

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2022-0009235
(43) 공개일자 2022년01월24일

(51) 국제특허분류(Int. Cl.)

G16H 50/80 (2018.01) G06N 20/00 (2019.01)

G06Q 10/04 (2012.01) G06Q 10/06 (2012.01)

G06Q 50/00 (2018.01) G16H 50/50 (2018.01)

(52) CPC특허분류

G16H 50/80 (2018.01)

G06N 20/00 (2021.08)

(21) 출원번호 10-2020-0087695

(22) 출원일자 2020년07월15일

심사청구일자 2020년07월15일

(71) 출원인

세종대학교산학협력단

서울특별시 광진구 능동로 209 (군자동, 세종대학교)

(72) 발명자

노승민

서울특별시 은평구 진관2로 111-7, 214동 805호
(진관동, 은평뉴타운우물골)

비밀 산무가나단

서울특별시 광진구 능동로 209, 세종대학교
대양AI 센터 621호

와이 해롤드 로빈슨

서울특별시 광진구 능동로 209, 세종대학교
대양AI 센터 621호

(74) 대리인

특허법인위더피플

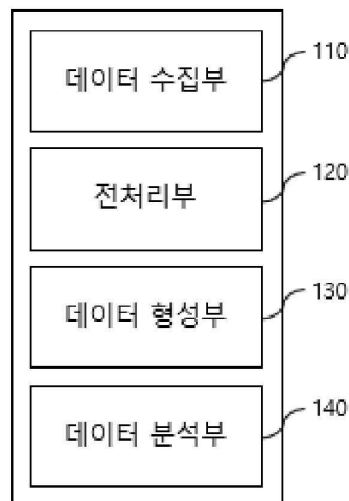
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치 및 방법

(57) 요약

본 발명은 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치 및 방법을 개시한다. 본 발명의 일 측면에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치는, 소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집하는 데이터 수집부; 상기 수집된 데이터를 전처리하는 전처리부; 상기 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성하는 데이터 형성부; 및 상기 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석하는 데이터 분석부;를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06Q 10/04 (2013.01)

G06Q 10/06375 (2013.01)

G06Q 50/01 (2013.01)

G16H 50/50 (2018.01)

G16H 50/70 (2018.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711106301

과제번호 2019R1F1A1060668

부처명 과학기술정보통신부

과제관리(전문)기관명 한국연구재단

연구사업명 개인기초연구(과기정통부)(R&D)

연구과제명 점진적 행동지식 학습 및 eXplainable 모델링을 통한 영유아/아동의 이상 질환 조기

예측 시스템 연구

기 여 율 1/1

과제수행기관명 세종대학교 산학협력단

연구기간 2020.03.01 ~ 2021.02.28

명세서

청구범위

청구항 1

소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집하는 데이터 수집부;

상기 수집된 데이터를 전처리하는 전처리부;

상기 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성하는 데이터 형성부; 및

상기 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석하는 데이터 분석부;를 포함하는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치.

청구항 2

제 1 항에 있어서,

상기 전처리부는,

데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하여 수집된 데이터를 전처리하는 것을 특징으로 하는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치.

청구항 3

제 1 항에 있어서,

상기 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함되는 것을 특징으로 하는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치.

청구항 4

제 1 항에 있어서,

상기 데이터 분석부는,

벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석하는 것을 특징으로 하는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치.

청구항 5

소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치에서의 인플루엔자 패턴 예측 방법에 있어서,

소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집하는 데이터 수집 단계;

상기 수집된 데이터를 전처리하는 전처리 단계;

상기 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성하는 데이터 형성 단계; 및

상기 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석하는 데이터 분석 단계;를 포함하는 인플루엔자 패턴 예측 방법.

청구항 6

제 5 항에 있어서,

상기 전처리 단계는,

데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하여 수집된 데이터를 전처리하는 것을 특징으로 하는 인

플루엔자 패턴 예측 방법.

청구항 7

제 5 항에 있어서,

상기 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함되는 것을 특징으로 하는 인플루엔자 패턴 예측 방법.

청구항 8

제 5 항에 있어서,

상기 데이터 분석 단계는,

벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석하는 것을 특징으로 하는 인플루엔자 패턴 예측 방법.

발명의 설명

기술 분야

[0001] 본 발명은 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치 및 방법에 관한 것으로, 더욱 상세하게는 트위터 등과 같은 소셜 네트워크 상의 데이터를 기반으로 인플루엔자 패턴을 예측하는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치 및 방법에 관한 것이다.

배경 기술

[0003] 최근에는 감염에 대한 적절한 의식없이 수많은 감염성 바이러스가 만들어져 빠르게 발달하고 퍼져나가고 있다.

[0004] 전염병의 출현은 중대한 보건 문제로서, 기후 및 환경 문제 등과 더불어 세계적인 이슈이다. 최근 발생한 신종 플루, 사스(SARS), 구제역 및 조류 인플루엔자 등의 사례에서 보는 것처럼, 운송 수단의 발달과 글로벌화로 인해 질병의 전염은 국경을 무시하고 빠른 시간 내에 전 세계적인 재앙으로 확산될 가능성이 매우 높다. 이전부터 이러한 전염병을 대비하기 위한 생물학적 백신 개발이 활발하게 진행되고 있으나, 사실상 전염병에 대한 대응책으로는 조기 예방이 무엇보다 중요하다.

[0005] 현대를 살아가는 대부분의 사람들은 트위터와 같은 소셜 네트워크를 통해 다양한 정보를 교환하며 소통한다. 따라서, 이러한 온라인 상의 정보들을 활용하여 개인의 정보를 분석한다면 전파력이 빠른 인플루엔자(독감)의 유행을 예측하여 전염병이 확산되는 것을 어느 정도 예방할 수 있을 것이다.

[0006] 따라서, 상술한 바와 같이 소셜 네트워크 정보를 활용하여 전염병을 예측하는 기술에 대한 연구가 필요한 실정이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국등록특허 제10-1960504호(2019.07.15 공고)

발명의 내용

해결하려는 과제

[0009] 본 발명은 상기와 같은 문제점을 해결하기 위해 제안된 것으로서, 온라인 상의 소셜 네트워크 정보를 수집하고

개인의 정보를 분석하여 전파력이 빠른 인플루엔자(독감)의 유행을 예측할 수 있는 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치 및 방법을 제공하는데 그 목적이 있다.

[0010] 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있으며, 본 발명의 일 실시예에 의해 보다 분명하게 알게 될 것이다. 또한, 본 발명의 목적 및 장점들은 특허청구범위에 나타난 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

[0012] 상기와 같은 목적을 달성하기 위한 본 발명의 일 측면에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치는, 소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집하는 데이터 수집부; 상기 수집된 데이터를 전처리하는 전처리부; 상기 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성하는 데이터 형성부; 및 상기 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석하는 데이터 분석부;를 포함한다.

[0013] 상기 전처리부는, 데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하여 수집된 데이터를 전처리하는 것을 특징으로 한다.

[0014] 상기 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함되는 것을 특징으로 한다.

[0015] 상기 데이터 분석부는, 벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석하는 것을 특징으로 한다.

[0016] 상기와 같은 목적을 달성하기 위한 본 발명의 다른 측면에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치에서의 인플루엔자 패턴 예측 방법은, 소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집하는 데이터 수집 단계; 상기 수집된 데이터를 전처리하는 전처리 단계; 상기 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성하는 데이터 형성 단계; 및 상기 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석하는 데이터 분석 단계;를 포함한다.

[0017] 상기 전처리 단계는, 데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하여 수집된 데이터를 전처리하는 것을 특징으로 한다.

[0018] 상기 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함되는 것을 특징으로 한다.

[0019] 상기 데이터 분석 단계는, 벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석하는 것을 특징으로 한다.

발명의 효과

[0021] 본 발명의 일 측면에 따르면, 온라인 상의 소셜 네트워크 정보를 수집하고 개인의 정보를 분석하여 전파력이 빠른 인플루엔자(독감)의 유행을 예측하고 이에 신속히 대응함으로써 전염병이 확산되는 것을 방지할 수 있는 효과가 있다.

[0022] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0024] 본 명세서에 첨부되는 다음의 도면들은 본 발명의 바람직한 실시예를 예시하는 것이며, 발명을 실시하기 위한 구체적인 내용들과 함께 본 발명의 기술사상을 더욱 이해시키는 역할을 하는 것이므로, 본 발명은 그러한 도면에 기재된 사항에만 한정되어 해석되어서는 아니 된다.

도 1은 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치의 개략적인 기

능 블록도,

도 2는 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치에서의 데이터 처리 흐름을 도시한 도면,

도 3은 본 발명의 일 실시예에 따른 무감독 기계 학습 알고리즘의 개략적인 구조도,

도 4는 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 방법의 개략적인 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 상술한 목적, 특징 및 장점은 첨부된 도면과 관련한 다음의 상세한 설명을 통하여 보다 분명해질 것이며, 그에 따라 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 또한, 본 발명을 설명함에 있어서 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 일 실시예를 상세히 설명하기로 한다.
- [0026] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 “포함” 한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 “...부” 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [0027] 오늘날 사람들은 소셜 미디어나 다른 형태의 소셜 플랫폼에 중독되어 있으며 인터넷이나 안드로이드 모바일 장치가 없는 사람은 없다. 따라서, 소셜 미디어는 대규모 데이터 세트 저장소를 형성하거나 고려했다. 따라서, 웹 기반 생활에 대한 정보를 모으기만 하면 트위터는 사용자가 트윗(#해시태그)을 기반으로 정보를 공유할 수 있는 가장 큰 데이터 저장소 중 하나이다.
- [0028] 조직화된 정보와 비정형화된 정보가 기하급수적으로 발전하는 한 가지 이유는, 서로 다른 온라인 생활과 블로그 목적지의 보편성을 확장하기 때문이다. 이 엄청난 양의 정보를 사용하여 지식의 일부를 형성하고 모든 공개 영역에서 중요한 활동을 수행할 수 있다. 트위터는 온라인 생활 단계에서 전문가들 사이에서 현명한 명성을 얻었으며 현재의 많은 부분에 중요한 지식을 제공하고 있다.
- [0029] 따라서, 후술하는 본 발명에 따르면 랜덤 포레스트 알고리즘(Random Forest Algorithm)에 기반한 분류를 조정된 AI 전략으로 적용하여 데이터 세트의 텍스트 데이터를 분석하고 분류에 따라 정확도 수준을 찾을 수 있다. 그후, 다양한 도표(상자 도표, 산점도, 어휘 분산 도표, 항 빈도, 도표 작성)(box plot, scatter plot, lexical dispersion plot, term frequency, plotting)를 사용하여 텍스트 분석을 수행할 수 있다. 본 발명에 따르면, 인플루엔자 패턴 예측 장치는 지리 공간 데이터 분석 개념을 사용하여 독감에 영향을 받는 영역을 예측하는데 사용하여 예측 정확도의 수준을 향상시킬 수 있다.
- [0030] 도 1은 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치의 개략적인 기능 블록도, 도 2는 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치에서의 데이터 처리 흐름을 도시한 도면, 도 3은 본 발명의 일 실시예에 따른 무감독 기계 학습 알고리즘의 개략적인 구조도이다.
- [0031] 도 1 내지 도 2를 참조하면, 본 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 장치는, 데이터 수집부(110), 전처리부(120), 데이터 형성부(130) 및 데이터 분석부(140)를 포함한다.
- [0032] 데이터 수집부(110)는 소셜 네트워크에서 수집하고자 하는 특정 해시태그에 대한 데이터를 수집한다. 즉, 데이터 수집부(110)는 특정 데이터베이스를 기초로 소셜 네트워크 서비스로부터 특정 해시태그를 기반으로하는 데이터를 수집한다. 다시 말해, 데이터 수집부(110)는 다양한 소셜 네트워크 서버에 접속하여 수집하고자 하는 특정 해시태그에 대한 데이터를 수집할 수 있다. 이때, 소셜 네트워크 서버로부터 제공되는 데이터는 가공되지 않은 상태의 데이터 즉, 미가공 데이터(raw data)이다. 그리고, 소셜 네트워크 서버로부터 제공되는 데이터는 복수의 계정별 사용자가 소셜 네트워크 서버 상에 업로드하는 텍스트 데이터를 의미할 수 있다. 한편, 데이터 수집부(110)는 인플루엔자 패턴 예측 분석을 위한 소셜 네트워크 상의 데이터를 효율적으로 검출하고 제공하기 위해, 원본 데이터를 관계형 데이터베이스 형태로 저장할 수 있다. 여기서, 소셜 네트워크 서버는 트위터와 같은 SNS 서버를 비롯하여 각종 블로그, 지식인(네이버), 카페(다음) 및 게시판과 같이 다양한 형식의 소셜 네트워크(소

설 채널이라고도 함) 서버를 모두 포함할 수 있으며, 본 실시예를 설명함에 있어서 소셜 네트워크는 트위터일 수 있다. 데이터 수집부(110)는 기능적으로 시스템 데이터(system data), 크롤링(crawling), 분류(classification) 및 기록 보관(archive) 기능을 포함하고, 이를 통해 온라인 사이트(예: 블로그, 카페, SNS 등) 실시간 수집 기술을 제공할 수 있다. 데이터 수집부(110)는 하나의 프로세서일 수 있으며, 트위터 API(Application Programming Interface)에서 특정(hash tag)(flu)에 대한 트위터 트윗을 수집하여 데이터를 수집할 수 있다.

[0033] 전처리부(120)는 수집된 데이터를 전처리한다. 전처리부(120)는 데이터 마이닝, 텍스트 마이닝, NLP 라이브러리 등과 같은 모듈을 사용하여 수집된 데이터를 전처리할 수 있다. 다시 말해, 전처리부(120)는 소셜 네트워크 서버로부터 수집된 미가공 데이터를 상술한 바와 같은 전처리 모듈을 이용하여 전처리 작업을 수행함으로써 가공된 데이터를 생성할 수 있다.

[0034] 본 발명에 따른 인플루엔자 패턴 예측 장치의 전처리부(120)의 보다 상세한 전처리 과정을 설명하면 다음과 같다.

[0035] 트위터는 가장 발전된 소셜 조직 웹 페이지이므로 지구상의 다른 지점에 연결된 다양한 온라인 저널로 구성된다. 따라서, 전체 온라인 저널을 가져오는 대신 특정 지점(예 : 인플루엔자)을 보고 당시에 모든 페이지를 다운로드하여 마이닝 장치를 사용해 콘텐츠 문서로 만들었다. 따라서, 트윗을 제거할 때 R-studio를 통해 트위터 API 코드를 실행해야 한다. 트위터 서치 응용 프로그램 인터페이스 코드는 R 콘솔에 실행해야 한다. 트위터 트윗에 대한 인터페이스를 소유하기 위해서는 트위터 사이트에 대한 연관성이 밝혀져야 한다. 그때 우리는 트윗을 스캔하여 CSV 레코드로 남겨두어야 한다. 정보 사전 준비가 수행된다. 트위터 API 프로세스에서 정보 사전 처리를 수행하려면 먼저 R의 도입 순서를 통해 수많은 R 번들을 도입하고 라이브러리 명령을 통해 R로 가져와야 한다. 데이터 처리는 텍스트 처리를 통해 이루어진다. R 패키지에 #influenza 태그가 제공되면 정확한 데이터(인플루엔자 독감 데이터)가 획득된다. 따라서, 품질 문장 설명이 획득된다. 수집된 정보의 준비에 대해 말한다. 트위터를 통해 지식을 얻었으면, 다음 단계에는 일반적으로 사전 준비, 제어, 청소, 디자인 및 분리가 포함된다. 결과를 회복하기 위해서는 전처리 전략이 필수적이다. 우리는 테스트되지 않은 정보로 더 활용되는 트위터로부터의 짧은 메시지를 위해 트윗을 추출한다. 더 나아가, 이것은 사전처리 될 수 있다.

[0036] 데이터 형성부(130)는 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성한다.

[0037] 본 발명에 따른 인플루엔자 패턴 예측 장치의 데이터 형성부(130)가 무감독 기계 학습 알고리즘(Unsupervised Machine Learning Algorithm)(도 3 참조)을 사용하여 텍스트 데이터를 훈련하고 테스트하는 절차는 다음과 같다.

- [0038] 1. 먼저, 텍스트 데이터에 시드를 설정한다.
- [0039] 2. 종자를 기준으로 텍스트 데이터를 훈련시킨다.
- [0040] 3. 머신 러닝을 사용하여 텍스트 데이터를 학습한 후
- [0041] 4. 그런 다음 훈련된 데이터를 테스트한다.
- [0042] 5. 분류를 위해 랜덤 포레스트 알고리즘을 적용한다.

[0044] 데이터 분석부(140)는 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석한다. 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함될 수 있다. 데이터 분석부(140)는 벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석할 수 있다. 지리 공간 데이터 분석(Geo-Spatial Data Analytics)은 일반적으로 인플루엔자 바이러스의 영향을 받는 영역을 예측하는데 도움이 된다. 지리적 관점에 있거나 위치가 있는 데이터를 지리 공간 정보라고 한다. 이러한 종류의 데이터에는 공간에서의 상황을 보여주는 정보 또는 관련 위치가 있다. 지리 공간 정보는 벡터 기반 및 래스터 기반의 두 가지 기본 구조로 제공된다. 시각적 가이드와 관련한 정보를 보면 상황이 어떻게 발전하고 있으며, 그러한 상황에 대응하는 방법이 더 명확 해진다. 공간 조건이 지속적으로 변화하는 방식을 살펴보면 보다 쉽게 변화를 계획하고 향후 활동을 결정할 수 있다.

[0045] 본 발명에 따른 인플루엔자 패턴 예측 장치의 데이터 분석부(140)에 대해 보다 자세히 설명하면 다음과 같다.

[0046] 데이터 분석부(140)는 훈련된 데이터를 얻은 부분으로 문서의 텍스트 데이터를 분석하기 위해 랜덤 포레스트 알고리즘(Random Forest Algorithm)을 적용한다. 훈련된 데이터를 사용하여 텍스트를 결정할 수 있다. 이때, 훈련된 데이터는 유사한 매트릭스 형식을 갖는다. 데이터 집합의 각 텍스트, 일반적으로 랜덤 포레스트 알고리즘(Random Forest Algorithm)은 유사성 매트릭스에서 작동하고 매트릭스를 기반으로 텍스트 리뷰를 계산한다. 이는 랜덤 포레스트 분류 예측 모델의 결과이므로, 결과적으로 우리는 총 계산된 유사성 행렬을 얻으며, 이러한 결과들과 함께 예측 모델의 확률 값이 결정될 수 있는 정확도를 찾기 위해 수동 계산이 수행될 수 있다. 한편, 벡터 기반 지리 공간 데이터 분석에서는, 점, 선, 또는 다각형 데이터 및 모양 파일과 통신하는 데이터 파일로 구성된다. 형태(shape) 파일 디자인은 데이터를 포커스, 선 및 다각형과 같은 조잡한 기하학적 구조로 저장한다. 각 구조와 관련된 데이터 속성과 결합된 이러한 구조는 지리 데이터를 나타낸다. 표현 형태 파일은 매우 정상적이지만 조직은 일반적인 파일 이름 접두사가 있는 여러 레코드로 구성되며 유사한 색인에 저장된다. 세 개의 필요한 레코드는 파일 이름 확장자가 .shp, .shx, 및 .dbf이다. 인증된 형태 문서는 명시적으로 .shp 레코드와 관련이 있지만 다른 지원 보고서가 필요하기 때문에 흩어지기에는 부족하다. 필수 문서(.shp)에는 형상 데이터가 포함되어 있다. 두 개의 접는 문서에는 단일 고정 길이 헤더와 임의의 속도로 가변 길이 레코드가 포함된다. 모든 가변 길이 레코드는 레코드 헤더 부분과 레코드 물질 세그먼트에 결합된다.

[0047] 임의 포레스트(random forests)에 대한 준비 계산은 부트스트랩 집계의 전체 방법과 관련이 있다. 응답이 $\beta = \beta_1, \dots, \beta_p$ 인 훈련 그룹 $a = a_1, \dots, a_i$ 이 별도로 구현되고 다음 예제로 설정된 준비를 대체하여 임의의 그림을 선택한다.

[0048] $= 1, \dots,$

[0049] $a_?, \beta_?$ 와 같은 a, β 에서 모델을 준비하여 n 을 시험한다. $a_?, \beta_?$ 에서 배열 또는 재발 나무 $?$ 를 훈련시킨다. 구성의 교반에서, 숨겨진 파라미터 x' 에 대한 전망은 x' 에 대한 모든 개별 재발 트리로부터 예측을 찾아서 빈번하게 이루어진다.

[0050]
$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

[0051] 이 부트 스트랩 시스템은 경향을 확장하지 않고 모델의 구별을 줄이므로 더 나은 모델 실행을 촉진한다. 이것은 고독한 나무의 추정치가 준비 그룹의 동요에 현저하게 반응하지만, 다양한 나무의 전형적인 것은 아니고 나무의 길이는 연결되어 있지 않다는 것을 암시한다. 기본적으로 단일 준비 그룹에서 다양한 나무를 준비하면 명백하게 관련된 나무(또는 준비 계산이 계산 가능한 경우 정기적으로 비교 나무)가 생길 수 있다. 부트 스트랩 분석은 나무를 확실한 준비 세트를 보여줌으로써 나무의 연결을 해제하는 전략이다.

[0052] 또한, x' 의 전체 문자 열화 나무와 예측값의 차이가 다음과 같은 점에 비추어 기대의 취약성에 대한 측정이 자주 이루어진다.

[0053]
$$\delta = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B-1}}$$

[0054] 테스트 수량 B 는 무료 매개 변수일 수 있다. 일반적으로 준비 세트에 따라 일부 트리가 사용된다. 완벽한 양의 나무 B 는 교차 승인을 통해 또는 포장을 벗어난 실수를 관찰함으로써 확립될 수 있다. 부트 스트랩 테스트에서 x_i 가 없는 나무만 모든 준비 테스트 x_i 에 대한 평균 예측 보치(botch)이다. 준비 및 테스트 실수는 귀중한 나무를 크게 떨어뜨린다.

[0055] 위의 전략은 나무의 1차 포장 계산을 나타낸다. 이 공통 배열에 따라 고독한 코스에서만 불규칙한 산림 차이: 그들은 학습 기법의 모든 지원자 분할에서 텍스처의 자기 주장적 하위 집합을 선택하는 변경된 트리 학습 계산을 사용한다. 이 전략은 현재 하이라이트 패킹이라고 불린다. 이 작업을 수행하는 이유는 일반적인 부트 스트랩 테스트에서 트리의 관계 때문이다. 하나 또는 두개 또는 세개의 기능이 대상 획득을 위한 엄청나게 강력한 포인 터인 경우, 이러한 하이라이트는 셀수없이 B 트리에서 선택되어 비교된다.

[0056] 다른 선택을 이해하고 활용하려면 등록에 대한 추가 데이터가 도움이 된다. 대부분의 대안은 예측할 수 없는 팀 버랜드(timberland)에 의해 생성된 2개의 데이터 개체에 의존한다. 현재 트리의 계획 그룹을 대체 테스트를 통해 측정할 때 약 33%의 사례가 모델과 별도로 유지된다. 이 oob 정보는 숲이 우거진 지역에 나무가 추가됨에 따라 게임 계획이 엉망이 되는 비판적인 조치를 처리하기 위해 측정된다. 또한, 가변적 포용성의 평가를 활용하는 것이 좋다. 모든 트리가 생성되면 정보의 합계가 트리 아래로 처리되고 모든 사례 그룹에 대해 영역이

계산된다. 2개의 케이스에 비교 터미널 중심점이 있으면 해당 영역이 확장된다. 이 과정이 완료되면 주변 지역은 나무의 양으로 분리되어 제도화된다. 영역은 누락된 정보를 재정의하고, 불일치를 발견하고, 정보에 대한 낮은 차원의 관점을 밝게 만드는데 사용된다.

- [0057] 상술한 바에 따르면, 인플루엔자 패턴 예측 장치는 데이터 분석 및 텍스트 데이터의 전처리를 위해 R-프로그래밍(이하, R)을 선택할 수 있다. R에서는 사전 처리를 위해 데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하고 문서에서 올바른 추천 텍스트를 얻기 위해 용어 빈도 IDF 및 코퍼스를 계산할 수 있다.
- [0059] 도 4는 본 발명의 일 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 방법의 개략적인 흐름도이다.
- [0060] 도 4를 참조하면, 본 실시예에 따른 소셜 네트워크 데이터에 기반한 인플루엔자 패턴 예측 방법은 데이터 수집 단계(S410), 전처리 단계(S420), 데이터 형성 단계(S430) 및 데이터 분석 단계(S440)를 포함할 수 있다.
- [0061] 데이터 수집 단계는, 소셜 네트워크에서 수집하고자 하는 특정 해쉬태그에 대한 데이터를 수집할 수 있다.
- [0062] 전처리 단계는, 수집된 데이터를 전처리할 수 있다. 전처리 단계는, 데이터 마이닝, 텍스트 마이닝, NLP 라이브러리를 사용하여 수집된 데이터를 전처리할 수 있다.
- [0063] 데이터 형성 단계는, 전처리된 데이터를 학습된 무감독 기계 학습 알고리즘을 이용하여 테스트하고 데이터 집합을 형성할 수 있다.
- [0064] 데이터 분석 단계는 형성된 데이터 집합을 랜덤 포레스트 알고리즘을 이용하여 데이터를 분석할 수 있다. 랜덤 포레스트 알고리즘에는, 지리 공간 데이터 분석을 위한 라이브러리가 포함될 수 있다. 데이터 분석 단계는, 벡터 기반 지리 공간 데이터 분석 방법을 활용하여 데이터를 분석할 수 있다.
- [0065] 본 발명의 실시예에 따른 방법들은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는, 본 발명을 위한 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0066] 본 명세서는 많은 특징을 포함하는 반면, 그러한 특징은 본 발명의 범위 또는 특허청구범위를 제한하는 것으로 해석되어서는 아니 된다. 또한, 본 명세서의 개별적인 실시예에서 설명된 특징들은 단일 실시예에서 결합되어 구현될 수 있다. 반대로, 본 명세서의 단일 실시예에서 설명된 다양한 특징들은 개별적으로 다양한 실시예에서 구현되거나, 적절히 결합되어 구현될 수 있다.
- [0067] 도면에서 동작들이 특정한 순서로 설명되었으나, 그러한 동작들이 도시된 바와 같은 특정한 순서로 수행되는 것으로 또는 일련의 연속된 순서, 또는 원하는 결과를 얻기 위해 모든 설명된 동작이 수행되는 것으로 이해되어서는 안 된다. 특정 환경에서 멀티태스킹 및 병렬 프로세싱이 유리할 수 있다. 아울러, 상술한 실시예에서 다양한 시스템 구성요소의 구분은 모든 실시예에서 그러한 구분을 요구하지 않는 것으로 이해되어야 한다. 상술한 앱 구성요소 및 시스템은 일반적으로 단일 소프트웨어 제품 또는 멀티플 소프트웨어 제품에 패키지로 구현될 수 있다.
- [0068] 이상에서 설명한 본 발명은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 있어 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 여러 가지 치환, 변형 및 변경이 가능하므로 전술한 실시예 및 첨부된 도면에 의해 한정되는 것은 아니다.

부호의 설명

[0070]

110 : 데이터 수집부

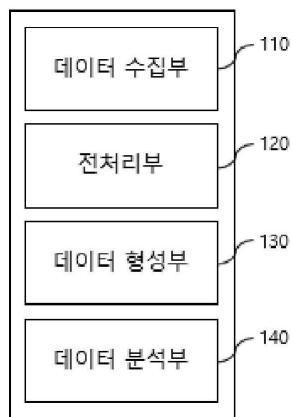
120 : 전처리부

130 : 데이터 형성부

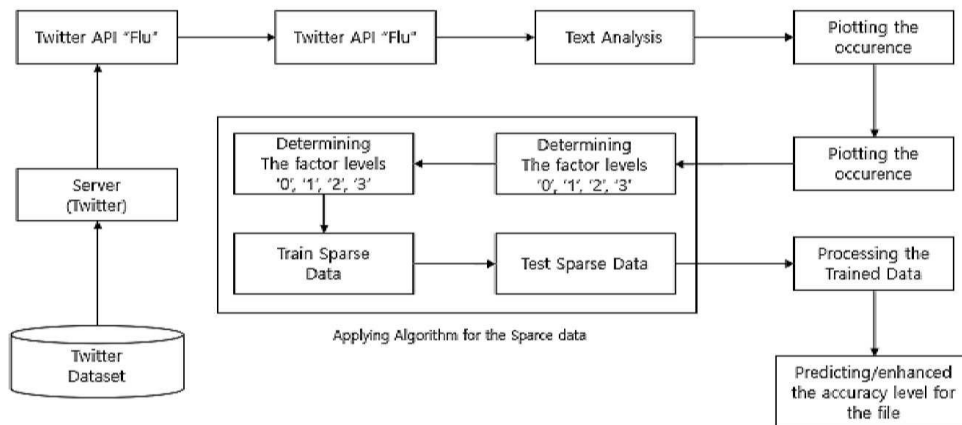
140 : 데이터 분석부

도면

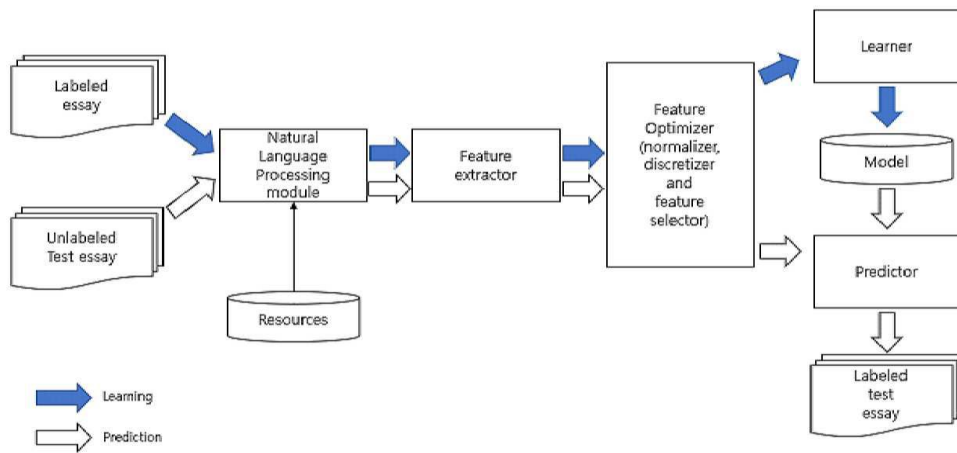
도면1



도면2



도면3



도면4

