-based Bitcoin block data prediction system and method that take into account time series distribution characteristics. The deep learning-based Bitcoin block data prediction system that takes into account time series distribution characteristics according to the present invention comprises: a data collection module for collecting a plurality of class data including block data, social media data, and price data; a pre-processing module which performs pre-processing for unifying the data formats of the plurality of collected class data as time series data, and clusters the time series data into time series data sets according to the distribution characteristics of the respective time series data for the plurality of pre-processed class data; a learning module which is trained on the plurality of clustered time series data sets through a deep learning-based model, and generates a plurality of prediction models according to the plurality of time series data sets; and a prediction module which evaluates the plurality of learned prediction models, and which, when new data is input, selects a prediction model with which to perform prediction from among the plurality of prediction models.

(57) **요약서:** 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템 및 방법이 제시된다. 본 발명에서 제안하는 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템은 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 데이터 수집 모듈, 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 전처리 모듈, 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 학습 모듈 및 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 예측 모듈을 포함한다.

[다음 쪽 계속]

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템

기술분야

[1] 본 발명은 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템 및 방법에 관한 것이다.

배경기술

[2] 비트코인은 사토시 나카모토에 의해 개발된 첫 번째 블록체인 기반 암호화폐이며 비트코인의 개발 이후 많은 블록체인 기반 암호화폐들이 등장하였다. 블록체인 기술이 지닌 데이터 무결성과 암호화폐의 익명성은 암호화폐 시장을 빠르게 성장시켰으며 현재 시장에서 약 5,800개 암호화폐들이 활발히 거래되고 있다. 또한, 시장에서 거래되고 있는 모든 암호화폐들의 시가총액은 약 390조 원에 육박하며 하루 80조 원 규모의 이루어지는 추세다. 암호화폐 시장이 약한 형태의 효율을 띈다는 연구에도 불구하고 암호화폐 시장의 급성장에 따라 암호화폐의 가격 및 변동 추세는 전 세계의 관심사가 되었으며 이를 예측하는 연구들이 수행되었다. 그러나 암호화폐의 가격은 다양한 요소에 의해 영향을 받고 있기에 직접적인 가격 혹은 변동 추세를 예측하는 것보단 이에 영향을 미칠 수 있는 요소들을 분석하고 예측하는 것이 중요하다. 또한, 암호화폐 가격에 영향을 미칠 수 있는 요소들을 예측하는 것은 가격과 변동 추세에 관심을 가지는 사용자에게 폭넓은 선택지를 제공할 뿐 아니라 블록체인 네트워크의 단기적인 데이터 흐름에 대한 통찰과 더 나아가 암호화폐와 관련된 서비스 생태계의 생존성에 대해 판단할 수 있는 밑거름을 제공한다.

[3] 종래기술에서는 다양한 암호화폐 중 가장 오래되었고 아직도 인기를 끌고 있는 비트코인을 분석하였다. 비트코인은 2009년 사토시 나카모토라는 가명의 개발자에 의해 개발되었으며 Peer-to-Peer(P2P) 기반 분산 데이터베이스 형태로 거래를 수행한다. 비트코인의 사용자인 노드들은 공개 키 암호 방식으로 거래를 수행하며 거래들은 작업증명(PoW)을 통해 블록에 담기게 되어 모든 네트워크로 전파된다. 한번 블록에 담겨 블록체인에 연결된 거래들은 임의로 변주할 수 없고 거래를 수행하는 사용자 주소는 익명성을 지니기 때문에 누가 어떤 거래를 했는지 특정할 수 없다. 비트코인과 관련된 다양한 서비스로는 비트코인 화폐를 거래할 수 있는 거래소, 전문적으로 암호화폐를 채굴하기 위한 일종의 조합인 마이닝풀이 있으며 그 외 전자지갑, 도박, 커뮤니티, 소셜 미디어 등이 있고 이를 비트코인 에코시스템이라고 정의한다.

[4] 비트코인 관련 데이터를 예측하는 종래기술은 데이터 셋과 예측 방법을 통해 분류할 수 있다. 먼저, 기존 주식시장에서 사용되는 통계 기반 예측 모델을

암호화폐 분야에 적용한 종래기술들이 있다. 기존 통계 기반 예측에서는 대부분 시장 데이터를 통해 가격 변화율을 예측하였다. 다음으로, 기계학습(딥러닝 제외)를 통해 가격을 예측하고자 하는 종래기술들은 Light GBM, RF 등 트리 기반 예측모델과 SVM 등을 사용했다. 딥러닝을 사용해 가격을 예측하고자 했던 종래기술들은 시계열 기반의 모델들인 LSTM과 GRU 등이 사용되었으며, LSTM과 GRU를 비교했을 때 GRU가 더 높은 성능을 나타낸다고 보고했다. 기존 통계 기반 예측 종래기술에서는 대부분 암호화폐 시장의 데이터를 사용하거나 주가 지수 등 다른 시장의 데이터를 혼합한 데이터를 사용했다. 기계 학습 및 딥러닝을 사용하여 가격 예측을 수행했던 종래기술에서는 시장 데이터와 Blockchain.com에서 제공하는 블록 데이터를 혼합하여 가격 예측을 수행했으며 Reddit, Twitter와 같은 소셜 미디어로부터 얻은 텍스트를 감성 분석을 통해 수치화하여 예측에 사용한 연구도 있다. 특히, 시장 데이터, 블록 데이터, 다른 시장 데이터, 다른 암호화폐 데이터 등 다양한 데이터를 혼합하고 이를 딥러닝 기반 시계열 예측 모델들을 통해 예측을 수행한 종래기술이 있다. 관련 종래기술은 다른 서비스의 데이터들이 비트코인 관련 데이터과 깊은 연관을 가진다는 것을 시사한다.

발명의 상세한 설명

기술적 과제

[5] 본 발명이 이루고자 하는 기술적 과제는 다양한 서비스들로부터 데이터를 수집하고 이를 딥러닝 기반 시계열 예측 모델 중 높은 예측 성능을 나타내는 LSTM, GRU 모델에 학습시켜 높은 예측 성능을 달성하기 위한 방법 및 장치를 제공하는데 있다.

기술적 해결방법

[6] 일 측면에 있어서, 본 발명에서 제안하는 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템은 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 데이터 수집 모듈, 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 전처리 모듈, 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 학습 모듈 및 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 예측 모듈을 포함한다.

[7] 데이터 수집 모듈은 블록의 헤더 부분인 블록 레벨 데이터, 실제 거래 데이터를 포함하는 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터를 포함하고, 블록체인 네트워크 내 거래 정보를 담고 있는

블록 데이터를 수집하는 블록 데이터 수집기, 블록체인 플랫폼을 통해 구현된 암호화폐로서 현금과의 환율(Exchange Rate)을 의미하는 가격 데이터를 수집하는 가격 데이터 수집기, 텍스트 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 텍스트 소셜 미디어 데이터를 수집하는 텍스트 소셜 미디어 데이터 수집기 및 시계열 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 시계열 소셜 미디어 데이터를 수집하는 시계열 소셜 미디어 데이터 수집기를 포함한다.

[8] 전처리 모듈은 블록 데이터 수집기로부터 수집된 블록 데이터의 입출력 레벨 데이터로부터 통계 데이터를 추출하여 트랜잭션 레벨 데이터와 병합하고, 트랜잭션 레벨 데이터로부터 통계 데이터를 추출하여 블록 레벨 데이터와 병합함으로써 블록 높이의 데이터 형식을 갖는 통계 데이터를 생성하는 통계 기반 특징 추출기, 텍스트 소셜 미디어 데이터 수집기로부터 수집된 텍스트 소셜 미디어 데이터에 대한 감성 분석을 통해 긍정적, 부정적 또는 중립적 감정으로 분류하여 감성분석 데이터를 생성하는 감성분석 기반 특징 추출기, 데이터들의 형식을 시계열 데이터로 통일 시키기 위해 통계 기반 특징 추출기로부터 획득된 통계 데이터에 대하여 미리 정해진 시간 내에 포함된 모든 통계 데이터의 블록 높이에 평균을 취함으로써 통계 데이터의 데이터의 형식을 블록 높이에서 시간 단위로 변환하고, 감성분석 기반 특징 추출기로부터 획득된 감성분석 데이터에 관한 복수의 시계열 데이터를 생성하여 통계 기반 특징 추출기, 가격 데이터 수집기, 감성분석 기반 특징 추출기 및 시계열 소셜 미디어 데이터 수집기로부터 획득된 데이터들을 시계열 데이터로 병합하는 시간 단위 변환 및 데이터 병합기, 시간 단위 변환 및 데이터 병합기로부터 획득된 시계열 데이터들에 대하여 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 데이터 간의 상관 계수를 계산하고 시계열 데이터들의 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 시계열 군집화기 및 시계열 군집화기로부터 획득된 시계열 데이터 셋에 대하여 예측 모델에 입력되어야 할 시계열 데이터 셋을 분류하는 시계열 분류기를 포함한다.

[9] 학습 모듈은 LSTM(Long short-term memory) 또는 GRU(Gated Recurrent Unit)를 사용하여 시계열 데이터의 시퀀스 길이, 학습 횟수를 포함하는 학습 변수를 변경하며, 전처리 모듈에서 군집화된 각각의 시계열 데이터 셋에 대응하는 예측 모델을 생성한다.

[10] 예측 모듈은 학습 모듈로부터 획득된 복수의 예측 모델에 전처리 모듈의 시계열 분류기로부터 획득된 시계열 데이터 셋의 학습 데이터셋, 검증 데이터셋, 테스트 데이터셋을 입력하여 복수의 예측 모델을 평가하고, 획득된 시계열 데이터 셋과 가장 가까운 클러스터 중심점을 검색하여 데이터를 입력할 예측 모델을 선택한다.

[11] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 방법은 데이터 수집 모듈이 블록

데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 단계, 전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 단계, 학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 단계 및 예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 단계를 포함한다.

발명의 효과

[12] 본 발명의 실시예들에 따르면 다양한 요소들의 영향을 받고 있는 비트코인의 가격 예측을 직접적인 가격 예측보단 가격 변동에 영향을 미칠 수 있는 요소들을 분석하여 예측할 수 있다. 이러한 요소들을 예측하는 것은 단기적인 비트코인 네트워크의 데이터 흐름에 대한 통찰을 제공할 수 있으며 더 나아가 비트코인 생태계 내 다양한 서비스들의 생존성을 판단할 수 있도록 도울 수 있다. 본 발명에서는 트랜잭션 개수, 수수료의 비율, 비트코인 거래량 등의 블록 데이터를 예측하고, 종래기술에서 고려되었던 데이터를 포함하여 세 가지 클래스의 데이터들을 수집하고 이를 시계열 예측에 뛰어난 성능을 보이는 장단기 메모리(LSTM)를 통해 학습하여 높은 예측 정확도를 달성할 수 있다.

도면의 간단한 설명

[13] 도 1은 본 발명의 일 실시예에 따른 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템의 구성을 나타내는 도면이다.

[14] 도 2는 본 발명의 일 실시예에 따른 데이터 수집 모듈의 구성을 나타내는 도면이다.

[15] 도 3은 본 발명의 일 실시예에 따른 블록 데이터의 특징 데이터 추출 과정을 설명하기 위한 도면이다.

[16] 도 4는 본 발명의 일 실시예에 따른 텍스트 소셜 미디어 데이터의 특징 데이터 추출 과정을 설명하기 위한 도면이다.

[17] 도 5는 본 발명의 일 실시예에 따른 블록 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[18] 도 6은 본 발명의 일 실시예에 따른 텍스트 소셜 미디어 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[19] 도 7은 본 발명의 일 실시예에 따른 가격 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[20] 도 8은 본 발명의 일 실시예에 따른 전체 클래스 데이터를 병합하는 과정을 설명하기 위한 도면이다.

[21] 도 9는 본 발명의 일 실시예에 따른 시계열 데이터 셋으로 군집화하는 과정을

설명하기 위한 도면이다.

[22] 도 10은 본 발명의 일 실시예에 따른 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 과정을 설명하기 위한 도면이다.

[23] 도 11은 본 발명의 일 실시예에 따른 새로운 데이터가 입력될 경우 예측 모델을 선택하는 과정을 설명하기 위한 도면이다.

[24] 도 12는 본 발명의 일 실시예에 따른 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 방법을 설명하기 위한 흐름도이다.

발명의 실시를 위한 형태

[25] 비트코인의 가격은 다양한 요소들의 영향을 받고 있기에 직접적인 가격 예측보단 가격 변동에 영향을 미칠 수 있는 요소들을 분석하는 것이 중요하다. 또한, 이러한 요소들을 예측하는 것은 단기적인 비트코인 네트워크의 데이터 흐름에 대한 통찰을 제공할 수 있으며 더 나아가 비트코인 생태계 내 다양한 서비스들의 생존성을 판단할 수 있도록 도울 수 있다. 본 발명에서는 트랜잭션 개수, 수수료의 비율, 비트코인 거래량 등의 블록 데이터를 예측하고자 한다. 종래기술에서 고려되었던 데이터를 포함하여 세 가지 클래스의 데이터들을 수집하고 이를 시계열 예측에 뛰어난 성능을 보이는 장단기 메모리(LSTM)를 통해 학습하여 높은 예측 정확도를 달성하였다.

[26] 본 발명에서는 비트코인 에코시스템에서 수집할 수 있는 데이터를 크게 3가지 클래스로 분류한다. 첫 번째 클래스는 비트코인 블록체인의 블록에 담긴 데이터이며 비트코인 네트워크 내 데이터의 흐름을 분석할 수 있다. 두 번째 클래스는 비트코인과 관련된 키워드를 언급한 소셜 미디어 데이터이며 비트코인에 대한 다양한 의견을 수집할 수 있는 트위터, 전 세계의 비트코인과 관련된 검색어의 인기를 분석하는 구글 트렌드 등이 있다. 마지막으로 비트코인 거래소에서 얻을 수 있는 가격 시계열 및 거래량 시계열 데이터이다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[27]

[28] 도 1은 본 발명의 일 실시예에 따른 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템의 구성을 나타내는 도면이다.

[29] 제안하는 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 시스템은 데이터 수집 모듈(110), 전처리 모듈(120), 학습 모듈(130) 및 예측 모듈(140)을 포함한다.

[30]

[31] 도 2는 본 발명의 일 실시예에 따른 데이터 수집 모듈의 구성을 나타내는 도면이다.

[32] 데이터 수집 모듈(110)은 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집한다.

[33] 데이터 수집 모듈(110)은 블록 데이터 수집기(111), 가격 데이터 수집기(112),

텍스트 소셜 미디어 데이터 수집기(113) 및 시계열 소셜 미디어 데이터 수집기(114)를 포함한다.

[34] 블록 데이터 수집기(111)는 블록의 헤더 부분인 블록 레벨 데이터, 실제 거래 데이터를 포함하는 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터를 포함하고, 블록체인 네트워크 내 거래 정보를 담고 있는 블록 데이터를 수집한다. 블록 데이터 수집기(111)는 블록체인 네트워크에서 블록 데이터를 수집하고, 데이터 형식은 JSON이며, 데이터 주기는 블록 높이이다.

[35] 본 발명의 실시예에 따르면, 비트코인 블록 데이터를 수집하기 위해 비트코인-노드(Bitcoind-node)를 설치하였으며 개발 환경에 상관 없이 도커(Docker) 기반으로 쉽게 설치할 수 있다. 노드와 RPC를 통해 통신할 수 있으며 getblockhash 명령어를 통해 특정 높이의 블록의 해쉬(hash) 값을 얻을 수 있고 getblock [hash] 명령어를 통해 해당 높이의 블록에 대한 데이터를 얻을 수 있다. 비트코인의 블록은 작업증명을 통해 생성되는 트랜잭션들의 집합이며 정성적인 데이터와 정량적인 데이터로 구성되어 있다. 정성적인 데이터는 블록의 해시, 생성 시간, 난스, 비트 등을 포함하며 정량적인 데이터는 블록의 크기, 가중치, 거래량, 트랜잭션의 개수 등을 포함한다. 또한, 블록 데이터는 세 가지 레벨로 분류할 수 있으며 블록의 헤더 부분인 블록 레벨 데이터, 실제 거래를 포함한 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터로 분류할 수 있다. 블록 데이터는 계층적인 구조로서 블록 레벨은 트랜잭션 레벨의 데이터를 포함하고 마찬가지로 트랜잭션 레벨은 입출력 레벨의 데이터를 포함한다.

[36] 가격 데이터 수집기(112)는 블록체인 플랫폼을 통해 구현된 암호화폐로서 현금과의 환율(Exchange Rate)을 의미하는 가격 데이터를 수집한다. 가격 데이터 수집기(112)는 암호화폐 거래소에서 가격 데이터를 수집하고, 데이터 형식은 시계열 데이터 이며, 데이터 주기는 분, 시간, 일이다.

[37] 본 발명의 실시예에 따른 가격 데이터 중 과거의 데이터는 캐글(Kaggle)에 등록된 비트코인 히스토리 데이터(Bitcoin Historical Data)이며 실시간으로 받아오는 데이터는 Blockchain.com으로부터 수집한다. 데이터는 분당 비트코인 가격을 나타내며 2012년 1월 1일부터 데이터를 수집하였고 각 행은 타임스탬프와 시작점, 고점, 저점, 끝점, 거래량 그리고 거래량에 대한 가격 가중 평균으로 구성되어 있으며 본 발명에서는 가격 가중 평균과 거래량을 사용한다. 수집한 데이터는 10분 단위의 가격 데이터를 포함하며 1시간 평균으로 변환하여 사용한다. 수집한 데이터는 표 1과 같다. 에 대한 분석을 수행하기 위한 텍스트 소셜 미디어 데이터를 수집한다. 텍스트 소셜 미디어 데이터 수집기(113)는 텍스트 소셜 미디어(예를 들어, 트위터)에서 텍스트 소셜 미디어 데이터를 수집하고, 데이터 형식은 텍스트이며, 데이터 주기는 없다. 본 발명의 실시예에 따른 텍스트 소셜 미디어 데이터를 수집하는 플랫폼은 트위터이다.

[38] [표1]

날짜	년-월-일 시
시간당 가격 가중 평균	해당 시간에 거래된 BTC의 평균 가격
거래량	해당 시간에 거래된 BTC양

[39] 트위터는 2006년 출시 된 이후 현재까지도 많은 인기를 끌고 있으며 3억 개의 활성화된 계정이 있으며 매일 5억 개의 트윗이 발생한다. 텍스트 소셜 미디어 데이터(예를 들어, 트위터 데이터)를 사람들이 특정 주제에 대해 어떻게 느끼는지에 대해 분석할 수 있는 풍부한 데이터 소스로 사용할 수 있으며 각 텍스트의 발생 시점을 볼 수 있으므로 시간이 지남에 따라 특정 주제에 대한 사용자들의 감정 변화를 관찰할 수 있다.

[40] 본 발명의 실시예에 따르면, 텍스트 API(Application Programming Interface)를 이용한 파이선(Python) 기반 텍스트 소셜 미디어 데이터 수집기(113)를 통해 실시간으로 "비트코인" 키워드에 대한 텍스트 소셜 미디어 데이터를 수집하는 모듈을 구현하였다.

[41] 시계열 소셜 미디어 데이터 수집기(114)는 시계열 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 시계열 소셜 미디어 데이터를 수집한다. 시계열 소셜 미디어 데이터 수집기(114)는 시계열 소셜 미디어에서 시계열 소셜 미디어 데이터를 수집하고, 데이터 형식은 시계열이며, 데이터 주기는 시이다.

[42] 본 발명의 실시예에 따른 시계열 소셜 미디어 데이터를 수집한 플랫폼은 구글 트렌드이다. 구글 트렌드는 다양한 지역 및 언어에서 구글 검색의 인기 검색어들의 인기를 분석하는 구글 플랫폼이다. 구글 트렌드는 그래프를 사용하여 시간에 따른 여러 검색어의 검색량을 비교한다. 구글 트렌드의 인기 수치는 0부터 100까지의 값으로 설정되어 있으며 기간 내 상대적인 인기 수치를 나타낸다.

[43] 본 발명의 실시예에 따르면, 검색 키워드별 시계열 소셜 미디어 데이터(예를 들어, 구글 트렌드)를 수집할 수 있는 파이선(Python) 기반의 수집 모듈을 구현하였으며 검색 키워드를 Blockchain, BTC, Bitcoin, Mining Pool, Cryptocurrency Exchange로 설정하고 수집하였다. 전체 기간에 대한 시간 별 데이터가 제공되지 않으므로 시간 별 데이터를 수집하기 위하여 1주일 간격으로 데이터를 수집한다. 수집된 데이터는 앞서 언급했듯이 기간 내 상대적인 수치이므로 전체 기간에 대한 절대적인 수치로 변환하는 과정이 필요하다. 따라서, 수집하는 단계에서 수집 간격을 정확히 1주일 간격으로 자르는 것이 전 주에 대한 데이터와 다음 주에 대한 데이터 사이에 일정한 시간이 겹치도록 수집하고 겹치는 시간을 이용하여 데이터를 스케일링한다. 본 발명의 실시예에 따르면 전 주와 다음 주에 사이에 1일이 겹치도록 수정하여 모든 데이터를 스케일링하였고 스케일링한 데이터는 표 2와 같다.

[44] [표2]

날짜	년-월-일 시
Blockchain	키워드 별 인기도
BTC	키워드 별 인기도
Bitcoin	키워드 별 인기도
Mining Pool	키워드 별 인기도
Cryptocurrency Exchange	키워드 별 인기도

[45] 전처리 모듈(120)은 데이터 수집 모듈(110)에서 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화한다. 전처리 모듈(120)은 통계 기반 특징 추출기(121), 감성분석 기반 특징 추출기(122), 시간 단위 변환 및 데이터 병합기(123), 시계열 군집화기(124) 및 시계열 분류기(125)를 포함한다.

[46]

[47] 도 3은 본 발명의 일 실시예에 따른 블록 데이터의 특징 데이터 추출 과정을 설명하기 위한 도면이다.

[48] 도 3a는 본 발명의 일 실시예에 따른 블록 통계 데이터 추출표를 나타내고, 도 3b는 본 발명의 일 실시예에 따른 블록 데이터 변환과정을 나타낸다.

[49] 통계 기반 특징 추출기(121)는 JSON 형태의 블록 데이터로부터 통계 기반 특징 시계열을 추출한다.

[50] 통계 기반 특징 추출기(121)는 블록 데이터 수집기로부터 수집된 블록 데이터의 입출력 레벨 데이터로부터 통계 데이터를 추출하여 트랜잭션 레벨 데이터와 병합하고, 트랜잭션 레벨 데이터로부터 통계 데이터를 추출하여 블록 레벨 데이터와 병합함으로써 블록 높이의 데이터 형식을 갖는 통계 데이터를 생성한다.

[51] 수집된 블록 데이터의 최하위 입출력 레벨로부터 바텀-업 방식으로 통계 데이터를 추출한다. 각 트랜잭션 내부에 포함된 다수의 입출력 레벨 데이터들로부터 통계 데이터를 추출하고 이는 트랜잭션 레벨의 데이터와 병합된다. 마찬가지로 각 블록 내부에 포함된 다수의 트랜잭션 레벨 데이터들로부터 통계 데이터를 추출하고 이는 블록 레벨의 데이터와 병합된다. 이 과정은 총 312개의 통계 데이터를 추출하며 시간 단위는 블록의 높이이다. 본 발명에서는 다른 클래스의 데이터와 시간 단위를 맞추기 위하여 1시간 내 포함된 모든 블록 높이의 데이터에 평균을 취함으로써 통계 데이터의 단위를 블록 높이에서 시 단위로 변환하였다. 추출과정에 대한 개요와 각 레벨의 데이터 개수는 도 3a에 나타나 있으며 통계 데이터의 형식은 표 3과 같다.

[52]

[표3]

날짜	년-월-일 시
통계 데이터#1	실수 값
통계 데이터#2	실수 값
...	...
통계 데이터#3	실수 값

[53] 도 3a를 참조하면, 1차와 2차 통계 추출에서 추출하는 통계의 항목은 동일하다. 블록 레벨 데이터는 추출과정을 거치지 않고 총 4(4*1)개의 특징으로 변환되며 트랜잭션 레벨 데이터는 2차 통계 추출과정만을 거쳐 총 66(6*11)개의 특징으로 변환된다. 입출력 레벨 데이터는 1차 통계 추출과정과 2차 통계 추출과정 모두를 거쳐 총 242(2*11*11)개의 특징으로 변환된다.

[54] 도 4는 본 발명의 일 실시예에 따른 텍스트 소셜 미디어 데이터의 특징 데이터 추출 과정을 설명하기 위한 도면이다.

[55] 도 4a는 본 발명의 일 실시예에 따른 감성 분석 예시를 나타내고, 도 4b는 본 발명의 일 실시예에 따른 감성 분석 결과를 나타낸다.

[56] 감성분석 기반 특징 추출기(122)는 텍스트 형태의 텍스트 소셜 미디어 데이터로부터 감성분석 기반 특징 시계열 데이터를 추출한다.

[57] 감성분석 기반 특징 추출기(122)는 텍스트 소셜 미디어 데이터 수집기로부터 수집된 텍스트 소셜 미디어 데이터에 대한 감성 분석을 통해 긍정적, 부정적 또는 중립적 감정으로 분류하여 감성분석 데이터를 생성한다.

[58] 수집된 텍스트 소셜 미디어 데이터들은 파이선의 TextBlob 패키지 기반 감성 분석 모듈을 통해 긍정적, 부정적 또는 중립적 감정으로 분류되며 감성 분석의 결과로 각 텍스트 소셜 미디어 데이터의 텍스트는 -1과 1 사이의 실수 값이 할당되며 텍스트 소셜 미디어 데이터 내 문자열들의 감성 수치가 0이 아닌 양의 값을 가지면 해당 데이터는 긍정적이라고 평가된다. 마찬가지로 텍스트 소셜 미디어 데이터 내 문자열들의 감성 수치가 0이 아닌 음의 값을 가지면 해당 데이터는 부정적이라고 평가되며 0의 값을 가지면 중립적이라고 평가된다. 결과적으로 감성 분석 모듈은 수집된 텍스트 소셜 미디어 데이터에 대하여 감성 분석을 수행하고 블록 데이터 예측에 적합한 6개의 시계열 데이터를 생성한다. 감성분석이 수행된 데이터는 표 4와 같다.

[59]

[표4]

날짜	년-월-일 시
긍정적 트윗 개수	해당 시간에 발생한 모든 긍정적 트윗 개수의 합
부정적 트윗 개수	해당 시간에 발생한 모든 부정적 트윗 개수의 합
중립적 트윗 개수	해당 시간에 발생한 모든 중립적 트윗 개수의 합
긍정 세기	해당 시간에 발생한 모든 긍정적 트윗의 감성 수치 합
부정 세기	해당 시간에 발생한 모든 부정적 트윗의 감성 수치 합
긍정 부정 세기 비율	해당 시간에 발생한 모든 긍정적 트윗의 감성 세기와 부정적 트윗의 감성 세기의 비율

[60]

[61] 시간 단위 변환 및 데이터 병합기(123)는 데이터들의 형식을 시계열 데이터로 통일 시키기 위해 통계 기반 특징 추출기(121)로부터 획득된 통계 데이터에 대하여 미리 정해진 시간 내에 포함된 모든 통계 데이터의 블록 높이에 평균을 취함으로써 통계 데이터의 데이터의 형식을 블록 높이에서 시간 단위로 변환한다. 그리고, 감성분석 기반 특징 추출기(122)로부터 획득된 감성분석 데이터에 관한 복수의 시계열 데이터를 생성하여 통계 기반 특징 추출기(121), 가격 데이터 수집기(112), 감성분석 기반 특징 추출기(122) 및 시계열 소셜 미디어 데이터 수집기(114)로부터 획득된 데이터들을 시계열 데이터로 병합한다.

[62] 다시 말해, 시간 단위 변환 및 데이터 병합기(123)는 예측 모델에 입력하기 위해 모든 데이터의 시간 단위를 통일한다. 예를 들어, 블록 데이터는 블록 높이 단위를 시 단위로 변환하고, 텍스트 소셜 미디어 데이터는 텍스트 형태의 텍스트 소셜 미디어 데이터로부터 감성분석 기반 특징 시계열 데이터를 추출하며, 가격 데이터는 분 단위를 시로 변환하여 모든 데이터의 시간 단위를 통일한다.

[63]

[64] 도 5는 본 발명의 일 실시예에 따른 블록 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[65] 도 5a는 본 발명의 일 실시예에 따른 시간 단위 변환 및 블록 통계 데이터를 설명하기 위한 도면이고, 도 5b는 본 발명의 일 실시예에 따른 시간 단위 변환,

블록 통계 데이터 및 데이터 관점을 설명하기 위한 도면이다.

[66] 블록 데이터의 단위는 블록 높이이며 이를 단위 시간 내 포함된 블록 통계 데이터의 평균으로 변환한다. 예를 들어, 블록의 평균 생성 주기가 10분일 경우 시간당 평균 6개를 생성할 수 있다.

[67]

[68] 도 6은 본 발명의 일 실시예에 따른 텍스트 소셜 미디어 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[69] 도 6a는 본 발명의 일 실시예에 따른 시간 단위 변환 및 텍스트 소셜 미디어 데이터를 설명하기 위한 도면이고, 도 6b는 본 발명의 일 실시예에 따른 시간 단위 변환, 텍스트 소셜 미디어 데이터 및 데이터 관점을 설명하기 위한 도면이다.

[70] 텍스트 소셜 미디어 데이터의 단위는 없으며 이를 단위 시간 내 발생한 텍스트 소셜 미디어 데이터의 감성 수치를 세거나 합하여 6개의 특징 데이터를 추출한다.

[71]

[72] 도 7은 본 발명의 일 실시예에 따른 가격 데이터의 데이터 형식 변환 과정을 설명하기 위한 도면이다.

[73] 도 7a는 본 발명의 일 실시예에 따른 시간 단위 변환 및 가격 데이터를 설명하기 위한 도면이고, 도 7b는 본 발명의 일 실시예에 따른 시간 단위 변환, 가격 데이터 및 데이터 관점을 설명하기 위한 도면이다.

[74] 가격 데이터의 단위는 분이며 이를 단위 시간 내 가격과 거래량의 평균을 취함으로써 단위를 시로 변환한다.

[75]

[76] 도 8은 본 발명의 일 실시예에 따른 전체 클래스 데이터를 병합하는 과정을 설명하기 위한 도면이다.

[77] 시간 단위 변환 및 데이터 병합기(123)는 시간 단위 변환 및 데이터 병합기(123)를 통해 시간 단위가 통일된 블록 통계 데이터(810), 시계열 소셜 미디어 데이터(820), 텍스트 소셜 미디어 데이터(다시 말해, 트윗 특징 데이터)(830) 및 가격 데이터(다시 말해, 거래 데이터)(840)를 병합하여 통합된 시계열 데이터(850)를 생성한다.

[78]

[79] 도 9는 본 발명의 일 실시예에 따른 시계열 데이터 셋으로 군집화하는 과정을 설명하기 위한 도면이다.

[80] 도 9a는 본 발명의 일 실시예에 따른 스피어만 순위 상관 계수의 예시를 나타내는 도면이고, 도 9b는 본 발명의 일 실시예에 따른 시계열 군집화 개요를 나타내는 도면이다.

[81] 시계열 데이터는 다양한 분포 특징을 가질 수 있으며 군집화한 데이터를 학습할 시 예측 성능을 향상 시킬 수 있다.

[82] 시계열 군집화기(124)는 시간 단위 변환 및 데이터 병합기로부터 획득된 시계열 데이터들에 대하여 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 데이터 간의 상관 계수를 계산하고 시계열 데이터들의 분포 특징에 따라 시계열 데이터 셋으로 군집화한다.

[83] 시계열 분류기(125)는 시계열 군집화기로부터 획득된 시계열 데이터 셋에 대하여 예측 모델에 입력되어야 할 시계열 데이터 셋을 분류한다.

[84] DDoS 공격이 발생한 후 대부분의 가격 손실은 이틀 내에 복구되며 5일 내에 복구되지 않는 경우도 있다고 보고되었다. 본 발명에서는 비트코인에 영향을 미치는 긍정적인 또는 부정적인 사건이 계속 발생하고 있으며 이러한 사건에 의해 비트코인 데이터가 영향을 받음에 따라 사건이 발생했을 때와 사건이 발생하지 않았을 때의 시계열 분포가 달라질 것이라고 가정한다. 본 발명에서는 가정에 따라 클러스터링 알고리즘을 통해 수집한 시계열 데이터들을 분포 특징을 고려하여 군집화하고 군집 별로 학습 모델을 생성하여 학습 과정과 테스트 과정의 오차를 최소화하고자 한다. 먼저, 다변량 시계열 데이터를 시퀀스 길이 s로 샘플링한다. d가 특징 벡터의 크기일 때, $a_t=[a_{t1},a_{t2},a_{t3},\cdots,a_{t(d-2)},a_{t(d-1)},a_{t(d)}]$는 특정 시간 t의 특징 벡터를 나타낸다. 그리고 A_j가 j번째 샘플링된 시계열 데이터라고 할 때, $A_j=a_j,a_{j+1},a_{j+2},\cdots,a_{j+(s-1)},a_{j+s}$로 나타낼 수 있다. 결과적으로 전체 데이터의 개수가 L일 때, 샘플링된 시계열 데이터의 개수는 L-s이다.

[85] 다음으로, 앞선 과정에서 얻은 L-s개의 데이터를 클러스터링하여 유사한 분포 특징을 가진 시계열끼리 군집화한다. 본 발명에서는 샘플링된 시계열들을 군집화하기 위하여 K-Medoids 알고리즘을 사용했다. K-Medoids 알고리즘은 거리 기반 비유사성과 같은 비용함수를 최소화하는 방향으로 동작한다. 본 발명에서 K-Medoids 알고리즘은 $(A_1,A_2,A_3,\cdots,A_{(L-s)-s},A_{(L-s)-1},A_{(L-s)})$와 같은 d차원인 시계열 객체가 주어질 때, 시계열 객체들을 $K(\leq(L-s))$개의 집합 $S_1,S_2,\cdots,S_K$로 나눈다. u_i가 집합 S_i의 중심일 때, 이 알고리즘은 집합에서 중심과 각 시계열 객체 사이의 최소 제곱 거리의 집합 S를 찾으며 수식은 다음과 같다:

[86] $$\underset{S}{\operatorname{Argmax}} \sum_{i=1}^{K} \sum_{A \in S_i} \|A - u_i\|^2$$

[87] 일반적으로 K-Medoids 알고리즘은 거리 측정 알고리즘으로 유클리드 거리를 사용한다. 하지만, 시계열들의 분포 특징을 유클리드 거리로 계산하는 것은 적합하지 않다. 본 발명에서는 시계열 간 선형 관계를 측정할 수 있는 피어슨 상관계수와 단조 관계를 측정할 수 있는 스피어만 상관계수를 고려한다. 피어슨 상관계수는 데이터에서 두 변수의 공분산을 표준 편차의 곱으로 나눈 값과

같으며 수식은 다음과 같다:

[88] $$r_{XY} = \frac{\sum_{i}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i}^{n}(Y_i - \overline{Y})^2}}$$

[89] 피어슨 상관계수는 사용하기 이전에 두 변수 중 적어도 하나의 변수는 정규분포여야 한다는 조건을 만족해야 하지만 수집한 대부분의 데이터가 정규성을 만족하지 못하므로 정규성 문제에서 자유로운 스피어만 순위 상관계수를 거리 측정 방법으로 사용한다. 스피어만 순위 상관계수는 순위가 매겨진 변수 간 피어슨 상관계수로 정의되며 X^{rank}가 X의 순위이고 Y^{rank}가 Y의 순위라고 가정할 때 스피어만 순위 상관계수 r^s는 다음과 같다:

[90] $$r^{S}_{X^{rank}Y^{rank}} = \frac{\sum_{i}^{n}(X_i^{rank} - \overline{X^{rank}})(Y_i^{rank} - \overline{Y^{rank}})}{\sqrt{\sum_{i}^{n}(X_i^{rank} - \overline{X}^{rank})^2}\sqrt{\sum_{i}^{n}(Y_i^{rank} - \overline{Y}^{rank})^2}}$$

[91] 스피어만 순위 상관 계수는 -1과 1 사이의 값을 갖고, 계수의 절댓값이 클수록 더 강한 상관 관계를 나타내며, 양의 상관관계와 음의 상관관계를 구분하지 않고 상관관계의 강함에 초점을 맞춘다. 두 데이터 객체 간 측정된 거리를 제곱 기존의 K-Medoids는 거리 측정값이 클수록 각 데이터 객체와의 거리가 멀어지고 이는 거리 측정값이 클수록 각 데이터 객체와의 거리가 가까워지는 스피어만 상관 계수와 반대로 동작한다. 따라서 두 데이터 객체 간 측정된 상관계수에 -1을 곱하여 부호를 전환하고 이 과정을 통해 스피어만 순위 상관계수는 기존의 거리 측정 알고리즘들과 동일하게 동작한다.

[92] 앞서 설명한 K-Medoids에 스피어만 순위 상관계수 기반 거리 측정 알고리즘을 적용했을 때 클러스터 중심과 각 시계열 객체의 최소 거리 집합 S를 찾는 수식은 다음과 같다:

[93] $$\underset{S}{\mathrm{Argmax}} \sum_{i=1}^{K} \sum_{A \in S_i} \sum_{m=1}^{d} r^s (A_m^{rank}, u_{im}^{rank})^2$$

[94] 이때, A_m^{rank}는 순위 시계열 객체 A의 m번째 특징을 나타내며 u_{im}^{rank}는 i번째 클러스터 중심점의 m번째 특징을 나타낸다. 이 클러스터링 알고리즘은 다음과 같이 동작한다:

[95] 1단계: 먼저, 각 클러스터의 초기 중심점을 데이터 객체 내에서 랜덤하게 개의 데이터로 설정한다.

[96] 2단계: 각 데이터 객체와 클러스터 객체 간 스피어만 순위 상관계수를 계산하여 해당 데이터 객체에서 가장 가까운 클러스터를 찾아 데이터를 할당하며 수식은 다음과 같다:

[97] $$S_i = \left\{ A_j : \sum_{m=1}^{d} r^s (A_{jm}^{rank}, u_{jm}^{rank})^2 \geq \sum_{m=1}^{d} r^s (A_{jm}^{rank}, u_{jm}^{rank})^2 \forall L, 1 \leq l \leq k \right\}$$

[98] 3단계: k개의 클러스터의 중심점 u를 각 클러스터 내 가장 가운데 위치한 객체로 설정하며 수식은 다음과 같다:

[99] $$u_i = \underset{X}{\text{Argmax}} \sum_{A \in S_i} \sum_{Y \in S_i, X \neq} \sum_{Ym=1}^{d} r^s(A_m^{rank}, u_{im}^{rank})^2$$

[100] 4단계: 모든 클러스터가 변하지 않을 때까지 2~3단계를 반복한다.

[101] 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 간 상관 계수를 계산하여 유사한 분포 특징을 가진 시계열들을 클러스터링한다. 클러스터링 결과는 클러스터 개수 k에 따라 다를 수 있기에 본 발명에서는 최적의 군집 수 k를 찾기 위하여 k를 2부터 12까지 변경하여 클러스터링을 수행했다. 또한, 동일 클러스터 개수 k에 대해서도 어떤 객체가 초기 중심점으로 설정되는가에 따라 클러스터링 결과가 달라질 수 있으므로 각 k마다 클러스터링을 5000번 동안 반복적으로 수행하고 그 중 실루엣 계수가 가장 높은 클러스터링 결과를 채택했다.

[102]

[103] 도 10은 본 발명의 일 실시예에 따른 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 과정을 설명하기 위한 도면이다.

[104] 학습 모듈(130)은 전처리 모듈(120)에서 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 예측 모델 생성기(131)를 통해 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성한다.

[105] 본 발명의 실시예에 따른 예측을 위한 모델로 순환신경망(RNN)의 변형인 LSTM(Long short-term memory)과 GRU(Gated Recurrent Unit)를 사용한다. 예측 모델은 파이선(Python)을 기반으로 하여 텐서플로(Tensorflow)의 래퍼(wrapper) 라이브러리인 케라스(Keras)로 구현되었다. LSTM은 순환신경망의 변형으로 RNN의 장기 의존성 문제를 해결했을 뿐만 아니라 학습 또한 빠르게 수렴한다.

[106] GRU는 LSTM의 변형으로 더 간단한 구조를 갖기 때문에 기존 LSTM보다 학습할 가중치가 적다는 이점을 가진다.

[107] 예측 모델의 입력은 다차원 벡터로 이루어진 시퀀스 길이 s의 시계열이며 출력은 다음 시간 t-1의 데이터 변동 방향, 변화율이며 예측 모델 최적화를 위해 시퀀스 길이, 학습 횟수를 포함하는 학습 변수를 변경하며 예측 모델을 생성하였다.

[108] 앞서 전체 시계열 데이터 셋을 클러스터링을 하고 유사한 분포 특징을 지닌 시계열 데이터의 셋을 생성했다. 각 시계열 데이터 셋에 대응하는 예측 모델을 생성하였으며 결과적으로 클러스터 개수 k개의 예측 모델이 생성된다.

[109]

[110] 도 11은 본 발명의 일 실시예에 따른 새로운 데이터가 입력될 경우 예측 모델을 선택하는 과정을 설명하기 위한 도면이다.

[111] 예측 모듈(140)은 학습 모듈(130)에서 학습된 복수의 예측 모델에 대하여

평가하고, 새로운 데이터가 입력될 경우, 데이터 예측기(141)를 통해 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택한다.

[112] 예측 모듈(140)은 학습 모듈(130)에서 생성된 복수의 예측 모델을 평가하는 방법과 새로운 데이터가 입력되었을 때 예측을 수행할 예측 모델을 선택한다. 본 발명의 실시예에 따르면, 군집화된 시계열 데이터 셋은 객관적인 평가를 위하여 6:2:2 비율의 학습 데이터셋, 검증 데이터셋, 테스트 데이터셋으로 분할된다. 클러스터링 결과에 따라 학습된 모델에 검증 데이터셋과 테스트 데이터셋을 입력하여 모델을 평가할 때 데이터 셋의 개별 시계열 데이터는 시계열 분류기를 거쳐 적합한 학습 모델에 입력되며 예측 모듈(140)은 입력된 시계열 데이터와 가장 가까운 클러스터 중심점을 검색하여 데이터를 입력할 예측 모델을 선택한다.

[113]

[114] 도 12는 본 발명의 일 실시예에 따른 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 방법을 설명하기 위한 흐름도이다.

[115] 제안하는 시계열 분포 특징을 고려한 딥러닝 기반 비트코인 블록 데이터 예측 방법은 데이터 수집 모듈이 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 단계(1210), 전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 단계(1220), 학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 단계(1230) 및 예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 단계(1240)를 포함한다.

[116] 단계(1210)에서, 데이터 수집 모듈이 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집한다.

[117] 블록 데이터 수집기가 블록의 헤더 부분인 블록 레벨 데이터, 실제 거래 데이터를 포함하는 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터를 포함하고, 블록체인 네트워크 내 거래 정보를 담고 있는 블록 데이터를 수집한다.

[118] 가격 데이터 수집기가 블록체인 플랫폼을 통해 구현된 암호화폐로서 현금과의 환율(Exchange Rate)을 의미하는 가격 데이터를 수집한다.

[119] 텍스트 소셜 미디어 데이터 수집기가 텍스트 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 텍스트 소셜 미디어 데이터를 수집한다.

[120] 시계열 소셜 미디어 데이터 수집기가 시계열 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 시계열 소셜 미디어 데이터를 수집한다.

[121] 단계(1220)에서, 전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을

시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화한다.

[122] 통계 기반 특징 추출기가 블록 데이터 수집기로부터 수집된 블록 데이터의 입출력 레벨 데이터로부터 통계 데이터를 추출하여 트랜잭션 레벨 데이터와 병합하고, 트랜잭션 레벨 데이터로부터 통계 데이터를 추출하여 블록 레벨 데이터와 병합함으로써 블록 높이의 데이터 형식을 갖는 통계 데이터를 생성한다.

[123] 감성분석 기반 특징 추출기가 텍스트 소셜 미디어 데이터 수집기로부터 수집된 텍스트 소셜 미디어 데이터에 대한 감성 분석을 통해 긍정적, 부정적 또는 중립적 감정으로 분류하여 감성분석 데이터를 생성한다.

[124] 시간 단위 변환 및 데이터 병합기가 데이터들의 형식을 시계열 데이터로 통일 시키기 위해 통계 기반 특징 추출기로부터 획득된 통계 데이터에 대하여 미리 정해진 시간 내에 포함된 모든 통계 데이터의 블록 높이에 평균을 취함으로써 통계 데이터의 데이터의 형식을 블록 높이에서 시간 단위로 변환한다. 그리고, 감성분석 기반 특징 추출기로부터 획득된 감성분석 데이터에 관한 복수의 시계열 데이터를 생성하여 통계 기반 특징 추출기, 가격 데이터 수집기, 감성분석 기반 특징 추출기 및 시계열 소셜 미디어 데이터 수집기로부터 획득된 데이터들을 시계열 데이터로 병합한다.

[125] 시계열 군집화기가 시간 단위 변환 및 데이터 병합기로부터 획득된 시계열 데이터들에 대하여 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 데이터 간의 상관 계수를 계산하고 시계열 데이터들의 분포 특징에 따라 시계열 데이터 셋으로 군집화한다.

[126] 시계열 분류기가 시계열 군집화기로부터 획득된 시계열 데이터 셋에 대하여 예측 모델에 입력되어야 할 시계열 데이터 셋을 분류한다.

[127] 단계(1230)에서, 학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성한다.

[128] LSTM(Long short-term memory) 또는 GRU(Gated Recurrent Unit)를 사용하여 시계열 데이터의 시퀀스 길이, 학습 횟수를 포함하는 학습 변수를 변경하며, 전처리 모듈에서 군집화된 각각의 시계열 데이터 셋에 대응하는 예측 모델을 생성한다.

[129] 단계(1240)에서, 예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택한다.

[130] 학습 모듈로부터 획득된 복수의 예측 모델에 전처리 모듈의 시계열 분류기로부터 획득된 시계열 데이터 셋의 학습 데이터셋, 검증 데이터셋, 테스트

데이터셋을 입력하여 복수의 예측 모델을 평가하고, 획득된 시계열 데이터 셋과 가장 가까운 클러스터 중심점을 검색하여 데이터를 입력할 예측 모델을 선택한다.

[131]

[132] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[133] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[134] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media),

CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[135] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[136] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

청구범위

[청구항 1] 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 데이터 수집 모듈;
수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 전처리 모듈;
군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 학습 모듈; 및
학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 예측 모듈
을 포함하는 딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 2] 제1항에 있어서,
데이터 수집 모듈은,
블록의 헤더 부분인 블록 레벨 데이터, 실제 거래 데이터를 포함하는 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터를 포함하고, 블록체인 네트워크 내 거래 정보를 담고 있는 블록 데이터를 수집하는 블록 데이터 수집기;
블록체인 플랫폼을 통해 구현된 암호화폐로서 현금과의 환율(Exchange Rate)을 의미하는 가격 데이터를 수집하는 가격 데이터 수집기;
텍스트 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 텍스트 소셜 미디어 데이터를 수집하는 텍스트 소셜 미디어 데이터 수집기; 및
시계열 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 시계열 소셜 미디어 데이터를 수집하는 시계열 소셜 미디어 데이터 수집기
를 포함하는 딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 3] 제2항에 있어서,
전처리 모듈은,
블록 데이터 수집기로부터 수집된 블록 데이터의 입출력 레벨 데이터로부터 통계 데이터를 추출하여 트랜잭션 레벨 데이터와 병합하고, 트랜잭션 레벨 데이터로부터 통계 데이터를 추출하여 블록 레벨 데이터와 병합함으로써 블록 높이의 데이터 형식을 갖는 통계 데이터를 생성하는 통계 기반 특징 추출기;
텍스트 소셜 미디어 데이터 수집기로부터 수집된 텍스트 소셜 미디어 데이터에 대한 감성 분석을 통해 긍정적, 부정적 또는 중립적 감정으로 분류하여 감성분석 데이터를 생성하는 감성분석 기반 특징 추출기; 및

데이터들의 형식을 시계열 데이터로 통일 시키기 위해 통계 기반 특징 추출기로부터 획득된 통계 데이터에 대하여 미리 정해진 시간 내에 포함된 모든 통계 데이터의 블록 높이에 평균을 취함으로써 통계 데이터의 데이터의 형식을 블록 높이에서 시간 단위로 변환하고, 감성분석 기반 특징 추출기로부터 획득된 감성분석 데이터에 관한 복수의 시계열 데이터를 생성하여 통계 기반 특징 추출기, 가격 데이터 수집기, 감성분석 기반 특징 추출기 및 시계열 소셜 미디어 데이터 수집기로부터 획득된 데이터들을 시계열 데이터로 병합하는 시간 단위 변환 및 데이터 병합기

를 포함하는 딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 4] 제3항에 있어서,

전처리 모듈은,

시간 단위 변환 및 데이터 병합기로부터 획득된 시계열 데이터들에 대하여 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 데이터 간의 상관 계수를 계산하고 시계열 데이터들의 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 시계열 군집화기; 및

시계열 군집화기로부터 획득된 시계열 데이터 셋에 대하여 예측 모델에 입력되어야 할 시계열 데이터 셋을 분류하는 시계열 분류기

를 포함하는 딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 5] 제1항에 있어서,

학습 모듈은,

LSTM(Long short-term memory) 또는 GRU(Gated Recurrent Unit)를 사용하여 시계열 데이터의 시퀀스 길이, 학습 횟수를 포함하는 학습 변수를 변경하며, 전처리 모듈에서 군집화된 각각의 시계열 데이터 셋에 대응하는 예측 모델을 생성하는

딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 6] 제1항에 있어서,

예측 모듈은,

학습 모듈로부터 획득된 복수의 예측 모델에 전처리 모듈의 시계열 분류기로부터 획득된 시계열 데이터 셋의 학습 데이터셋, 검증 데이터셋, 테스트 데이터셋을 입력하여 복수의 예측 모델을 평가하고, 획득된 시계열 데이터 셋과 가장 가까운 클러스터 중심점을 검색하여 데이터를 입력할 예측 모델을 선택하는

딥러닝 기반 비트코인 블록 데이터 예측 시스템.

[청구항 7] 데이터 수집 모듈이 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 단계;

전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을 시계열

데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 단계;
학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 단계; 및
예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 단계
를 포함하는 딥러닝 기반 비트코인 블록 데이터 예측 방법.

[청구항 8] 제7항에 있어서,
데이터 수집 모듈이 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 단계는,
블록 데이터 수집기가 블록의 헤더 부분인 블록 레벨 데이터, 실제 거래 데이터를 포함하는 트랜잭션 레벨 데이터, 트랜잭션 내부에 포함된 입력과 출력의 집합인 입출력 레벨 데이터를 포함하고, 블록체인 네트워크 내 거래 정보를 담고 있는 블록 데이터를 수집하고;
가격 데이터 수집기가 블록체인 플랫폼을 통해 구현된 암호화폐로서 현금과의 환율(Exchange Rate)을 의미하는 가격 데이터를 수집하고;
텍스트 소셜 미디어 데이터 수집기가 텍스트 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 텍스트 소셜 미디어 데이터를 수집하며;
시계열 소셜 미디어 데이터 수집기가 시계열 소셜 미디어로부터 키워드에 대한 분석을 수행하기 위한 시계열 소셜 미디어 데이터를 수집하는
딥러닝 기반 비트코인 블록 데이터 예측 방법.

[청구항 9] 제8항에 있어서,
전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터 셋으로 군집화하는 단계는,
통계 기반 특징 추출기가 블록 데이터 수집기로부터 수집된 블록 데이터의 입출력 레벨 데이터로부터 통계 데이터를 추출하여 트랜잭션 레벨 데이터와 병합하고, 트랜잭션 레벨 데이터로부터 통계 데이터를 추출하여 블록 레벨 데이터와 병합함으로써 블록 높이의 데이터 형식을 갖는 통계 데이터를 생성하고;
감성분석 기반 특징 추출기가 텍스트 소셜 미디어 데이터 수집기로부터 수집된 텍스트 소셜 미디어 데이터에 대한 감성 분석을 통해 긍정적,

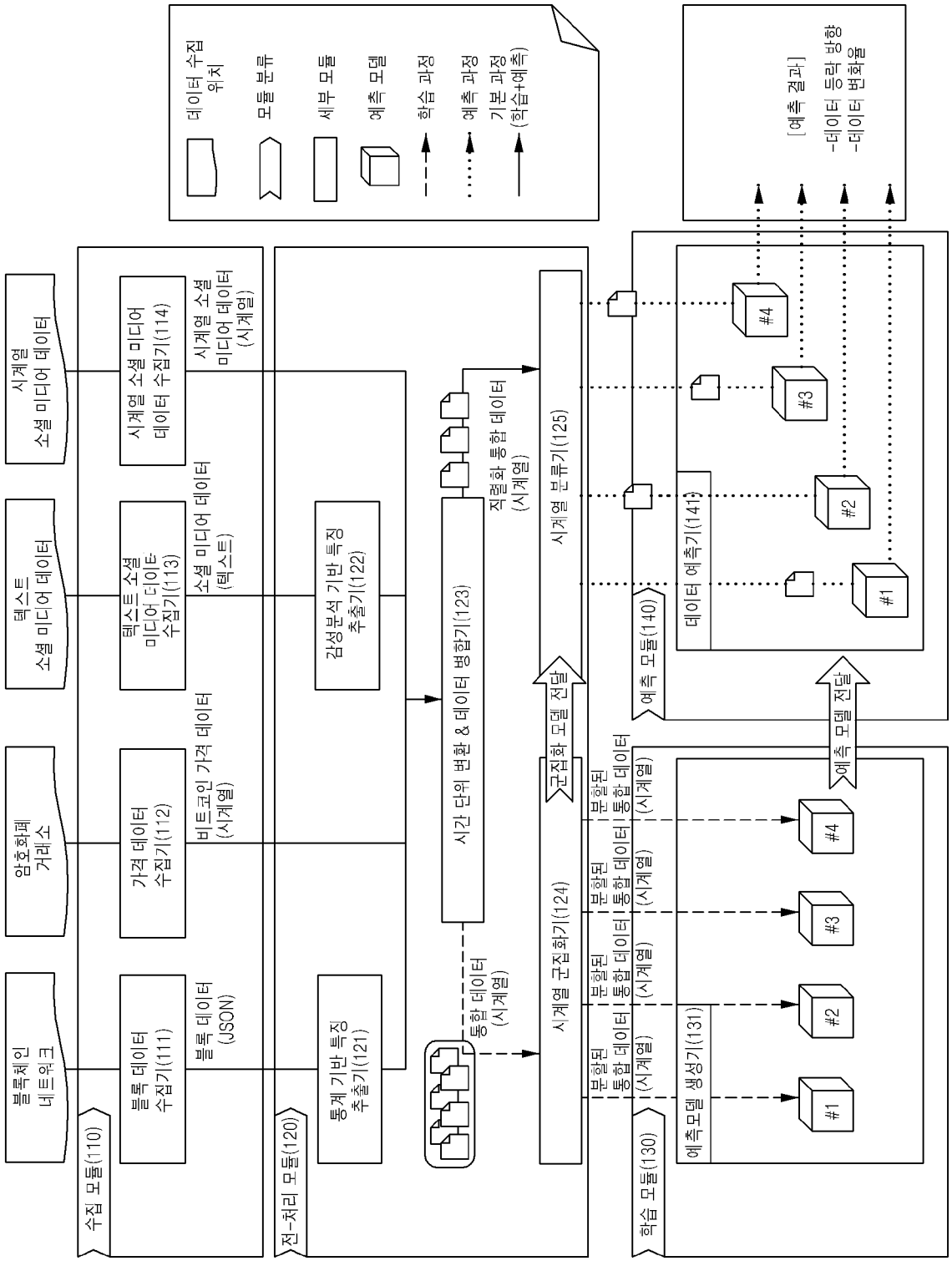
부정적 또는 중립적 감정으로 분류하여 감성분석 데이터를 생성하며;
시간 단위 변환 및 데이터 병합기가 데이터들의 형식을 시계열 데이터로 통일 시키기 위해 통계 기반 특징 추출기로부터 획득된 통계 데이터에 대하여 미리 정해진 시간 내에 포함된 모든 통계 데이터의 블록 높이에 평균을 취함으로써 통계 데이터의 데이터의 형식을 블록 높이에서 시간 단위로 변환하고, 감성분석 기반 특징 추출기로부터 획득된 감성분석 데이터에 관한 복수의 시계열 데이터를 생성하여 통계 기반 특징 추출기, 가격 데이터 수집기, 감성분석 기반 특징 추출기 및 시계열 소셜 미디어 데이터 수집기로부터 획득된 데이터들을 시계열 데이터로 병합하는
딥러닝 기반 비트코인 블록 데이터 예측 방법.

[청구항 10] 제9항에 있어서,
시계열 군집화기가 시간 단위 변환 및 데이터 병합기로부터 획득된 시계열 데이터들에 대하여 스피어만 순위 상관계수를 거리측정 알고리즘으로 사용한 K-Medoids 클러스터링 알고리즘을 통해 시계열 데이터 간의 상관 계수를 계산하고 시계열 데이터들의 분포 특징에 따라 시계열 데이터 셋으로 군집화하고;
시계열 분류기가 시계열 군집화기로부터 획득된 시계열 데이터 셋에 대하여 예측 모델에 입력되어야 할 시계열 데이터 셋을 분류하는
딥러닝 기반 비트코인 블록 데이터 예측 방법.

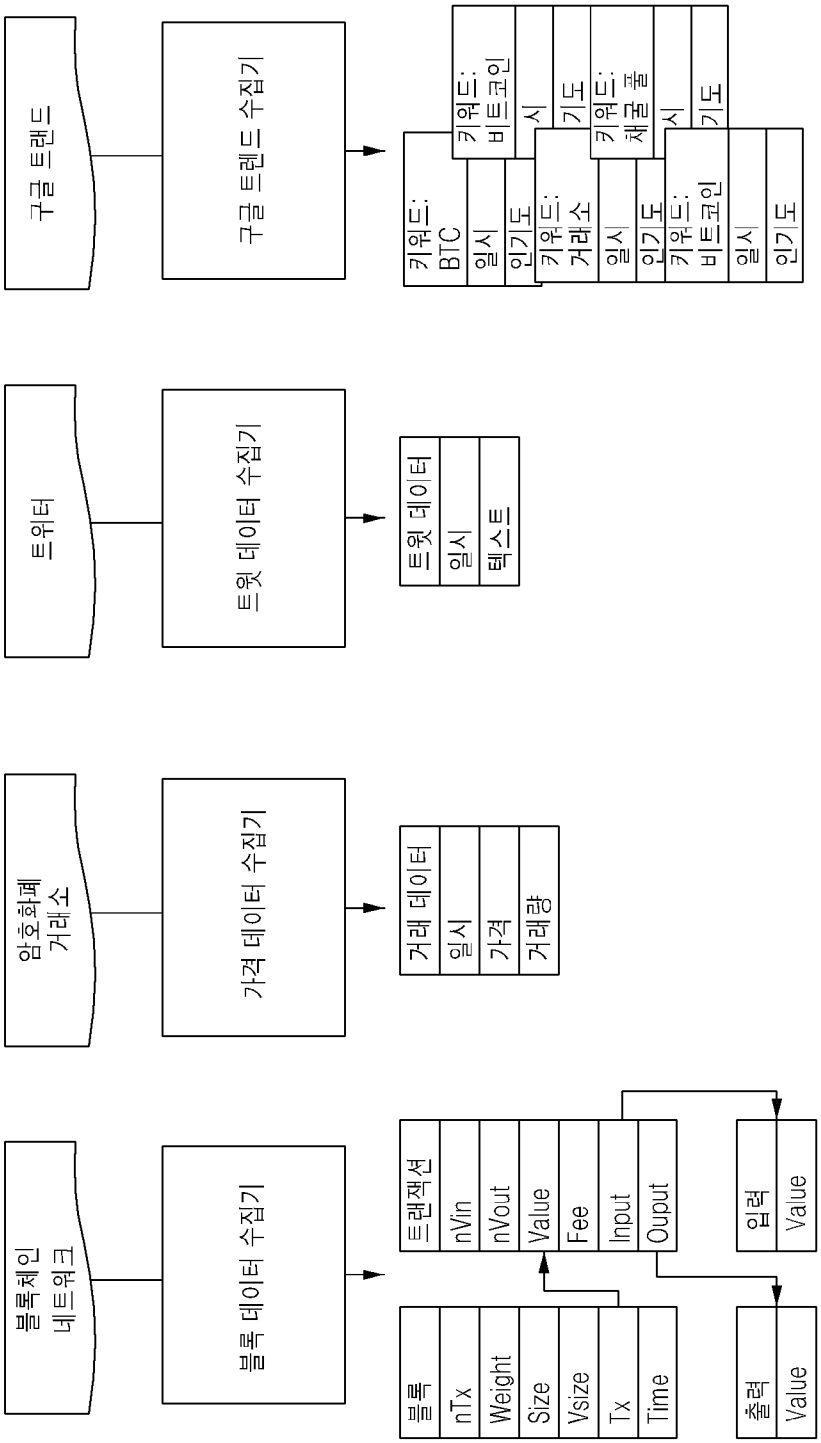
[청구항 11] 제7항에 있어서,
학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 단계는,
LSTM(Long short-term memory) 또는 GRU(Gated Recurrent Unit)를 사용하여 시계열 데이터의 시퀀스 길이, 학습 횟수를 포함하는 학습 변수를 변경하며, 전처리 모듈에서 군집화된 각각의 시계열 데이터 셋에 대응하는 예측 모델을 생성하는
딥러닝 기반 비트코인 블록 데이터 예측 방법.

[청구항 12] 제7항에 있어서,
예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 선택하는 단계는,
학습 모듈로부터 획득된 복수의 예측 모델에 전처리 모듈의 시계열 분류기로부터 획득된 시계열 데이터 셋의 학습 데이터셋, 검증 데이터셋, 테스트 데이터셋을 입력하여 복수의 예측 모델을 평가하고, 획득된 시계열 데이터 셋과 가장 가까운 클러스터 중심점을 검색하여 데이터를 입력할 예측 모델을 선택하는
딥러닝 기반 비트코인 블록 데이터 예측 방법.

[도1]


데이터 수집 위치
모듈 분류
세부 모듈
예측 모델
학습 과정
예측 과정
기본 과정 (학습+예측)
[예측 결과]
-데이터 등락 방향
-데이터 변화율
블록체인 네트워크
암호화폐 거래소
텍스트 소셜 미디어 데이터
시계열 소셜 미디어 데이터
수집 모듈(110)
블록 데이터 수집기(111)
가격 데이터 수집기(112)
텍스트 소셜 미디어 데이터 수집기(113)
시계열 소셜 미디어 데이터 수집기(114)
블록 데이터 (JSON)
비트코인 가격 데이터 (시계열)
소셜 미디어 데이터 (텍스트)
시계열 소셜 미디어 데이터 (시계열)
전-처리 모듈(120)
통계 기반 특징 추출기(121)
감성분석 기반 특징 추출기(122)
통합 데이터 (시계열)
시간 단위 변환 & 데이터 병합기(123)
직렬화 통합 데이터 (시계열)
시계열 군집 화기(124)
군집화 모델 전달
시계열 분류기(125)
학습 모듈(130)
예측모델 생성기(131)
분할된 통합 데이터 (시계열)
#1
#2
#3
#4
예측 모델 전달
예측 모듈(140)
데이터 예측기(141)

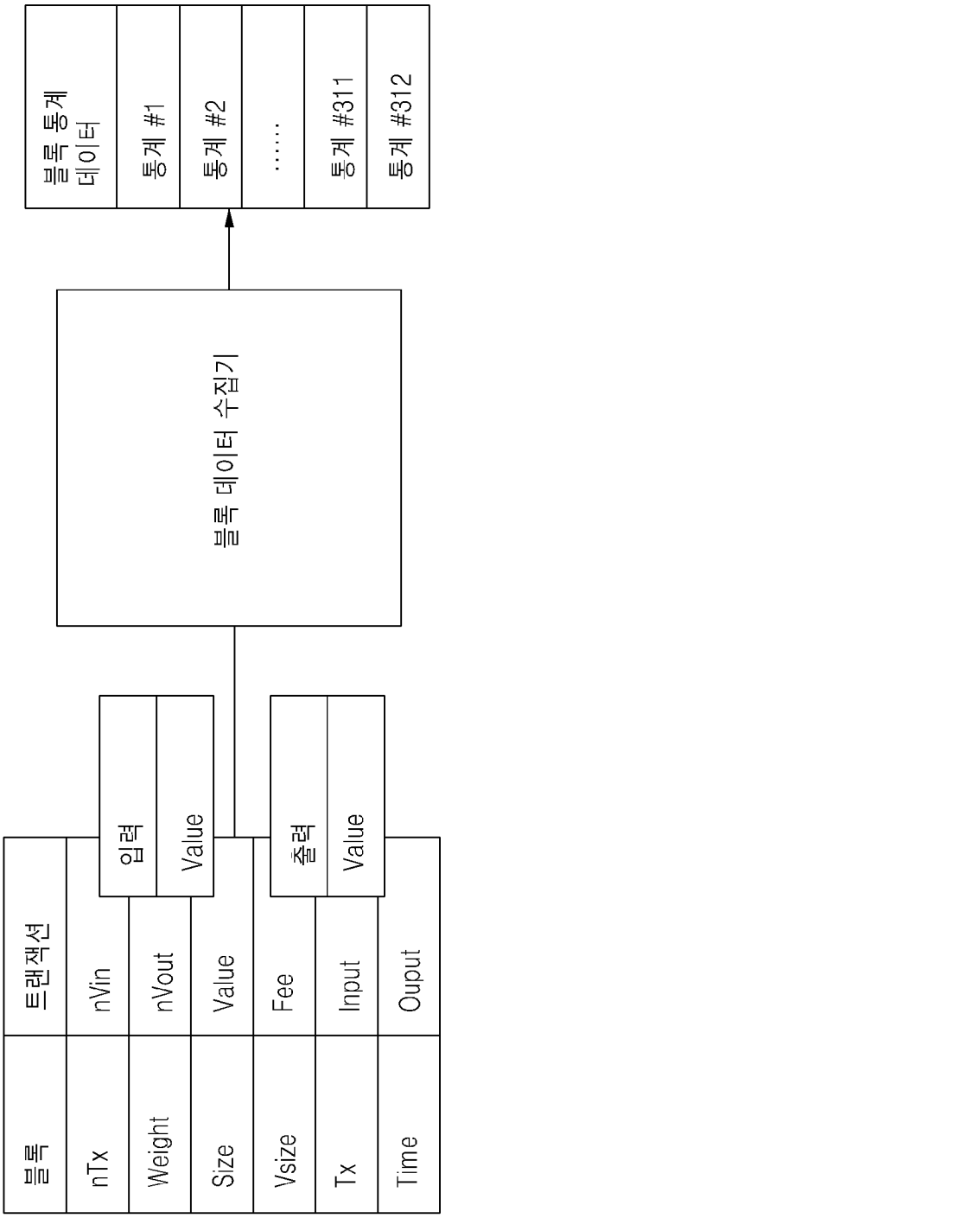
[도2]


구글 트랜드
구글 트렌드 수집기
키워드: BTC
일시
인기도
키워드: 비트코인
시
기도
키워드: 거래소
일시
인기도
키워드: 채굴 풀
시
기도
키워드: 비트코인
일시
인기도
트위터
트윗 데이터 수집기
트윗 데이터
일시
텍스트
암호화폐 거래소
가격 데이터 수집기
거래 데이터
일시
가격
거래량
블록체인 네트워크
블록 데이터 수집기
블록
nTx
Weight
Size
Vsize
Tx
Time
트랜잭션
nVin
nVout
Value
Fee
Input
Ouput
입력
Value
출력
Value

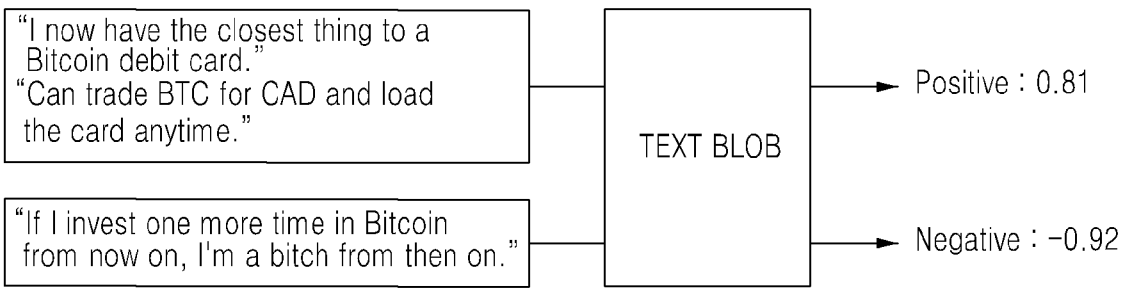
[도3a]

레벨	데이터	1차 추출	2차 추출	데이터 개수
블록	nix			1
	weight			1
	size			1
	vsize			1
트랜잭션	nvin	1차 통계 추출*	2차 통계 추출: summation, maximum, minimum, mean, standard deviation, first quartile(1Q), 2Q, 3Q, inter-quartile range(IQR), skewness, kurtosis	11
	nvout			11
	value			11
	fee			11
	tx_size			11
	tx_vise			11
입출력	vin			121
	vout			121

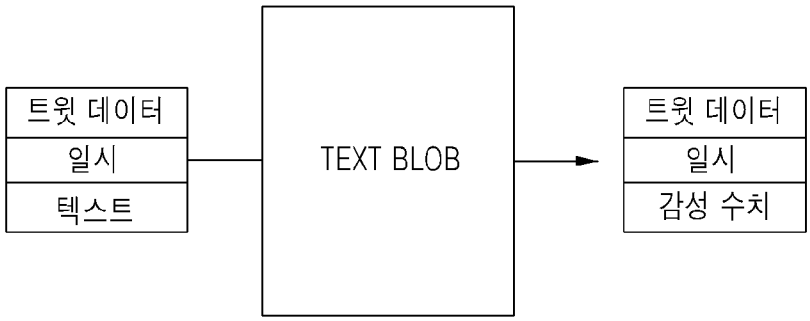
[도3b]


블록 통계 데이터
통계 #1
통계 #2
.....
통계 #311
통계 #312
블록 데이터 수집기
트랜잭션
입력
Value
출력
Value
블록
nTx
Weight
Size
Vsize
Tx
Time
nVin
nVout
Value
Fee
Input
Ouput

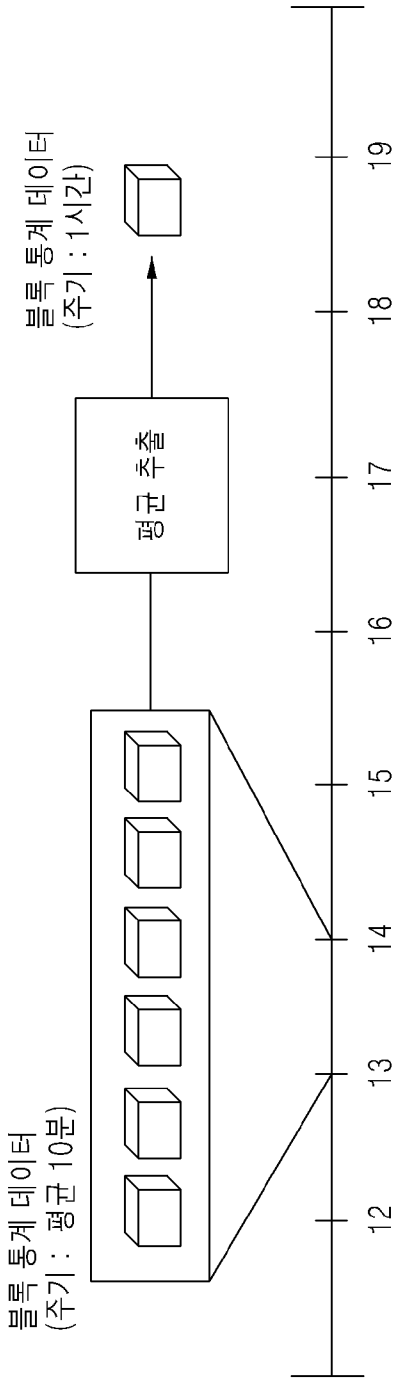
[도4a]


"I now have the closest thing to a Bitcoin debit card."
"Can trade BTC for CAD and load the card anytime."
"If I invest one more time in Bitcoin from now on, I'm a bitch from then on."
TEXT BLOB
Positive : 0.81
Negative : -0.92

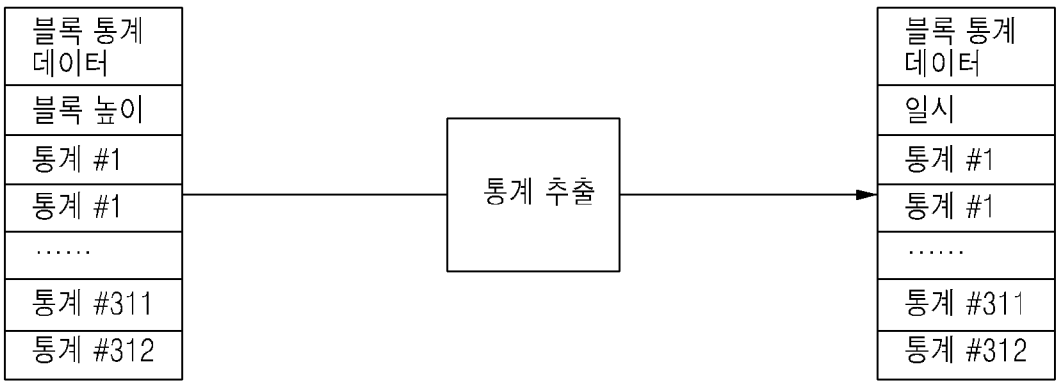
[도4b]


트윗 데이터
일시
텍스트
TEXT BLOB
트윗 데이터
일시
감성 수치

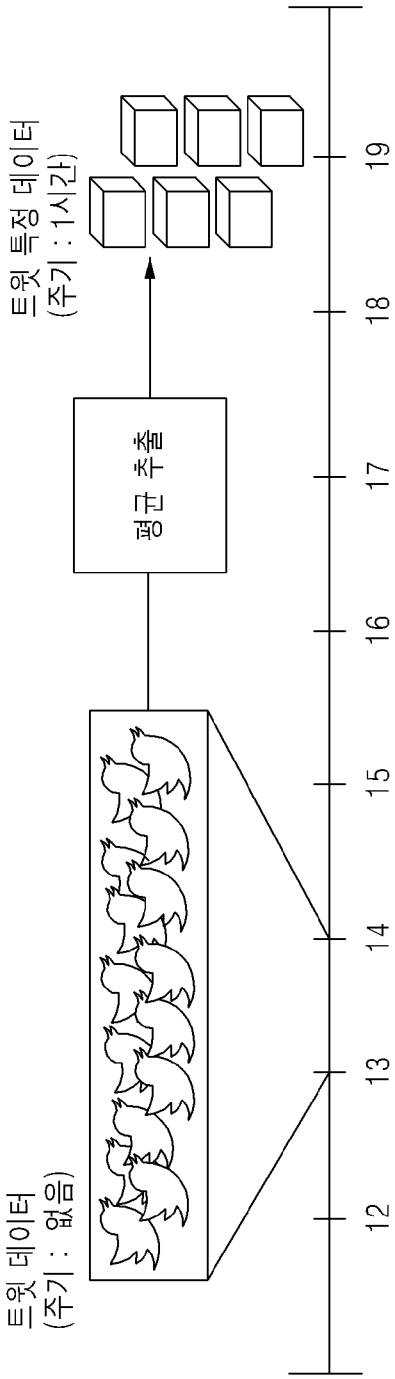
[도5a]


블록 통계 데이터
(주기 : 1시간)
평균 추출
블록 통계 데이터
(주기 : 평균 10분)
12
13
14
15
16
17
18
19

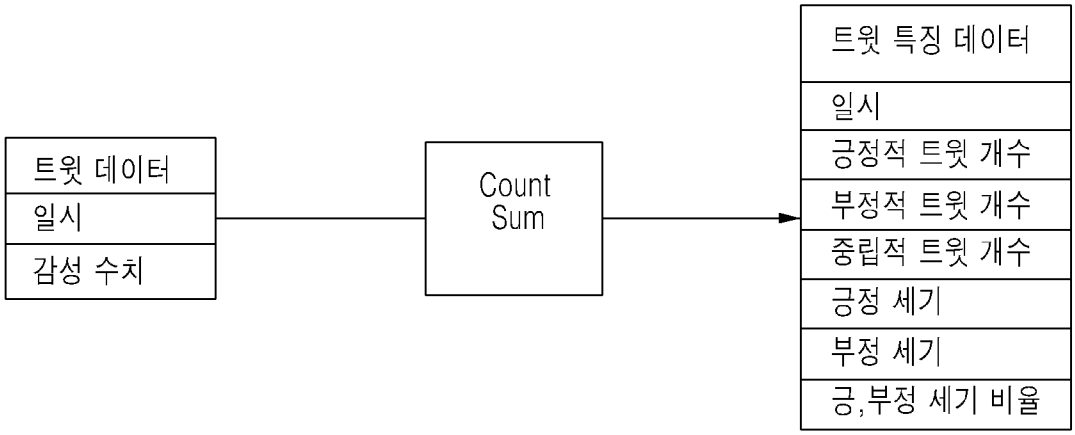
[도5b]


블록 통계 데이터
블록 높이
통계 #1
통계 #1
……
통계 #311
통계 #312
통계 추출
블록 통계 데이터
일시
통계 #1
통계 #1
……
통계 #311
통계 #312

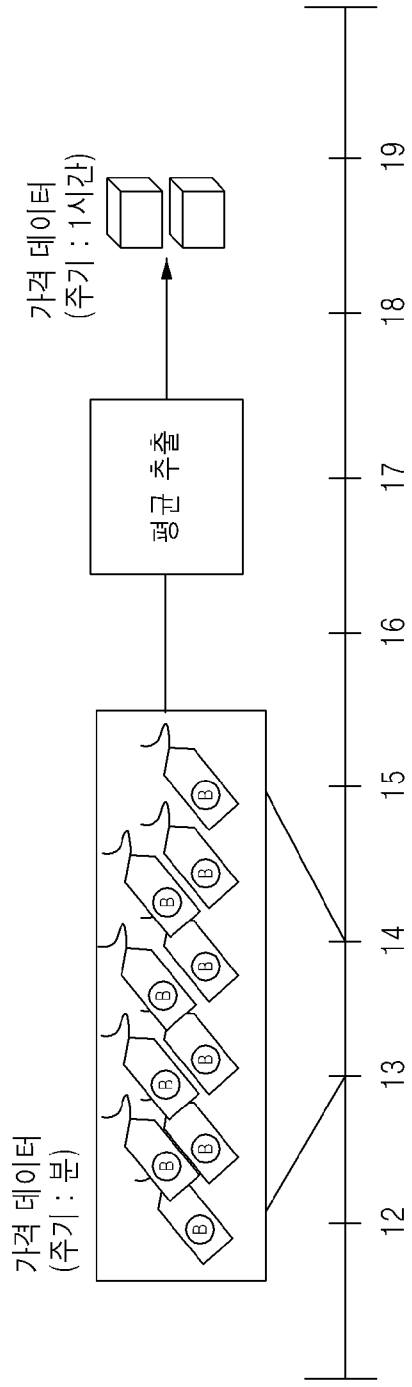
[도6a]


트윗 특정 데이터 (주기 : 1시간)
없음 대기 추출
트윗 데이터 (주기 : 없음)
12
13
14
15
16
17
18
19

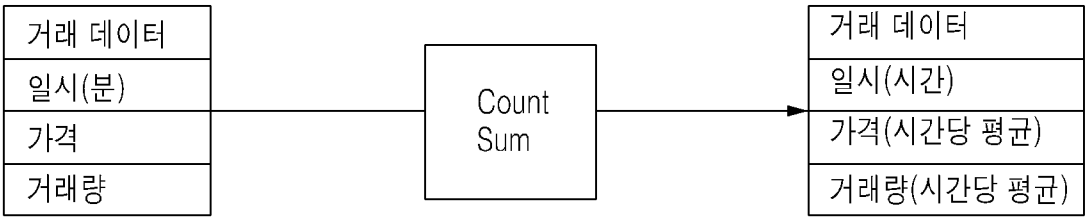
[도6b]


트윗 데이터
일시
감성 수치
Count
Sum
트윗 특징 데이터
일시
긍정적 트윗 개수
부정적 트윗 개수
중립적 트윗 개수
긍정 세기
부정 세기
긍,부정 세기 비율

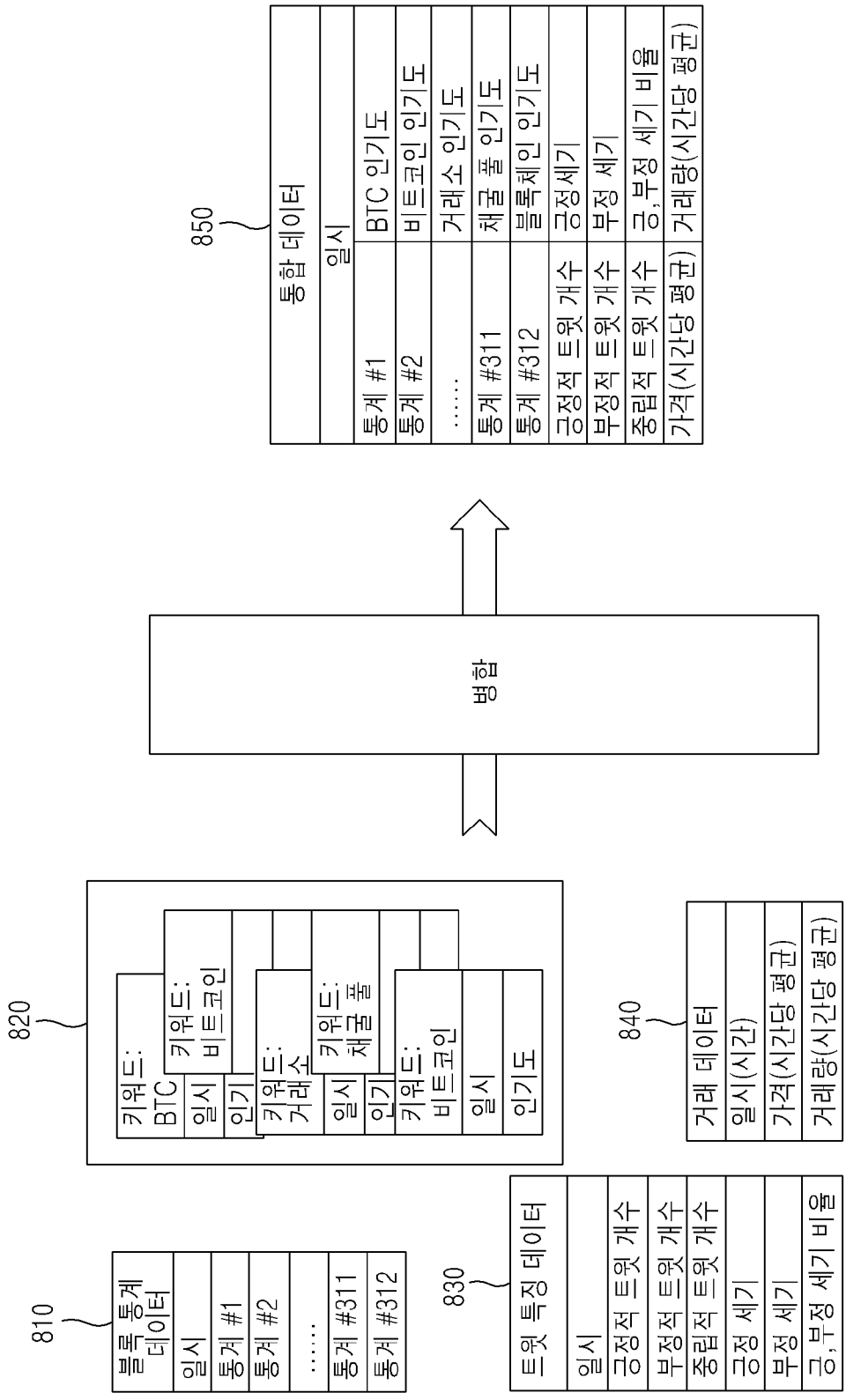
[도7a]


가격 데이터
(주기 : 분)
영역 구간 추출
가격 데이터
(주기 : 1시간)
12
13
14
15
16
17
18
19

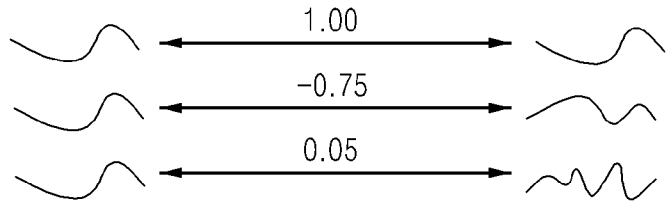
[도7b]



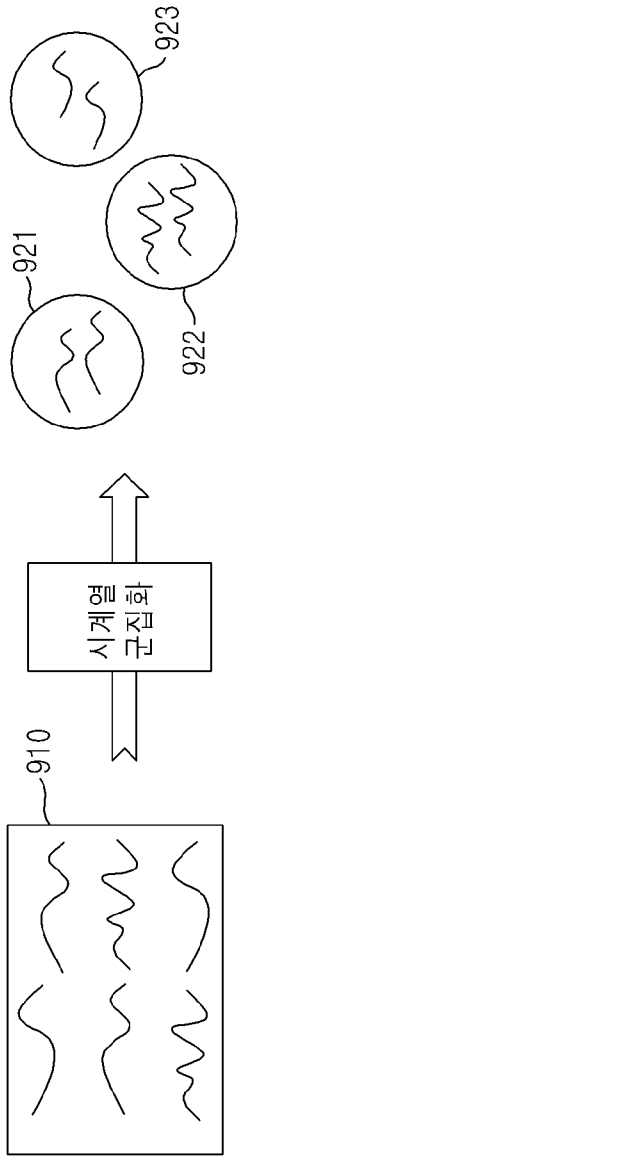
[도8]


850
통합 데이터
일시
통계 #1
BTC 인기도
통계 #2
비트코인 인기도
.....
거래소 인기도
통계 #311
채굴 풀 인기도
통계 #312
블록체인 인기도
긍정적 트윗 개수
긍정세기
부정적 트윗 개수
부정 세기
중립적 트윗 개수
긍,부정 세기 비율
가격(시간당 평균)
거래량(시간당 평균)
병합
820
키워드: BTC
일시
인기
키워드: 비트코인
키워드: 거래소
일시
인기
키워드: 채굴 풀
키워드: 비트코인
일시
인기도
840
거래 데이터
일시(시간)
가격(시간당 평균)
거래량(시간당 평균)
810
블록 통계 데이터
일시
통계 #1
통계 #2
.....
통계 #311
통계 #312
830
트윗 특징 데이터
일시
긍정적 트윗 개수
부정적 트윗 개수
중립적 트윗 개수
긍정 세기
부정 세기
긍,부정 세기 비율

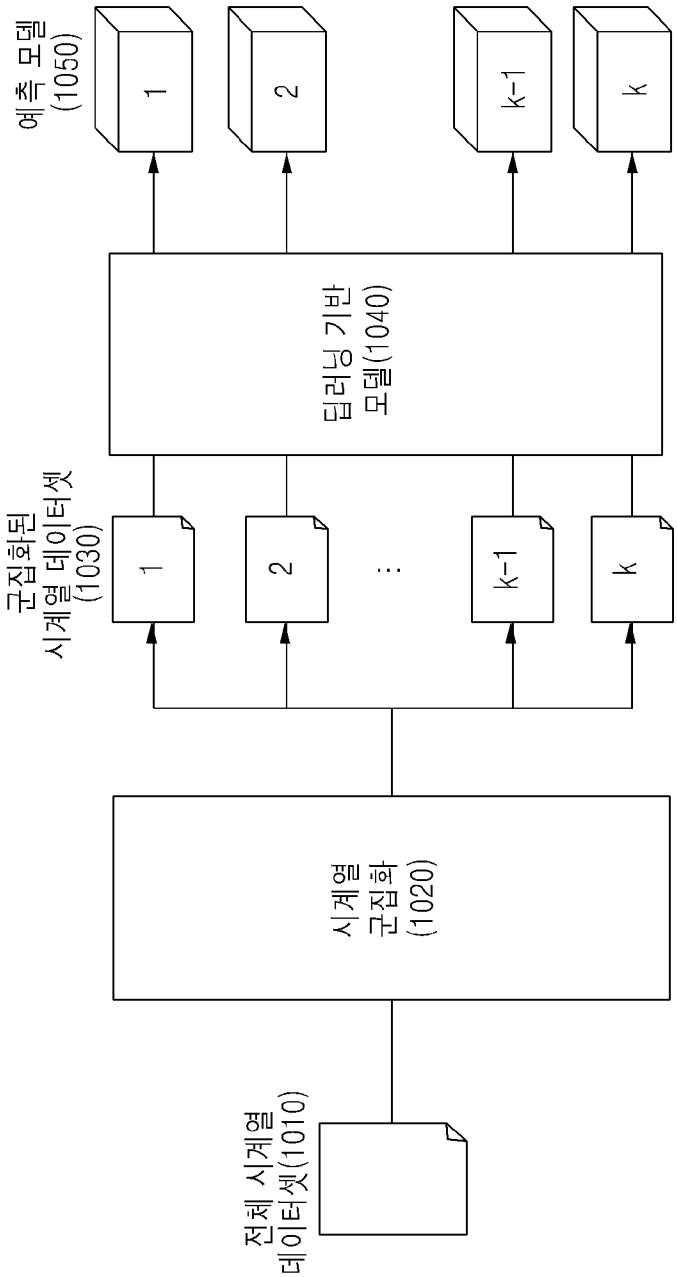
[도9a]



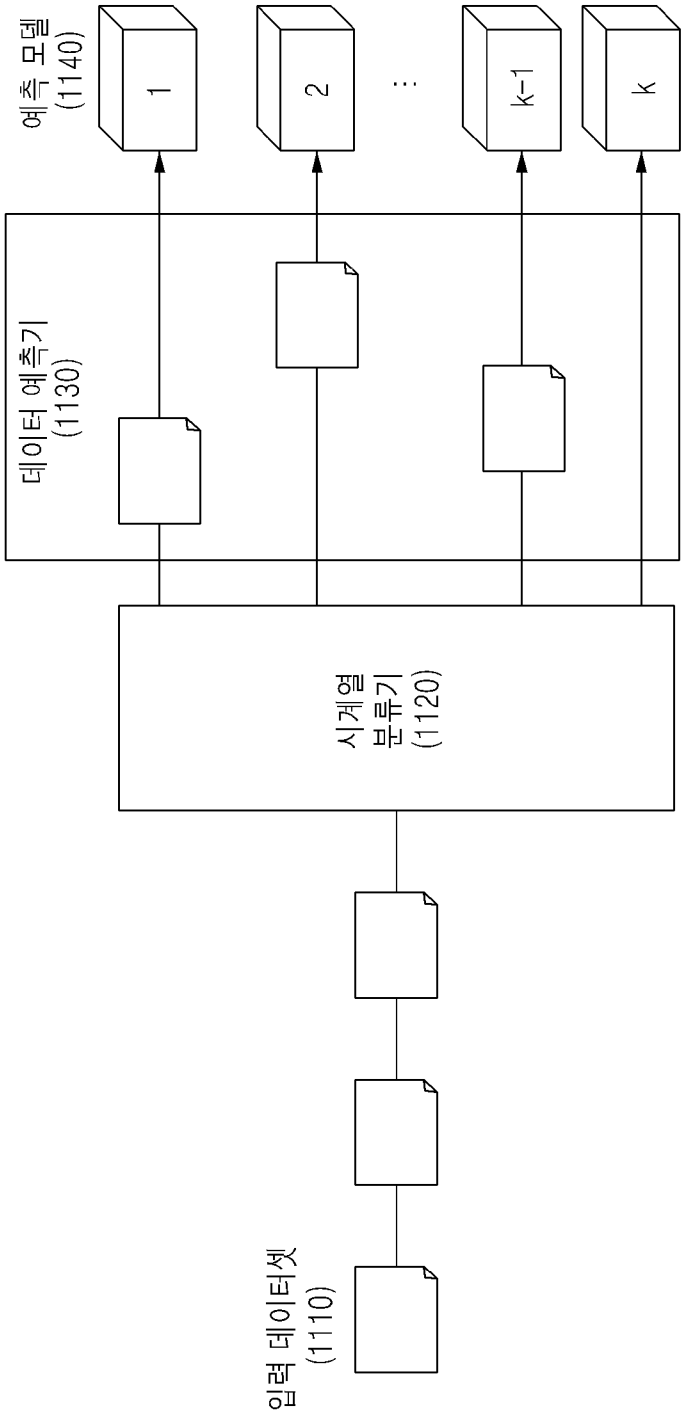
[도9b]



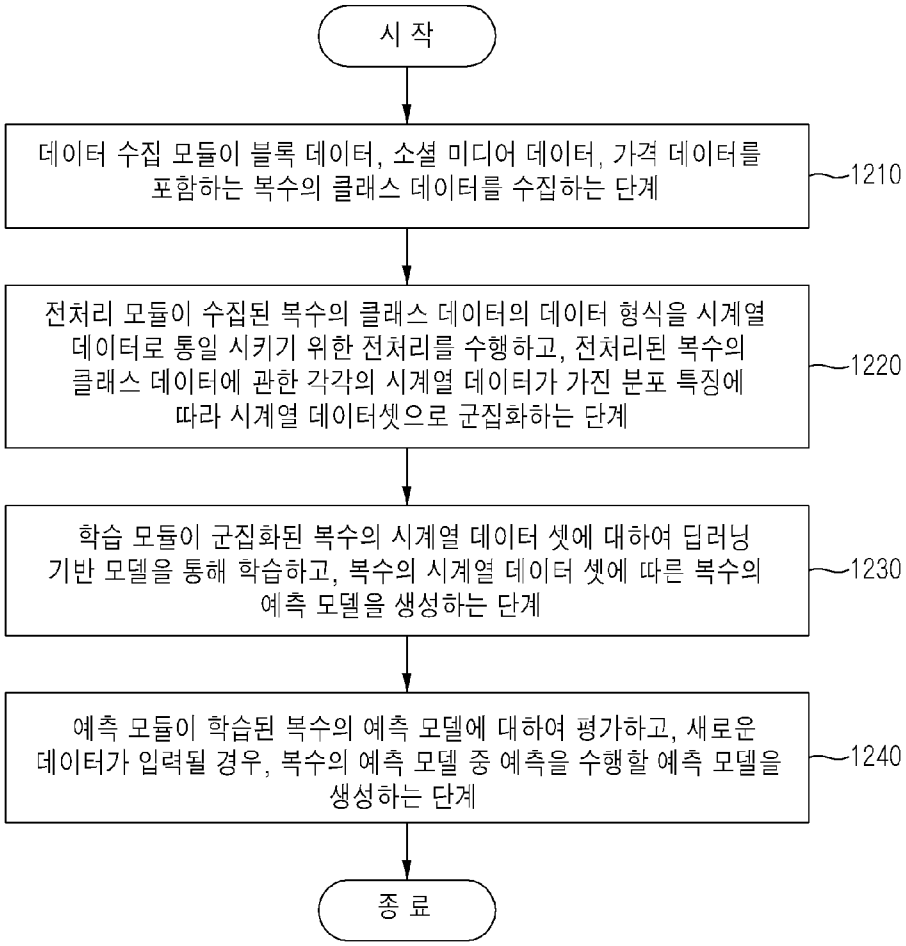
[도10]


전체 시계열 데이터셋(1010)
시계열 군집화 (1020)
군집화된 시계열 데이터셋 (1030)
1
2
...
k-1
k
딥러닝 기반 모델(1040)
예측 모델 (1050)
1
2
k-1
k

[도11]



[도12]


시 작
데이터 수집 모듈이 블록 데이터, 소셜 미디어 데이터, 가격 데이터를 포함하는 복수의 클래스 데이터를 수집하는 단계
1210
전처리 모듈이 수집된 복수의 클래스 데이터의 데이터 형식을 시계열 데이터로 통일 시키기 위한 전처리를 수행하고, 전처리된 복수의 클래스 데이터에 관한 각각의 시계열 데이터가 가진 분포 특징에 따라 시계열 데이터셋으로 군집화하는 단계
1220
학습 모듈이 군집화된 복수의 시계열 데이터 셋에 대하여 딥러닝 기반 모델을 통해 학습하고, 복수의 시계열 데이터 셋에 따른 복수의 예측 모델을 생성하는 단계
1230
예측 모듈이 학습된 복수의 예측 모델에 대하여 평가하고, 새로운 데이터가 입력될 경우, 복수의 예측 모델 중 예측을 수행할 예측 모델을 생성하는 단계
1240
종 료

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2020/017919

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 10/06(2012.01)i; **G06Q 10/04**(2012.01)i; **G06Q 20/06**(2012.01)i; **G06F 40/30**(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06Q 10/06(2012.01); G06N 20/00(2019.01); G06N 99/00(2010.01); G06Q 10/04(2012.01); G06Q 20/06(2012.01); G06Q 20/36(2012.01); G06Q 30/02(2012.01); G06Q 50/02(2012.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & keywords: 시계열(time-series), 예측 모델(prediction model), 딥러닝(deep-learning), 블록체인(blockchain), 데이터(data), 비트코인(bitcoin), 전처리(preprocess)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KR 10-2020-0115708 A (KEPCO KDN CO., LTD.) 08 October 2020 (2020-10-08) See paragraphs [0004], [0038] and [0041], claims 1, 3 and 5 and figures 1-3.	1-12
Y	KR 10-1862000 B1 (POPFUNDING) 29 May 2018 (2018-05-29) See claim 1.	1-12
Y	KR 10-2019-0013038 A (BIGTREE) 11 February 2019 (2019-02-11) See claims 1, 2, 4 and 6.	1-12
Y	KR 10-2020-0036219 A (CHUNGBUK NATIONAL UNIVERSITY INDUSTRY-ACADEMIC COOPERATION FOUNDATION) 07 April 2020 (2020-04-07) See paragraphs [0053]-[0060] and figures 7, 10 and 11.	5,11
Y	KR 10-2020-0005206 A (AIM SYSTEM, INC. et al.) 15 January 2020 (2020-01-15) See paragraph [0087] and figure 27.	6,12

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 January 2022	**20 January 2022**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2020/017919

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2020-0115708 A	08 October 2020	KR 10-2205215 B1	19 January 2021
KR 10-1862000 B1	29 May 2018	JP 2020-500339 A	09 January 2020
		JP 6703239 B2	03 June 2020
		WO 2019-103262 A1	31 May 2019
KR 10-2019-0013038 A	11 February 2019	None	
KR 10-2020-0036219 A	07 April 2020	KR 10-2137583 B1	24 July 2020
KR 10-2020-0005206 A	15 January 2020	None	

국 제 조 사 보 고 서

국제출원번호
PCT/KR2020/017919

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06Q 10/06(2012.01)i; **G06Q 10/04**(2012.01)i; **G06Q 20/06**(2012.01)i; **G06F 40/30**(2020.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06Q 10/06(2012.01); G06N 20/00(2019.01); G06N 99/00(2010.01); G06Q 10/04(2012.01); G06Q 20/06(2012.01); G06Q 20/36(2012.01); G06Q 30/02(2012.01); G06Q 50/02(2012.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 시계열(time-series), 예측 모델(prediction model), 딥러닝(deep-learning), 블록체인(blockchain), 데이터(data), 비트코인(bitcoin), 전처리(preprocess)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	KR 10-2020-0115708 A (한전케이디엔주식회사) 2020.10.08 단락 4, 38, 41, 청구항 1, 3, 5 및 도면 1-3 참조.	1-12
Y	KR 10-1862000 B1 (팝펀딩 주식회사) 2018.05.29 청구항 1 참조.	1-12
Y	KR 10-2019-0013038 A (주식회사 빅트리) 2019.02.11 청구항 1, 2, 4, 6 참조.	1-12
Y	KR 10-2020-0036219 A (충북대학교 산학협력단) 2020.04.07 단락 53-60 및 도면 7, 10, 11 참조.	5,11
Y	KR 10-2020-0005206 A (에임시스템 주식회사 등) 2020.01.15 단락 87 및 도면 27 참조.	6,12

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☑ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌
"D" 본 국제출원에서 출원인이 인용한 문헌
"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌
"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌
"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌
"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.
"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2022년01월20일(20.01.2022)	**2022년01월20일(20.01.2022)**

ISA/KR의 명칭 및 우편주소

대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)

팩스 번호 **+82-42-481-8578**

심사관

박혜련

전화번호 **+82-42-481-3463**

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2020/017919

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2020-0115708 A	2020/10/08	KR 10-2205215 B1	2021/01/19
KR 10-1862000 B1	2018/05/29	JP 2020-500339 A	2020/01/09
		JP 6703239 B2	2020/06/03
		WO 2019-103262 A1	2019/05/31
KR 10-2019-0013038 A	2019/02/11	없음	
KR 10-2020-0036219 A	2020/04/07	KR 10-2137583 B1	2020/07/24
KR 10-2020-0005206 A	2020/01/15	없음	