(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0032312
(43) 공개일자 2023년03월07일

(51) 국제특허분류(Int. Cl.)
G06F 16/33 (2019.01) *G06F 16/34* (2019.01)
G06F 16/35 (2019.01) *G06F 17/16* (2006.01)
G06F 18/00 (2023.01)
(52) CPC특허분류
G06F 16/3347 (2019.01)
G06F 16/34 (2019.01)
(21) 출원번호 10-2021-0115007
(22) 출원일자 2021년08월30일
심사청구일자 2021년08월30일

(71) 출원인
엘아이지넥스원 주식회사
경기도 용인시 기흥구 마북로 207 (마북동)
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
김근형
경기도 성남시 분당구 판교로 333 LIG넥스원
박승환
대전광역시 유성구 대학로 99 충남대학교
(74) 대리인
전경석

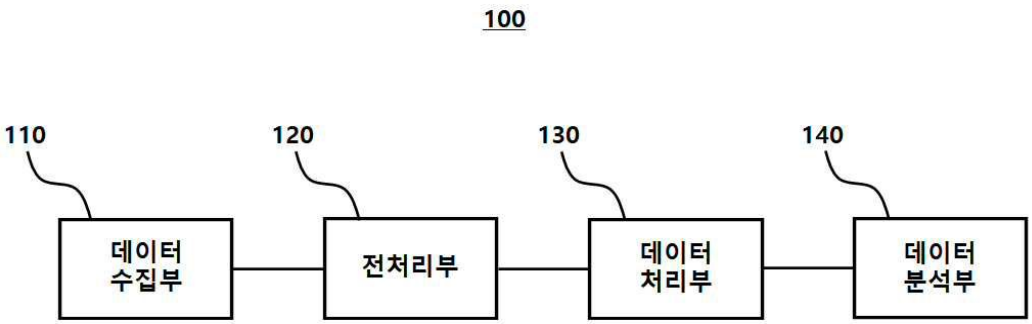
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **비정형 데이터 기반 데이터 처리 시스템 및 데이터 처리 방법**

(57) 요 약

비정형 데이터 기반 데이터 처리 시스템 및 데이터 처리 방법이 개시될 수 있다. 일 실시예에 따른 비정형 데이터 기반 데이터 처리 시스템은 비정형 데이터이며 분석하고자 하는 복수의 데이터를 수집하는 데이터 수집부; 복수의 데이터에 대해 전처리를 수행하는 전처리부; 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 데이터 처리부; 및 벡터 데이터에 군집화 및 시각화를 수행하여 상기 복수의 데이터에 대한 새로운 그룹화를 수행하는 데이터 분석부를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/35 (2019.01)
G06F 17/16 (2013.01)
G06F 18/23213 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711133632
과제번호	2019R1G1A1004084
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	스마트 제조를 위한 빅 데이터 애널리틱스 기반 첨단 공정관리 기법 개발
기 여 율	1/1
과제수행기관명	충남대학교
연구기간	2020.03.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

비정형 데이터 기반 데이터 처리 시스템으로서,

비정형 데이터이며 분석하고자 하는 복수의 데이터를 수집하는 데이터 수집부;

상기 데이터 수집부에 연결되고, 상기 복수의 데이터에 대해 전처리를 수행하는 전처리부;

상기 전처리부에 연결되고, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 데이터 처리부; 및

상기 데이터 처리부에 연결되고, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 복수의 데이터에 대한 새로운 그룹화를 수행하는 데이터 분석부

를 포함하는 데이터 처리 시스템.

청구항 2

제1항에 있어서, 상기 데이터는 고장원인 상세 데이터를 포함하는 데이터 처리 시스템.

청구항 3

제2항에 있어서, 상기 데이터 처리부는 상기 전처리된 데이터에 Doc2vec 알고리즘을 적용하여 상기 정형화된 벡터 데이터로 변환하는 데이터 처리 시스템.

청구항 4

제2항에 있어서, 상기 데이터 분석부는 상기 벡터 데이터의 차원을 축소시키는 데이터 처리 시스템.

청구항 5

제4항에 있어서, 상기 데이터 분석부는 상기 벡터 데이터를 t-SNE(stochastic neighbor embedding) 알고리즘에 적용하여 상기 벡터 데이터의 차원을 축소시키는 데이터 처리 시스템.

청구항 6

제5항에 있어서, 상기 데이터 분석부는

상기 t-SNE 알고리즘을 이용하여 상기 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하고,

상기 특징 벡터의 군집화(Clustering)를 통해 군집(Cluster) 특성을 분석하는 데이터 처리 시스템.

청구항 7

제6항에 있어서, 상기 데이터 분석부는 K-means 알고리즘을 이용하여 상기 군집화를 수행하는 데이터 처리 시스템.

청구항 8

비정형 데이터 기반 데이터 처리 시스템에서의 데이터 처리 방법으로서,

상기 데이터 처리 시스템의 데이터 수집부에서, 비정형 데이터이며 분석하고자 하는 복수의 데이터를 수집하는 단계;

상기 데이터 처리 시스템의 전처리부에서, 상기 복수의 데이터에 대해 전처리를 수행하는 단계;

상기 데이터 처리 시스템의 데이터 전처리부에서, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 단계; 및

상기 데이터 처리 시스템의 데이터 분석부에서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 복수의 데이터에 대한 새로운 그룹화를 수행하는 단계

를 포함하는 데이터 처리 방법.

청구항 9

제8항에 있어서, 상기 데이터는 고장원인 상세 데이터를 포함하는 데이터 처리 방법.

청구항 10

제9항에 있어서, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 단계는

상기 전처리된 데이터에 Doc2vec 알고리즘을 적용하여 상기 정형화된 벡터 데이터로 변환하는 단계

를 포함하는 데이터 처리 방법.

청구항 11

제9항에 있어서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 고장원인 상세 데이터에 대한 새로운 그룹화를 수행하는 단계는

상기 벡터 데이터의 차원을 축소시키는 단계

를 포함하는 데이터 처리 방법.

청구항 12

제11항에 있어서, 상기 벡터 데이터의 차원을 축소시키는 단계는

상기 벡터 데이터를 t-SNE 알고리즘에 적용하여 상기 벡터 데이터의 차원을 축소시키는 단계

를 포함하는 데이터 처리 방법.

청구항 13

제12항에 있어서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 고장원인 상세 데이터에 대한 새로운 그룹화를 수행하는 단계는

상기 t-SNE 알고리즘을 이용하여 상기 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하는 단계; 및

상기 특징 벡터의 군집화를 통해 군집 특성을 분석하는 단계

를 포함하는 데이터 처리 방법.

청구항 14

제13항에 있어서, 상기 특징 벡터의 군집화를 통해 군집 특성을 분석하는 단계는

K-means 알고리즘을 이용하여 상기 군집화를 수행하는 단계

를 포함하는 데이터 처리 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 비정형 데이터 기반 데이터 처리 시스템 및 데이터 처리 방법에 관한 것이다.

배 경 기 술

[0002] 국방 분야에서도 축적되는 데이터를 통해 의미 있는 정보를 추출하여 미래 무기체계 발전을 위한 정책제정, 연구개발, 군수품 품질보증 활동 등에 활용하는 것이 필요한 실정이다. 특히, 국방품질 정책 수립 활용을 위해 군수품의 전반적인 품질 문제의 동향 파악을 위해 국방 분야의 빅데이터 활용이 필요하나 활용 사례가 적다.

[0003] 한편, 정비 이력 데이터는 양적으로 팽창되고 있으나, 아직 데이터의 의미 있는 정보를 추출하여 개발 단계의 활용은 미흡한 현실이다. 이는 야전 운용 제원이 데이터 분석용이 아닌 이력 관리용으로 수집하여 수작업으로 입력되고, 작업자의 주관이 반영된 비정형 데이터가 대부분으로 이루어진다.

[0004] 종래에는 텍스트 마이닝을 이용한 야전 운용 제원의 비정형 데이터를 정형화 분석을 통해 고장 원인 분류를 수행하고, 분석 결과를 활용하여 무기 체계 품질 향상 방안을 제시하고 있다. 종래의 데이터 분석들은 대부분 정형화된 데이터를 수집 및 정제하여 분석하는 방식이었으며, 비정형 데이터를 동일한 방식으로 적용하는데 한계가 있는 문제점이 있다. 따라서 많은 양의 비정형 데이터를 분석하기 위해서는 분석 목적에 적합한 데이터 전처리(Pre-Processing) 과정을 통해 데이터를 정형화시키는 과정이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 비정형 데이터인 고장원인 상세 데이터를 정형화된 벡터 데이터로 변환하고 벡터 데이터의 군집화 및 시각화를 수행하는 비정형 데이터 기반 데이터 처리 시스템 및 데이터 처리 방법을 제공할 수 있다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따르면, 비정형 데이터 기반 데이터 처리 시스템은 비정형 데이터이며 분석하고자 하는 복수의 데이터를 수집하는 데이터 수집부; 상기 데이터 수집부에 연결되고, 상기 복수의 데이터에 대해 전처리를 수행하는 전처리부; 상기 전처리부에 연결되고, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 데이터 처리부; 및 상기 데이터 처리부에 연결되고, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 복수의 데이터에 대한 새로운 그룹화를 수행하는 데이터 분석부를 포함할 수 있다.

[0007] 일 실시예에 있어서, 상기 데이터는 고장원인 상세 데이터를 포함할 수 있다.

[0008] 일 실시예에 있어서, 상기 데이터 처리부는 상기 전처리된 데이터에 Doc2vec 알고리즘을 적용하여 상기 정형화된 벡터 데이터로 변환할 수 있다.

[0009] 일 실시예에 있어서, 상기 데이터 분석부는 상기 벡터 데이터의 차원을 축소시킬 수 있다.

[0010] 일 실시예에 있어서, 상기 데이터 분석부는 상기 벡터 데이터를 t-SNE(stochastic neighbor embedding) 알고리즘에 적용하여 상기 벡터 데이터의 차원을 축소시킬 수 있다.

[0011] 일 실시예에 있어서, 상기 데이터 분석부는 상기 t-SNE 알고리즘을 이용하여 상기 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하고, 상기 특징 벡터의 군집화(Clustering)를 통해 군집(Cluster) 특성을 분석할 수 있다.

[0012] 일 실시예에 있어서, 상기 데이터 분석부는 K-means 알고리즘을 이용하여 상기 군집화를 수행할 수 있다.

[0013] 본 개시의 일 실시예에 따르면, 비정형 데이터 기반 데이터 처리 시스템에서의 데이터 처리 방법이 개시될 수 있다. 일 실시예에 따른 데이터 처리 방법은 상기 데이터 처리 시스템의 데이터 수집부에서, 비정형 데이터이며 분석하고자 하는 복수의 데이터를 수집하는 단계; 상기 데이터 처리 시스템의 전처리부에서, 상기 복수의 데이터에 대해 전처리를 수행하는 단계; 상기 데이터 처리 시스템의 데이터 전처리부에서, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 단계; 및 상기 데이터 처리 시스템의 데이터 분석부에서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 복수의 데이터에 대한 새로운 그룹화를 수행하는 단계를 포함할 수 있다.

[0014] 일 실시예에 있어서, 상기 데이터는 고장원인 상세 데이터를 포함할 수 있다.

[0015] 일 실시예에 있어서, 상기 전처리된 데이터를 정형화된 벡터 데이터로 변환하는 단계는 상기 전처리된 데이터에 Doc2vec 알고리즘을 적용하여 상기 정형화된 벡터 데이터로 변환하는 단계를 포함할 수 있다.

[0016] 일 실시예에 있어서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 고장원인 상세 데이터에 대한 새로운 그룹화를 수행하는 단계는 상기 벡터 데이터의 차원을 축소시키는 단계를 포함할 수 있다.

[0017] 일 실시예에 있어서, 상기 벡터 데이터의 차원을 축소시키는 단계는 상기 벡터 데이터를 t-SNE 알고리즘에 적용하여 상기 벡터 데이터의 차원을 축소시키는 단계를 포함할 수 있다.

[0018] 일 실시예에 있어서, 상기 벡터 데이터에 군집화 및 시각화를 수행하여 상기 고장원인 상세 데이터에 대한 새로

운 그룹화를 수행하는 단계는 상기 t-SNE 알고리즘을 이용하여 상기 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하는 단계; 및 상기 특징 벡터의 군집화를 통해 군집 특성을 분석하는 단계를 포함할 수 있다.

[0019] 일 실시예에 있어서, 상기 특징 벡터의 군집화를 통해 군집 특성을 분석하는 단계는 K-means 알고리즘을 이용하여 상기 군집화를 수행하는 단계를 포함할 수 있다.

발명의 효과

[0020] 본 발명의 다양한 실시예에 따르면, 무기 체계 야전 운용 제원의 비정형 데이터를 텍스트 마이닝을 활용하여 고장 원인을 분석할 수 있다.

[0021] 또한, 고차원 벡터로부터 저차원의 특징을 추출하여 군집화 및 시각화를 통해 고장 원인에 대한 새로운 분류 유형을 제시할 수 있다.

[0022] 또한, 야전 운용 제원의 실질적 문제점을 분석하고 이를 활용하여 신무기 개발 단계에 환류하면 무기 체계의 품질 향상을 도모할 수 있다.

[0023] 더욱이, 비정형 데이터의 정형화 분석은 고장원인에 따른 수리내용의 확인 뿐만 아니라 어떠한 부품 및 무기 체계가 빈번하게 발생하는지 고장 횟수 및 고장기간을 확인할 수 있어, 사전 고장을 예방할 수 있다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 일 실시예에 따른 비정형 데이터 기반 데이터 처리 시스템을 개략적으로 나타낸 도면이다.

도 2는 본 발명의 일 실시예에 따른 자연어를 벡터화하는 Doc2vec 알고리즘의 작동 방식을 나타낸 도면이다.

도 3은 본 발명의 일 실시예에 따른 t-SNE 알고리즘을 적용한 결과를 나타낸 도면이다.

도 4는 본 발명의 일 실시예에 따라 비정형 데이터 기반 데이터 처리 방법을 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0025] 본 발명의 실시예들은 본 발명의 기술적 사상을 설명하기 위한 목적으로 예시된 것이다. 본 발명에 따른 권리범위가 이하에 제시되는 실시예들이나 이들 실시예들에 대한 구체적 설명으로 한정되는 것은 아니다.

[0026] 본 발명에 사용되는 모든 기술적 용어들 및 과학적 용어들은, 달리 정의되지 않는 한, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 일반적으로 이해되는 의미를 갖는다. 본 발명에 사용되는 모든 용어들은 본 발명을 더욱 명확히 설명하기 위한 목적으로 선택된 것이며 본 발명에 따른 권리범위를 제한하기 위해 선택된 것이 아니다.

[0027] 본 발명에서 사용되는 "포함하는", "구비하는", "갖는" 등과 같은 표현은, 해당 표현이 포함되는 어구 또는 문장에서 달리 언급되지 않는 한, 다른 실시예를 포함할 가능성을 내포하는 개방형 용어(open-ended terms)로 이해되어야 한다.

[0028] 본 발명에서 기술된 단수형의 표현은 달리 언급하지 않는 한 복수형의 의미를 포함할 수 있으며, 이는 청구범위에 기재된 단수형의 표현에도 마찬가지로 적용된다.

[0029] 본 발명에서 사용되는 "제1", "제2" 등의 표현들은 복수의 구성요소들을 상호 구분하기 위해 사용되며, 해당 구성요소들의 순서 또는 중요도를 한정하는 것은 아니다.

[0030] 본 발명에서 사용되는 용어 "부"는, 소프트웨어, 또는 FPGA(field-programmable gate array), ASIC(application specific integrated circuit)과 같은 하드웨어 구성요소를 의미한다. 그러나, "부"는 하드웨어 및 소프트웨어에 한정되는 것은 아니다. "부"는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고, 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서, "부"는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스, 함수, 속성, 프로시저, 서브루틴, 프로그램 코드의 세그먼트, 드라이버, 펌웨어, 마이크로코드, 회로, 데이터, 데이터베이스, 데이터 구조, 테이블, 어레이 및 변수를 포함한다. 구성요소와 "부" 내에서 제공되는 기능은 더 작은 수의 구성요소 및 "부"로 결합되거나 추가적인 구성요소와 "부"로 분리될 수 있다.

[0031] 본 발명에서 사용되는 "~에 기초하여"라는 표현은, 해당 표현이 포함되는 어구 또는 문장에서 기술되는, 결정 판단의 행위 또는 동작에 영향을 주는 하나 이상의 인자를 기술하는데 사용되며, 이 표현은 결정, 판단의 행위

또는 동작에 영향을 주는 추가적인 인자를 배제하지 않는다.

[0032] 본 발명에서, 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 경우, 어떤 구성요소가 다른 구성요소에 직접적으로 연결될 수 있거나 접속될 수 있는 것으로, 또는 새로운 다른 구성요소를 매개로 하여 연결될 수 있거나 접속될 수 있는 것으로 이해되어야 한다.

[0033] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예들을 설명한다. 첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응하는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

[0034] 도 1은 본 발명의 일 실시예에 따른 비정형 데이터 기반 데이터 처리 시스템을 개략적으로 나타낸 도면이다. 도 1을 참조하면, 비정형 데이터 기반 데이터 처리 시스템(100)은 데이터 수집부(110), 전처리부(120), 데이터 처리부(130) 및 데이터 분석부(140)를 포함할 수 있다.

[0035] 데이터 수집부(110)는 분석하고자 하는 복수의 데이터를 수집할 수 있다. 일 실시예에 있어서, 복수의 데이터는 고장원인 상세 데이터를 포함할 수 있다. 예를 들면, 고장원인 상세 데이터는 비정형 데이터이며, 정비 부품에 관한 정보(이하, "정비 부품 정보"라 함), 정비 부품에 대한 정비 접수 번호 정보, 정비 부품에 대한 고장원인 정보, 및 고장 원인 정보에 대한 수리 내용 정보를 포함할 수 있다. 그러나, 고장원인 상세 데이터는 반드시 이에 한정되는 것은 아니다.

[0036] 전처리부(120)는 데이터 수집부(110)에 연결될 수 있다. 전처리부(120)는 데이터 수집부(110)에 의해 수집된 복수의 데이터에 대해 전처리(pre-processing)를 수행할 수 있다. 일 실시예에 있어서, 전처리부(120)는 고장원인 상세 데이터에 대해 전처리를 수행할 수 있다.

[0037] 데이터 처리부(130)는 전처리부(120)에 연결될 수 있다. 데이터 처리부(130)는 전처리부(120)에 의해 전처리된 데이터를 정형화된 벡터 데이터로 변환할 수 있다.

[0038] 일 실시예에 있어서, 데이터 처리부(130)는 비정형 데이터의 고장원인 상세 데이터를 정형화된 벡터 데이터로 변환하기 위한 알고리즘을 적용할 수 있다. 예를 들면, 데이터 처리부(130)는 Doc2vec 알고리즘을 적용하여 비정형 텍스트 데이터를 정형화된 벡터 데이터로 변환할 수 있다. 여기서, Doc2vec 알고리즘은 규모가 큰 텍스트 블록에 대한 연속적인 표현을 비지도 학습하여 벡터를 추출할 수 있다. 특히, Doc2vec 알고리즘은 복잡한 문서 데이터의 의미를 보존하면서 차원을 축소하는데 효과적이다.

[0039] 데이터 분석부(140)는 데이터 처리부(130)에 연결될 수 있다. 데이터 분석부(140)는 데이터 처리부(130)에 의해 변환된 벡터 데이터에 기초하여 고장 원인 정보 간 유사도를 산출하여 시각화를 수행할 수 있다. 일 실시예에 있어서, 데이터 분석부(140)는 변환된 벡터 데이터에 대해 군집화 및 시각화를 수행하여 고장원인 상세 데이터에 대한 새로운 그룹화를 수행할 수 있다.

[0040] 일 실시예에 있어서, 데이터 분석부(140)는 데이터 처리부(130)에 의해 고차원으로 벡터화된 데이터를 기계 학습, 시각화 및 분석을 위해 차원을 축소하는 알고리즘에 적용할 수 있다. 예를 들면, 데이터 분석부(140)는 고차원으로 벡터화된 데이터를 t-SNE(stochastic neighbor embedding) 알고리즘에 적용하여 고차원의 벡터 데이터를 저차원의 벡터 데이터로 축소시킬 수 있다.

[0041] t-SNE(stochastic neighbor embedding) 알고리즘은 고차원의 데이터를 저차원으로 축소하는 알고리즘이다. 차원 축소의 한 방법인 주성분 분석(PCA)은 선형 결합 방식으로 축소 과정에서 데이터가 겹치는 현상이 발생할 수 있지만, t-SNE 알고리즘은 고차원에서의 거리를 저차원에서도 보존하면서 차원을 축소하기 때문에 데이터가 겹치는 현상이 적다. 이런 장점으로 문서를 벡터화한 고차원의 데이터를 저차원으로 축소하여 시각화에 활용하고 있다.

[0042] t-SNE는 자유도가 1인 t분포를 사용하여 고/저차원의 거리를 조건부 확률을 사용한 유사도를 사용한다. 고차원 공간상에서 유사한 대상들은 이웃으로 선택될 확률이 높고 그렇지 않은 대상들은 이웃으로 선택될 확률이 낮아지도록 하는 유사도 확률 분포를 생성한다. 그 수식은 아래의 수학식 1과 같다.

수학식 1

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)}$$

[0043]

[0044] 수학식 1에 있어서, $P_{j|i}$는 주어진 대상 x_i가 있을 때 x_i의 이웃으로 x_j가 선택될 조건부 확률을 의미한다. 이때 x_i의 이웃으로 선택될 확률의 분포는 평균 x_i, 표준편차 σ_i인 정규분포(Gaussian distribution)를 따른다고 가정하게 된다.

[0045] 수학식 1에서의 조건부 확률 $P_{j|i}$의 경우, x_i와 x_j중에서 어떤 x를 기준으로 확률을 계산하는지에 따라 값이 달라지는 특성이 있다. 즉, 대상 x_i에 대해 x_j와의 유사도 확률($P_{j|i}$)을 계산할 때와 대상 x_j에 대해 x_i와의 유사도 확률($P_{i|j}$)을 계산할 때, 그 결과가 달라질 수 있다(수학식 1에서 분모가 달라짐). 이러한 특성은 고차원 공간상에서 대상 x_i가 이상치(outlier)인 경우, $P_{j|i}$가 모든 다른 대상들 x_j에 대해 작은 값을 가지므로 비용함수에 주는 영향이 작아, 고차원 공간상의 대상 x_j에 대응하는 저차원 공간상의 대상 y_j가 어느 위치에 존재해야 하는지를 판단하기 어렵게 만든다. 이 때문에 아래의 수학식 2에 따라 조건부 확률을 대칭화한 결합 확률(joint probability) P_{ij}를 대신 사용하게 된다. P_{ij}를 사용하면 모든 대상 x_i에 대해 $\sum_j P_i$가 항상 1/2n보다 큰 값을 갖게 되기 때문에 모든 대상 x_i 각각이 비용함수에 유의미한 영향을 줄 수 있게 된다.

수학식 2

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

[0046]

[0047] 또한, 고차원에서와 같이 저차원에서도 저차원 공간상의 유사도 확률 분포 $q_{j|i}$대신 q_{ji}를 사용한다. 이에 따라 비용함수는 아래의 수학식 3과 같이 나타낼 수 있다.

수학식 3

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

[0048]

[0049] 저차원 공간상의 대상들 간의 유사도 확률 분포를 계산할 때 정규분포 대신 Student t-분포를 사용한다는 점이다. 정규분포의 경우 양극단의 확률 값이 매우 작기 때문에 서로 간의 거리가 가까운 대상과 적당히(moderately) 떨어진 대상에 대해 유사도 확률 값이 크게 차이가 나지 않는 특성이 있다. 이는 서로 간에 적당히 떨어진 대상들을 저차원 공간상으로 투영시킬 때 서로 간의 거리가 가까운 대상에 비해 충분히 넓은 저차원 공간을 활용할 수 없음을 의미한다. 이를 crowding problem이라 부른다. 이를 해결하기 위해 t-SNE에서는 정규분포 대신 Student t-분포를 사용하여 적당히 떨어진 대상들이 저차원 공간상에서 충분히 멀리 떨어져 존재할 수 있도록 하게 된다. Student t-분포를 사용한 결합 확률 q_{ji}는 아래의 수학식 4와 같이 정의될 수 있다.

수학식 4

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

[0050]

[0051] 이하 첨부된 도면을 참조하여 벡터 데이터 변환 및 데이터 시각화를 수행하는 방법에 대해 설명하기로 한다. 도 2는 본 발명의 일 실시예에 따른 자연어를 벡터화하는 Doc2vec 알고리즘의 작동 방식을 나타낸 도면이다.

[0052] 일반적으로, 컴퓨터에서 자연어를 처리하기 위해서는 정량화하여야 한다. 다양한 정량화 방법 중 벡터로 표현하는 벡터화가 있고, 그 중에서 단어의 순서를 고려하여 문서를 벡터화하는 Doc2vec 알고리즘의 활용이 많아지고 있다. Doc2vec은 비슷한 문서끼리 높은 유사도가 나오도록 특정 차원으로 벡터화할 수 있다. 이와 같이, 문서 간의 유사도를 활용하여 군집화, 머신러닝 등 다양한 분석에 적용할 수 있다.

[0053] 도 2를 참조하면, 데이터 처리부(130)는 문서의 식별자인 접수번호를 One-Hot 인코딩하여 입력노드로 사용하고, 설정한 윈도우의 크기만큼의 단어를 순차적으로 출력노드로 설정하여 원하는 n차원의 은닉층 1개를 가진 인공신경망에서 학습을 수행시킬 수 있다. 문서의 식별자는 학습된 은닉층의 가중치를 통해 문장을 n차원으로 임베딩하면 n차원의 특징벡터가 생성되고, 이를 문서 유사도 비교에 활용될 수 있다.

[0054] 도 3은 본 발명의 일 실시예에 따른 t-SNE 알고리즘을 적용한 결과를 나타낸 도면이다. 야전 운용 제원은 고장 원인 상세 내용에 대해 작업자가 기존에 정해진 4가지 고장 타입 중 한 개를 선택함으로써 기록된다. 예를 들면, 고장 타입은 설계, 자재, 작업 및 기타를 포함할 수 있다. 데이터 분석부(140)는 군집화 및 시각화를 통해 고장 원인 상세내용에 대한 새로운 그룹화를 수행할 수 있다.

[0055] 일 실시예에 있어서, 데이터 분석부(140)는 데이터 처리부(130)에 의해 Doc2vec 알고리즘에 적용된 고장원인 상세 데이터(즉, 벡터 데이터)를 시각화하여 분석을 수행할 수 있다. 여기서, 데이터 처리부(130)의 Doc2vec 알고리즘에 의해 생성된 벡터 데이터는 500차원의 고차원 데이터이기 때문에 시각화가 어렵다. 따라서, 데이터 분석부(140)는 t-SNE 알고리즘을 이용하여 500차원의 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하고, 생성된 특징 벡터의 군집화(Clustering)를 통해 군집(Cluster) 특성을 분석할 수 있다.

[0056] 일 실시예에 있어서, 군집화 알고리즘은 계산이 빠르고 높은 성능을 보여주는 K-means 알고리즘을 사용할 수 있다. 도 3에서는 실루엣 계수를 통해 군집의 개수를 6으로 설정하고 있다. 도 3은 야전 운용 제원의 t-SNE 특징 벡터에 기초하여 6개의 군집(A~F)을 생성한 예를 나타낸다.

[0057] 본 발명에 도시된 흐름도에서 프로세스 단계들, 방법 단계들, 알고리즘들 등이 순차적인 순서로 설명되었지만, 그러한 프로세스들, 방법들 및 알고리즘들은 임의의 적합한 순서로 작동되도록 구성될 수 있다. 다시 말하면, 본 발명의 다양한 실시예들에서 설명되는 프로세스들, 방법들 및 알고리즘들의 단계들이 본 발명에서 기술된 순서로 수행될 필요는 없다. 또한, 일부 단계들이 비동시적으로 수행되는 것으로서 설명되더라도, 다른 실시예에서는 이러한 일부 단계들이 동시에 수행될 수 있다. 또한, 도면에서의 묘사에 의한 프로세스의 예시는 예시된 프로세스가 그에 대한 다른 변화들 및 수정들을 제외하는 것을 의미하지 않으며, 예시된 프로세스 또는 그의 단계들 중 임의의 것이 본 발명의 다양한 실시예들 중 하나 이상에 필수적임을 의미하지 않으며, 예시된 프로세스가 바람직하다는 것을 의미하지 않는다.

[0058] 도 4는 본 발명의 일 실시예에 따라 비정형 데이터 기반 데이터 처리 방법을 나타낸 흐름도이다. 도 4를 참조하면, 단계 S402에서, 데이터 수집부(110)는 분석하고자 하는 복수의 데이터를 수집할 수 있다. 일 실시예에 있어서, 데이터 수집부(110)는 비정형 데이터인 고장원인 상세 데이터를 수집할 수 있다.

[0059] 단계 S404에서, 전처리부(120)는 수집된 고장원인 상세 데이터에 대해 전처리를 수행할 수 있다. 일 실시예에 있어서, 전처리부(120)는 데이터 수집부(110)로부터 고장원인 상세 데이터를 수신하고, 수신된 고장원인 상세 데이터에 전처리를 수행할 수 있다.

[0060] 단계 S406에서, 데이터 처리부(130)는 전처리부(120)에 의해 전처리된 데이터를 정형화된 벡터 데이터로 변환할 수 있다. 일 실시예에 있어서, 데이터 처리부(130)는

[0061] 데이터 처리부(130)는 전처리부(120)에 연결될 수 있다. 데이터 처리부(130)는 전처리부(120)에 의해 전처리된 데이터를 정형화된 벡터 데이터로 변환할 수 있다. 일 실시예에 있어서, 데이터 처리부(130)는 Doc2vec 알고리즘을 적용하여 비정형 텍스트 데이터를 정형화된 벡터 데이터로 변환할 수 있다.

[0062] 단계 S408에서, 데이터 분석부(140)는 데이터 처리부(130)에 의해 변환된 벡터 데이터에 기초하여 고장 원인 정보 간 유사도를 산출하여 시각화를 수행할 수 있다. 일 실시예에 있어서, 데이터 분석부(140)는 변환된 벡터 데이터의 군집화 및 시각화를 통해 고장 원인 상세내용에 대한 새로운 그룹화를 수행할 수 있다. 예를 들면, 데이터 분석부(140)는 t-SNE 알고리즘을 이용하여 벡터 데이터를 대표할 수 있는 특징 벡터를 생성하고, 생성된 특징 벡터의 군집화를 통해 군집 특성을 분석할 수 있다.

[0063] 위 방법은 특정 실시예들을 통하여 설명되었지만, 위 방법은 또한 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다. 그리고, 위 실시예들을 구현하기 위한 기능적인(functional) 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있다.

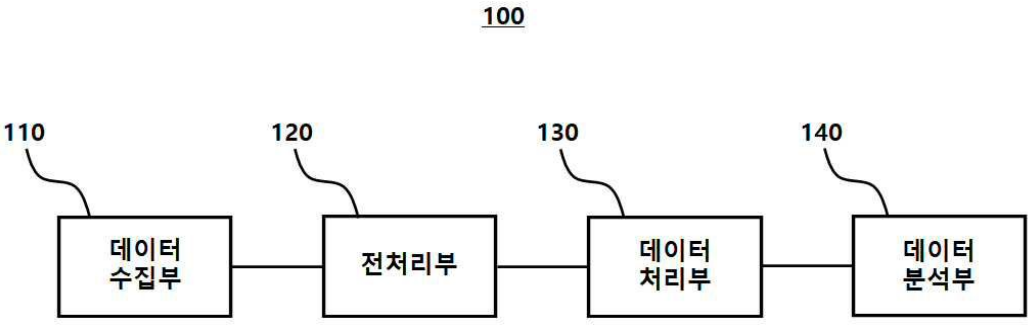
[0064] 이상 일부 실시예들과 첨부된 도면에 도시된 예에 의해 본 발명의 기술적 사상이 설명되었지만, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 이해할 수 있는 본 발명의 기술적 사상 및 범위를 벗어나지 않는 범위에서 다양한 치환, 변형 및 변경이 이루어질 수 있다는 점을 알아야 할 것이다. 또한, 그러한 치환, 변형 및 변경은 첨부된 청구범위 내에 속하는 것으로 생각되어야 한다.

부호의 설명

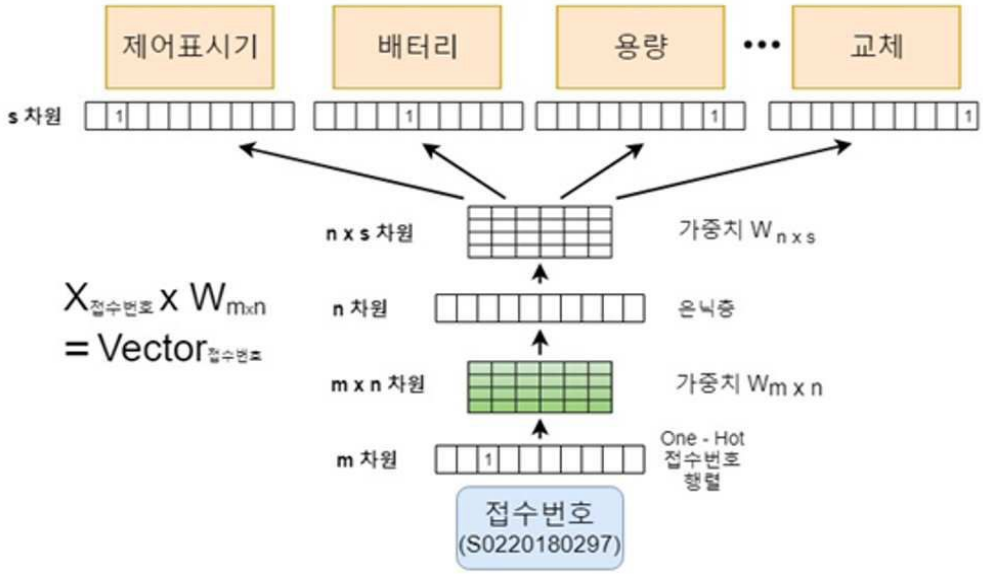
[0065] 100: 비정형 데이터 기반 데이터 처리 시스템, 110: 데이터 수집부, 120: 전처리부, 130: 데이터 처리부, 140: 데이터 분석부

도면

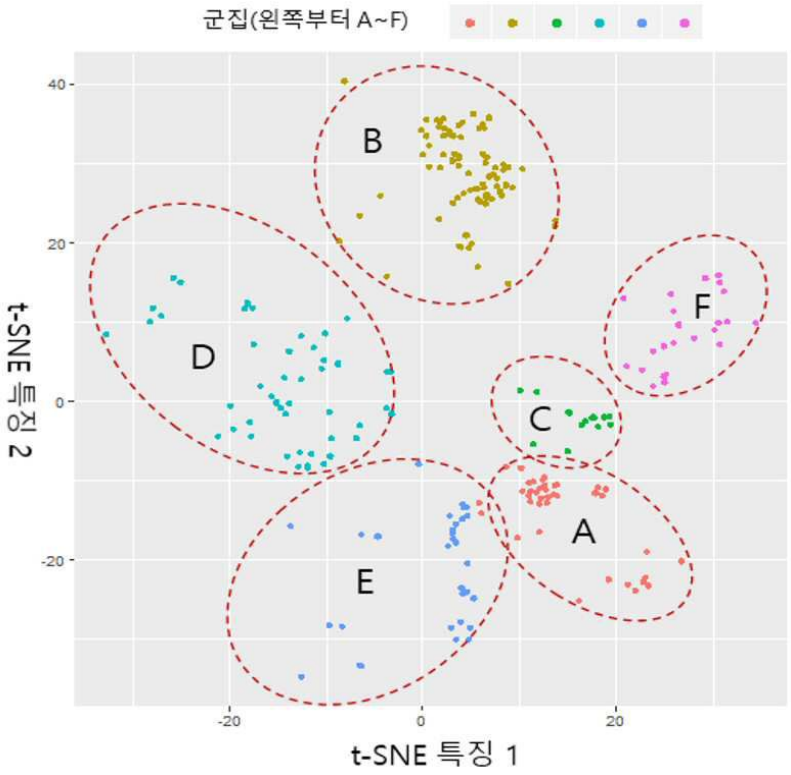
도면1



도면2


제어표시기
배터리
용량
교체
s 차원
n x s 차원
가중치 W n x s
X 접수번호 x W mxn
= Vector 접수번호
n 차원
은닉층
m x n 차원
가중치 W m x n
m 차원
One - Hot
접수번호
행렬
접수번호
(S0220180297)

도면3


군집(왼쪽부터 A~F)
B
D
F
C
A
E
t-SNE 특징 2
t-SNE 특징 1
40
20
0
-20
-20
0
20

도면4

