(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2019 년 1 월 17 일 (17.01.2019)

WIPO | PCT

(10) 국제공개번호
WO 2019/013376 A1

(51) 국제특허분류:
G06F 17/18 (2006.01) *G06F 17/30* (2006.01)

(21) 국제출원번호: PCT/KR2017/007873

(22) 국제출원일: 2017 년 7 월 21 일 (21.07.2017)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2017-0089573 2017 년 7 월 14 일 (14.07.2017) KR

(71) 출원인: 한국과학기술원 **(KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY)** [KR/KR]; 34141 대전시 유성구 대학로 291, Daejeon (KR).

(72) 발명자: **문일철 (MOON, Il-Chul)**; 34141 대전시 유성구 대학로 291(구성동), Daejeon (KR). **박준건 (PARK, Junkeon)**; 34141 대전시 유성구 대학로 291(구성동), Daejeon (KR).

(74) 대리인: **특허법인신우 (SHINWOO PATENT)**; 06132 서울시 강남구 테헤란로 21길 8, 6층(역삼동), 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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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).

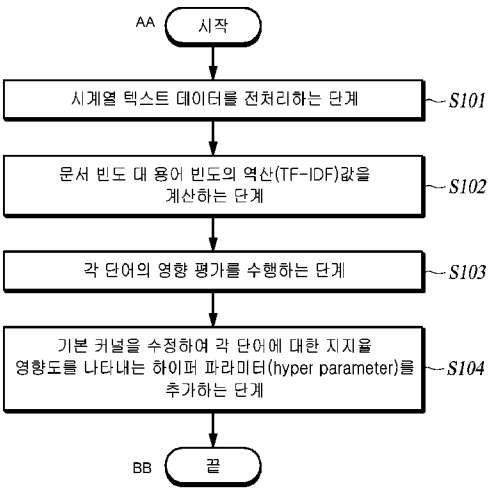
공개:
— 국제조사보고서와 함께 (조약 제21조(3))

WO 2019/013376 A1

(54) Title: METHOD AND DEVICE FOR PREDICTING APPROVAL RATING BY USING TEXT-COMPENSATED AUTOMATIC STATISTICAL MODEL

(54) 발명의 명칭: 텍스트 보완된 자동 통계학자 모델을 이용한 지지율 예측 방법 및 장치

[도1]



S101 ... Step of pre-processing sequential text data
S102 ... Step of calculating term frequency-inverse document frequency (TF-IDF) value
S103 ... Step of evaluating impact of each word
S104 ... Step of adding hyper parameter indicating approval rating impact for each word by correcting basic kernel
AA ... Start
BB ... End

(57) Abstract: Provided is a method by which a computer predicts an approval rating by applying a multi-dimensional Gaussian process to sequential text data. The method of the present invention comprises the steps of: pre-processing the sequential text data; and applying the multi-dimensional Gaussian process to the pre-processed sequential text data, wherein the step of pre-processing the sequential text data comprises the steps of: calculating a term frequency-inverse document frequency (TF-IDF) value; evaluating an impact of each word; and adding a hyper parameter indicating an approval rating impact for each word by correcting a basic kernel.

(57) 요약서: 컴퓨터에 의해 수행되는, 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하여 지지율을 예측하는 방법이 제공된다. 본 발명의 방법은, 시계열 텍스트 데이터를 전처리하는 단계와, 상기 전처리된 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하는 단계를 포함하고, 시계열 텍스트 데이터를 전처리하는 단계는 문서 빈도 대 용어 빈도의 역산(TF-IDF) 값을 계산하는 단계와, 각 단어의 영향 평가를 수행하는 단계와, 기본 커널을 수정하여 각 단어에 대한 지지율 영향도를 나타내는 하이퍼 파라미터(hyper parameter)를 추가하는 단계를 포함한다.

명세서

발명의 명칭: 텍스트 보완된 자동 통계학자 모델을 이용한 지지율 예측 방법 및 장치

기술분야

[1] 본 발명은 지지율 예측 방법 및 장치에 관한 것으로, 보다 구체적으로는 시계열 텍스트 데이터를 다차원 가우시안 프로세스에 적용하여 정치인의 지지율을 예측하는 방법 및 장치에 관한 것이다.

배경기술

[2] 정치인의 지지율을 예측하는 것은 대중들이 지속적으로 관심을 가지는 분야이다. 선거 기간 동안의 지지율 예측은 다른 사람들의 의견을 알고 싶어하는 유권자나 정치인의 캠페인 전략에 영향을 미치게 된다. 대통령 후보 지지율은 직접적으로는 공개되는 공개 설문 조사는 가장 좋은 방법이지만, 공개 설문 조사는 시간과 비용이 상당히 들기 때문에 간접적으로 지지율을 알 수 있는 방법이 필요하다.

[3] 과거의 연구에서는 이렇게 특정 인물의 지지율과 상관되는 변수들을 모델링하여 지지율을 예측해보는 연구는 있었지만 그 변수들은 특정 도메인에만 한정되었다. 또한 그 변수들과 지지율의 상관관계를 확인하였지만 지지율에 대한 수치적인 접근법이 부족하였다. 따라서, 뉴스 기사와 같은 지지율 이외의 출처로부터 복잡한 비선형 경향을 포착할 수 있으며 정치적 맥락을 연결시킬 수 있는 예측 모델을 개발할 필요성이 있는데, 자동 통계학자 모델은 변화가 발생한 원인에 대해 설명하는 것이 어렵다는 문제점이 있다.

발명의 상세한 설명

기술적 과제

[4] 따라서, 본 개시는 커널 구성 프로세서 상에서 텍스트 정보를 보완함으로써 자동 통계학자 모델을 확장하여 정치인 지지율을 예측하는데 적용가능하다.

[5] 본 개시는 자동 베이지안 공분산 발견(automatic bayesian covariance discovery)의 이점을 이용함으로써 복잡하고 비선형적인 트랜드에 적용이 가능하다.

[6] 본 개시는 텍스트 보완을 하는데 어떤 커널 기능을 포함함에 있어 어떤 텍스트 입력(단어)이 중요한지를 제공할 수 있다.

과제 해결 수단

[7] 본 발명의 일 실시에 따르면, 컴퓨터에 의해 수행되는, 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하여 지지율을 예측하는 방법이 제공된다. 본 방법은 시계열 텍스트 데이터를 전처리하는 단계와, 상기 전처리된 시계열 텍스트 데이터를 다차원 가우시안 프로세스에 적용하는 단계를 포함할 수 있다.

[8] 일 실시예에서, 시계열 텍스트 데이터를 전처리하는 단계는 문서 빈도 대 용어

빈도의 역산(TF-IDF) 값을 계산하는 단계와, 각 단어의 영향 평가를 수행하는 단계와, 기본 커널을 수정하여 각 단어에 대한 지지율 영향도를 나타내는 하이퍼 파라미터(hyper parameter)를 추가하는 단계를 포함할 수 있다.

[9] 본 발명의 다른 측면에서, 컴퓨터에 의해 수행되는, 다차원 가우시안 프로세스를 적용하여 지지율을 예측하는 방법이 제공된다. 본 방법은 문서 빈도 대 용어 빈도의 역산(TF-IDF) 값을 계산하는 단계와, 각 단어의 영향 평가를 수행하는 단계와, 상기 영향 평가를 수행한 각 단어를 다차원 가우시안 프로세스에 적용하는 단계를 포함할 수 있다.

[10] 일 실시예에서, 상기 기본 커널은 Constant (CON) 커널이고, 수정된 CON 커널은 다차원 각각에 대한 상수를 지정할 수 있다.

[11] 일 실시예에서, 상기 기본 커널은 PER 커널이고, 수정된 PER 커널은 각각의 단어에 대한 주기적이고 길이 스케일의 하이퍼 파라미터를 지정하고, 하이퍼 파라미터로 구성된 PER 커널 함수의 용어를 곱일 수 있다.

[12] 일 실시예에서, 상기 기본 커널은 LIN 커널이고, 수정된 LIN 커널은 각 단어에 하이퍼 파라미터를 할당하고, 하이퍼 파라미터로 구성된 LIN 커널 함수의 용어를 추가할 수 있다.

[13] 일 실시예에서, 상기 각 단어의 영향 평가를 수행하는 단계는 상기 문서 빈도 대 용어 빈도의 역산(TF-IDF) 값이 상위인 사전 정의된 수의 단어를 선정하여 영향 평가를 수행할 수 있다.

[14] 본 방법의 다른 실시예에 따르면, 컴퓨터에 의해 수행되는, 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하는 방법이 제공된다. 본 방법은 시계열 텍스트 데이터를 전처리하는 단계와, 상기 전처리된 시계열 텍스트 데이터를 다차원 가우시안 프로세스에 적용하는 단계와 베이지안 정보 기준(BIC) 측정 값을 사용하여 모델을 정규화하는 단계를 포함할 수 있다.

[15] 본 방법의 다른 측면에 있어서, 컴퓨터 장치로서, 상기 기재된 방법을 실행하도록 동작 가능한 컴퓨터 장치가 제공될 수 있다.

[16] 본 방법의 또 다른 측면에 있어서, 하나 이상의 명령어가 저장된 컴퓨터 판독 가능한 저장 매체로서, 상기 하나 이상의 명령어는 컴퓨터에 의해 실행되는 경우, 컴퓨터로 하여금 상기 방법을 실행하게 하는, 컴퓨터 판독 가능한 저장 매체가 제공될 수 있다.

발명의 효과

[17] 본 개시에 의하면, 자동 베이지안 공분산 발견의 이점을 이용함으로써 복잡하고 비선형적인 트랜드에 적용 가능한 정치인 지지율 예측 방법이 제공될 수 있다.

[18] 본 개시에 의하면, 텍스트 보완을 하는 커널 기능을 포함함에 있어 어떤 텍스트 입력(단어)이 중요한지를 제공하는 정치인 지지율 예측 방법이 제공될 수 있다.

도면의 간단한 설명

[19] 도 1은 본 발명의 일 실시예에 따라, 본 발명의 방법이 수행되는 흐름도를 도시하는 흐름도이다.

[20] 도 2는 종례의 SE 커널, PER 커널 및 LIN 커널 함수의 일 예를 도시하는 그래프이다.

[21] 도 3은 본 발명의 일 실시예에 따라, 4명의 후보 각각 대한 지지율 예측을 나타내는 그래프이다.

[22] 도 4a 내지 도 4c는 본 발명의 일 실시예에 따라, 피팅 결과값의 정량적인 평가값을 도시하는 그래프이다.

[23] 도 5a는 본 발명의 일 실시예에 따라, 후보 1의 하이퍼 파라미터 분석을 나타내며, 하이퍼 파라미터 값의 히스토그램이다.

[24] 도 5b는 본 발명의 일 실시예에 따라, 후보 1의 하이퍼 파라미터 분석을 나타내며, 각각의 입력에 대한 하이퍼 파라미터 값의 그래프이다.

발명의 실시를 위한 형태

[25] 이하, 첨부 도면을 참조하여 본 발명의 실시예에 관하여 상세히 설명한다. 이하에서는, 본 발명의 요지를 불필요하게 흐릴 우려가 있다고 판단되는 경우, 이미 공지된 기능 및 구성에 관한 구체적인 설명을 생략한다. 또한, 이하에서 설명하는 내용은 어디까지나 본 발명의 일 실시예에 관한 것일 뿐 본 발명이 이로써 제한되는 것은 아님을 알아야 한다.

[26] 본 명세서에서 사용되는 용어는 단지 특정한 실시예를 설명하기 위해 사용되는 것으로 본 발명을 한정하려는 의도에서 사용된 것이 아니다. 예를 들면, 단수로 표현된 구성요소는 문맥상 명백하게 단수만을 의미하지 않는다면 복수의 구성요소를 포함하는 개념으로 이해되어야 한다. 또한, 본 발명의 명세서에서, '포함하다' 또는 '가지다' 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것일 뿐이고, 이러한 용어의 사용에 의해 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하려는 것은 아니다.

[27] 본 명세서에 기재된 실시예에 있어서 '블록' 또는 '부'는 적어도 하나의 기능이나 동작을 수행하는 기능적 부분을 의미하며, 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다. 또한, 복수의 '블록' 또는 '부'는, 특정한 하드웨어로 구현될 필요가 있는 '블록' 또는 '부'를 제외하고는, 적어도 하나의 소프트웨어 모듈로 일체화되어 적어도 하나의 프로세서로 구현될 수 있다.

[28] 덧붙여, 달리 정의되지 않는 한 기술적 또는 과학적인 용어를 포함하여, 본 명세서에서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의된 용어들은, 관련 기술의 문맥상 의미와

일치하는 의미를 갖는 것으로 해석되어야 하며, 본 발명의 명세서에서 명백하게 달리 정의하지 않는 한 과도하게 제한 또는 확장하여 해석되지 않는다는 점을 알아야 한다.

[29] 이하, 첨부된 도면을 참조하여, 본 발명의 실시예에 대해 구체적으로 설명하기로 한다.

[30] 본 발명은 정치와 관련된 뉴스 기사를 사용하여 정치인들의 지지율을 예측하는 방법에 관한 것이다.

[31] 도 1은 본 발명의 일 실시예에 따른, 뉴스 기사를 사용하여 정치인들의 지지율을 예측하는 방법의 흐름도이다.

[32] 단계(S101)에서, 시계열 텍스트 데이터를 전처리할 수 있다. 일 실시예에서, 시계열 텍스트 데이터의 전처리는 스피치 태깅, POS 태깅을 일부 사용하여, 뉴스 기사에서 명사를 뽑아내는 것일 수 있다. 일 실시예에서, 키워드를 포함한 신문 기사 중 연관된 키워드를 포함한 신문 기사만을 수집하여 형태소 분석을 통해 명사에 해당하는 단어만을 추출할 수 있다.

[33] 이후 단계(S102)에서는 각각의 명사에 대해서, 문서 빈도 대 용어 빈도의 역산(TF-IDF) 값을 계산할 수 있다. 여기서, TF-IDF는 코퍼스에 있는 문서에서 주어진 단어의 중요성을 나타내는 점수를 나타낸다. 높은 TF-IDF 값을 가지는 상위 500개의 단어가 자동 베이지안 공분산 발견(automatic bayesian covariance discovery: ABCD)의 입력으로 선택될 수 있는데, 입력은 높은 TF-IDF 값에 의해 선택된 단어의 정규화된 발생 빈도 값이다.

[34] 단계(S103)에서, 각 단어의 영향 평가를 수행할 수 있다. 일 실시예에서, 각 단어의 영향 평가는 텍스트 입력을 ABCD 모델에 적용하여 수행될 수 있다. 본 개시의 단계(S104)에서는 기본 커널을 수정하여 ABCD 모델을 적용한다. 수정된 기본 커널은 기본 커널과 유사한 형태를 가지면서, 차별점으로 수정된 기본 커널은 각 단어에 대해 지지율에 영향을 나타내는, 즉 대통령 후보 지지율 예측하기 위한 각 단어의 영향 평가하는 하이퍼 파라미터(hyper-parameter)를 가지고 있다.

[35] 본 발명의 일 실시예에서, SE(squared exponential) 커널의 평탄화(smoothness)로 인하여 단어의 영향력이 줄어드는 것을 방지하기 위해 SE 커널은 제외할 수 있다.

[36] 본 발명은 기본 커널 함수로 Constant (CON) 커널, PER 커널 및 LIN 커널을 수정하여 사용할 수 있고, 데이터에 대한 커널의 최상의 구조를 찾기 위해 커널 연산에 덧셈과 곱셈을 적용하여 커널을 확장할 수 있다.

[37] 일 실시예에 있어서, CON 커널 함수는 상수 함수이다. 다차원 공간으로 CON 커널을 수정하기 위해서, 우리는 각 차원에 대한 상수를 지정하고, 상수를 추가할 수 있다. 그러나 본 발명에서는 훈련 기간(training period)에서는 상수를 차별화하지 않았고, 상수는 모든 차원에 걸쳐 동일하게 두었다. 본 발명에서 각각의 차원에 대해서 차별화된 훈련을 각각 실험하였으나, 그 차이는 크지

않았다.

[38]

$$k_{con}(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{D} \sigma_i^2 = \sigma_{total}^2$$

[39] 여기서, X1와 X2는 커널함수의 입력값으로 하루동안 생성된 문서에서 선택된 단어의 빈도수이다. 또한 i는 각 차원(혹은 단어)의 인덱스이며 D는 총 차원수(선택된 단어의 수)이고, σ_i 는 그 차원의 상수값을 의미한다.

[40] 일 실시예에 있어서, PER 커널 함수는 함수의 반복 동작을 의미할 수 있다. 각 차원의 다른 동작을 포착하기 위해, 각각의 단어에 대한 주기적이고 길이 스케일의 하이퍼 파라미터를 지정하고, 하이퍼 파라미터로 구성된 PER 커널 함수의 용어를 곱한다. 각각의 단어의 주기적인 하이퍼 파라미터는 각 단어의 주기를 결정하고, 각각의 길이 스케일의 하이퍼 파라미터는 각 단어의 함수의 변화의 길이를 결정할 수 있다. 다음 수식은 수정된 다차원 PER 커널을 나타낸다.

[41]

$$k_{per}(\mathbf{x_1}, \mathbf{x_2}) = \prod_{i=1}^{D} \sigma_i^2 \exp\left(-\frac{2\sin^2(\pi|\mathbf{x_{1i}} - \mathbf{x_{2i}}|/p_i)}{l_i^2}\right)$$

$$= \sigma_{total}^2 \exp\left(\sum_{i=1}^{D} -\frac{2\sin^2(\pi|\mathbf{x_{1i}} - \mathbf{x_{2i}}|/p_i)}{l_i^2}\right)$$

[42] 여기서, X1와 X2는 커널함수의 입력값으로 하루 동안 생성된 문서에서 선택된 단어의 빈도수이다. 또한, i는 각 차원(혹은 단어)의 인덱스이며 D는 총 차원수(선택된 단어의 수)이다. σ_i 는 그 차원의 가중치를 의미하며, Pi 는 주기를 나타내고, Ii는 진폭을 의미한다.

[43] 일 실시예에서, LIN 커널 함수는 함수의 선형성을 의미할 수 있다. 마찬가지로 각 단어의 선형성을 포착하기 위해 각 단어에 하이퍼 파라미터를 할당하고, 하이퍼 파라미터로 구성된 LIN 커널 함수의 용어를 추가할 수 있다. 기울기 파라미터는 단어 빈도가 어떻게 지지율에 미치는 영향을 변하는지를 제공할 수 있다. 다음의 수식은 수정된 다차원 LIN 커널을 나타낸다.

[44]

$$k_{lin}(\mathbf{x_1}, \mathbf{x_2}) = \sum_{i=1}^{D} \sigma_{bi}^2 + \sigma_{vi}^2(\mathbf{x_{1i}} - c_i)(\mathbf{x_{2i}} - c_i)$$

$$= \sigma_{b,total}^2 + \sum_{i=1}^{D} \sigma_{vi}^2(\mathbf{x_{1i}} - c_i)(\mathbf{x_{2i}} - c_i)$$

[45] 여기서, X1과 X2는 커널함수의 입력값으로 하루동안 생성된 문서에서 선택된 단어의 빈도수이다. 또한 i는 각 차원(혹은 단어)의 인덱스이며 D는 총 차원수(선택된 단어의 수)이다. σ_{bi} 와 Ci 는 y-절편과 x-절편을 의미하며 σ_{vi} 는 기울기를 나타낸다.

[46] 일 실시예에 있어서, 주어진 데이터에 대해 가장 적합한 커널 구조를 선택하고 확장하기 위해서, 트리 검색 방법 및 베이지안 정보 기준(BIC) 측정 값이 사용될 수 있다. 일 실시예에서, 최고의 커널 구조를 위해, 트리 검색의 경우, 모든 가능한 기본 커널 함수의 후보 및 연산이 고려되었고, 검색은 가장 낮은 BIC 를 가지는 후보들 중 커널 구조를 선택했다. 모델은 모델이 복잡한 구조를 가질 때 학습 데이터가 더 잘 맞아, 과적합(over-fitting)문제를 방지하기 위해 모델을 정규화하였다. BIC 는 여러 파라미터를 사용하여 정규화를 고려하는 방법 중 하나이다. BIC는 공식적으로 다음과 같이 정의된다.

[47]

$$BIC = -2\ln(\hat{L}) + \ln(n)k$$

[48] 여기서 $\hat{L}$ 은 모델에 대한 가능성의 최대값이고, n은 훈련 데이터 중 여러 데이터 포인트를 나타내며, k는 모델에 포함된 파라미터의 개수이다.

[49]

[50] 본 발명의 실험 결과는 아래와 같다.

[51] A. 데이터 셋

[52] 2017년 1월 1일부터 2017년 2월 7일까지의 대통령 후보의 지지율을 수집하고, 같은 기간 YTN의 정치 신문 기사 11476건을 수집하였다. POS 태커로부터, 상기 신문기사로부터 18622 개의 명사를 추출하고, 각 단어의 TF-IDF 값을 계산하였다, 상기 명사 중에 우리는 높은 TF-IDF 값을 가지는 상위 500 개의 단어를 시계열 정보의 입력으로 선택하였다. 마지막으로, 2017 한국 대통령 선거의 상위 4명의 대통령 후보를 선택하여 우리의 모델의 성능을 테스트하였다.

[53] 본 발명의 모델의 성능을 테스트 하기 위하여, 학습 기간과 테스트 기간을 나누었다. 전체 기간의 80%를 학습기간으로 하였고, 2017년 1월 1일부터 2017년 2월 1일까지이다. 학습기간의 데이터를 이용하여 모델을 학습시켰고, 모델의

결과를 테스트 기간이 실제 값들과 비교하였다.

[54] B. 정량적 평가

[55] 도 2는 종례의 SE 커널, PER 커널 및 LIN 커널 함수의 일예를 도시하는 그래프이다. x 축은 입력값이고, y축은 커널 함수에 x의 값을 넣었을 때 나오는 확률이다. 서로 다른 색은 각 입력값을 넣었을 때의 결과값을 의미한다.

[56] 도 3은 전체 기간의 실제 값, 예측값 및 신뢰구간(confidence interval)을 도시한다. 설명하면, 도 3은 피팅 결과의 정량적인 평가를 나타낸다. 예상대로, 입력 단어의 수가 커지면 BIC 점수는 작아지나, 단어 사이즈가 100 이상이 될 때까지는 BIC 점수가 높게 유지된다. 이는 지지율을 예측하기 위해 의미 있는 단어의 수와 리스트를 나타낼 수 있다.

[57] 테스트 기간 동안, 실제값의 대부분은 후보 3을 제외하고 신뢰구간에 포함된된다. 또한, 후보 1 은 주기적으로 볼 수 있는 변동에서 가장 높은 지지율을 보였고 우리의 모델은 주기적 움직임을 커널 함수 확장 후에 주기적 커널 함수로서 캡처할 수 있었다.

[58]

	Kernel Structure	Input Size
Candidate 1	PER + CON	101
Candidate 2	LIN * CON	301
Candidate 3	LIN	301
Candidate 4	PER * CON	201

[59] 표 1: 각 후보에게 있어 가장 성능이 좋은 커널 구조와 입력 크기를 나타낸다. 가장 성능이 좋은 커널 구조는 가장 작은 BIC 점수를 제공하고 입력 크기가 가장 적절한 것은 실험 데이터에 있어 가장 작은 평균절대오차(Mean Absolute Error: MAE)값을 제공한다.

[60] 또한, 평균절대오차(MAE) 및 평균제곱근오차(Root Mean Square Error: RMSE)는 ABCD 모델에 단어를 통합할 때 상대적인 성능을 나타낸다. 입력 사이즈가 하나라는 것은 입력이 텍스트 입력을 제외한 단순히 시계열적 정보라는 것을 의미한다. MAE 및 RMSE는 일반적으로 학습 시기에 단어의 수를 증가시킬 때 일반적으로 감소한다. 그러나 테스트 시기의 MAE 및 RMSE는 단어의 수가 특정 지점(즉, 후보 1의 경우 11 단어)이 될 때까지 상대적으로 평평한 구조를 나타낸다. 이는 적절한 단어의 셋을 초과할 때 해당 모델이 오퍼빗될 수 있음을 의미할 수 있다. 일부는 이 점이 본 결과의 타당성을 부정한다고 주장할 수 있겠으나, 테스트는 단어의 집합 크기가 제한될 때 상대적으로 만족된다라고 해석되어야 한다. 또한, 이 모델의 진정한 가치는 주기적으로 선평적이며 자동적인 트랜드 구조가 식별된 테이블 1에서 나온다. 또한, 가치는 도 4의 학습된 하이퍼 파라미터의 해석으로부터 나온다.

[61] C. 정성 분석

[62] 도 4a 내지 도 4c는 최적화된 하이퍼 파라미터 변수의 값과 분포를 나타낸다. 하이퍼 파라미터는 후보 1에 사용된 주기적 커널의 길이 파라미터이다. 값이 작으면 영향력은 짧은 주기로 나타나고 하이퍼 파라미터의 값이 크면 대응하는 차원의 영향도는 긴 주기의 사이클을 보인다. 하이퍼 파라미터의 가장 작은 값을 실험했을 때, 대응하는 차원은 최신 정보조사, 헌법 개정 등이었다. 일부 단어 (즉, 최신)는 정치적 의미 보다는 어휘적 의미가 있었으므로 이를 고려해야 한다.

[63] 도 4a에서 입력값의 크기가 50 이상이 되면 BIC가 줄어드는 경향을 보인다. 도 4b에서 점선은 훈련구간으로 20이상이면 안정적임을 보이며, 실선은 실험구간으로 입력값에 큰 영향이 없는 것으로 보임을 알 수 있다. 도 4c에서 훈련구간에서 50 이상이면 안정적임을 보이며, 실선은 실험구간으로 입력값에 큰 영향이 없는 것으로 (즉, 경향이 없는 것으로) 보인다.

[64] 그러나, 일부 단어 (즉, 헌법 수정)는 주요한 정치적 의미를 가지므로 우리의 분석은 이러한 단어는 후보 1의 지지율에 근시적인 영향을 미친다. 헌번 개정이 논의되고 후보 1이 개정안에 동의하지 않았다는 것은 사실이다 또한 개정 노력은 곧 사그러들며 이것이 영향의 짧은 주기를 만들어낸다는 것도 사실이다. 우리가 길이 스케일의 파라미터의 가장 높은 값을 실험할 때, 대응하는 차원이 귀향, 집회당, 재벌 등과 같은 단어였다. 유사하게, 일부 단어(재벌)는 한국의 사업 형태를 생성하고 후보 1에 대한 장기간 영향도에 있어서 주요한 정치적 의미를 가진다. 후보 1은 재벌이 한국 경제 발전에 방해가 된다고 주장해 왔다. 후보 1은 영향력의 긴 주기를 생성하는 1월 초에 재벌 개혁 정책을 약속했다.

[65] 도 5a는 본 발명의 일 실시예에 따라, 후보 1의 하이퍼 파라미터 분석을 나타내며, 하이퍼 파라미터 값의 빈도수를 그린 히스토그램이다.

[66] 도 5b는 본 발명의 일 실시예에 따라, 후보 1의 하이퍼 파라미터 분석을 나타내며, 각각의 입력에 대한 하이퍼 파라미터 값의 그래프이다. 23.27.39 의 단어가 나왔을 때는 하이퍼 파라미터의 값이 클수록 큰 주기의 움직임을 보인다. 작을수록 단기적 변화가 생긴다. 각 단어의 차원에 대응시켜서 봤을 때, 23, 평택, 천안함은 하이퍼 파라미터가 큰데, 이 정치인의 성향과 잘 들어 맞는다는 것을 의미한다.

[67] 텍스트 증가된 ABCD 모델로 분석한 정치적 지지율에 관한 예시를 제시한다. 원래 ABCD 모델은 시계열적 정보만을 사용하였으나, 본 발명에서는 원문 정보(textual information)를 사용하여 분석에 의미론적인 의미와 합리적 근거를 추가하였다. 본 발명의 실험은 텍스트 추가된 ABCD가 고유한 지지율의 주기적, 선형적 등의 동향을 포착할 수 있다는 것을 보여준다. 또한, 본 발명은 이러한 동향을 설명하는 키워드를 포착하고 이러한 포착된 키워드는 주어진 정치적 문맥에서 해석될 수 있다.

[68] 당업자라면 알 수 있듯이 본 발명은 본 명세서에서 기술된 예시에 한정되는 것이 아니라 본 발명의 범주를 벗어나지 않는 범위 내에서 다양하게 변형, 재구성 및 대체될 수 있다. 예를 들어, 본원에 기술된 다양한 기술들은 하드웨어

또는 소프트웨어, 또는 하드웨어와 소프트웨어의 조합에 의해 구현될 수 있다. 따라서, 본원에 따른 소프트웨어 안전성 분석을 위한 분석 머신의 특정한 양태나 부분은 범용 또는 전용 마이크로프로세서, 마이크로-컨트롤러 등에 의해 실행 가능한 하나 이상의 컴퓨터 프로그램으로 구현될 수 있다. 본 발명의 일 실시예에 따른 컴퓨터 프로그램은, 컴퓨터 프로세서 등에 의해 판독 가능한 저장 매체, 예컨대 EPROM, EEPROM, 플래시 메모리장치와 같은 비휘발성 메모리, 내장형 하드 디스크와 착탈식 디스크 같은 자기 디스크, 광자기 디스크, 및 CDROM 디스크 등을 포함한 다양한 유형의 저장 매체에 저장된 형태로 구현될 수 있다. 또한, 프로그램 코드(들)는 어셈블리어나 기계어로 구현될 수 도 있다.

[69] 본 발명의 진정한 사상 및 범주에 속하는 모든 변형 및 변경을 이하의 특허청구범위에 의해 모두 포괄하고자 한다.

청구범위

[청구항 1] 컴퓨터에 의해 수행되는, 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하여 지지율을 예측하는 방법으로서,
시계열 텍스트 데이터를 전처리하는 단계와,
상기 전처리된 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하는 단계
를 포함하는 지지율 예측 방법.

[청구항 2] 제1항에 있어서,
시계열 텍스트 데이터를 전처리하는 단계는
문서 빈도 대 용어 빈도의 역산(TF-IDF) 값을 계산하는 단계와,
각 단어의 영향 평가를 수행하는 단계와,
기본 커널을 수정하여 각 단어에 대한 지지율 영향도를 나타내는 하이퍼 파라미터(hyper parameter)를 추가하는 단계
를 포함하는 지지율 예측 방법.

[청구항 3] 컴퓨터에 의해 수행되는, 다차원 가우시안 프로세스를 적용하여 지지율을 예측하는 방법으로서,
문서 빈도 대 용어 빈도의 역산(TF-IDF) 값을 계산하는 단계와,
각 단어의 영향 평가를 수행하는 단계와,
상기 영향 평가를 수행한 각 단어에 다차원 가우시안 프로세스를 적용하는 단계
를 포함하는 지지율 예측 방법.

[청구항 4] 제2항에 있어서,
상기 기본 커널은 Constant (CON) 커널이고, 수정된 CON 커널은 다차원 각각에 대한 상수를 지정하는 지지율 예측 방법.

[청구항 5] 제2항에 있어서,
상기 기본 커널은 PER 커널이고, 수정된 PER 커널은 각각의 단어에 대한 주기적이고 길이 스케일의 하이퍼 파라미터를 지정하고, 하이퍼 파라미터로 구성된 PER 커널 함수의 용어를 곱하는 지지율 예측 방법.

[청구항 6] 제2항에 있어서,
상기 기본 커널은 LIN 커널이고, 수정된 LIN 커널은 각 단어에 하이퍼 파라미터를 할당하고, 하이퍼 파라미터로 구성된 LIN 커널 함수의 용어를 추가하는 지지율 예측 방법.

[청구항 7] 제2항에 있어서,
상기 각 단어의 영향 평가를 수행하는 단계는
상기 문서 빈도 대 용어 빈도의 역산(TF-IDF) 값이 상위인 사전 정의된 수의 단어를 선정하여 영향 평가를 수행하는 것인 지지율 예측 방법.

[청구항 8] 컴퓨터에 의해 수행되는, 시계열 텍스트 데이터에 다차원 가우시안

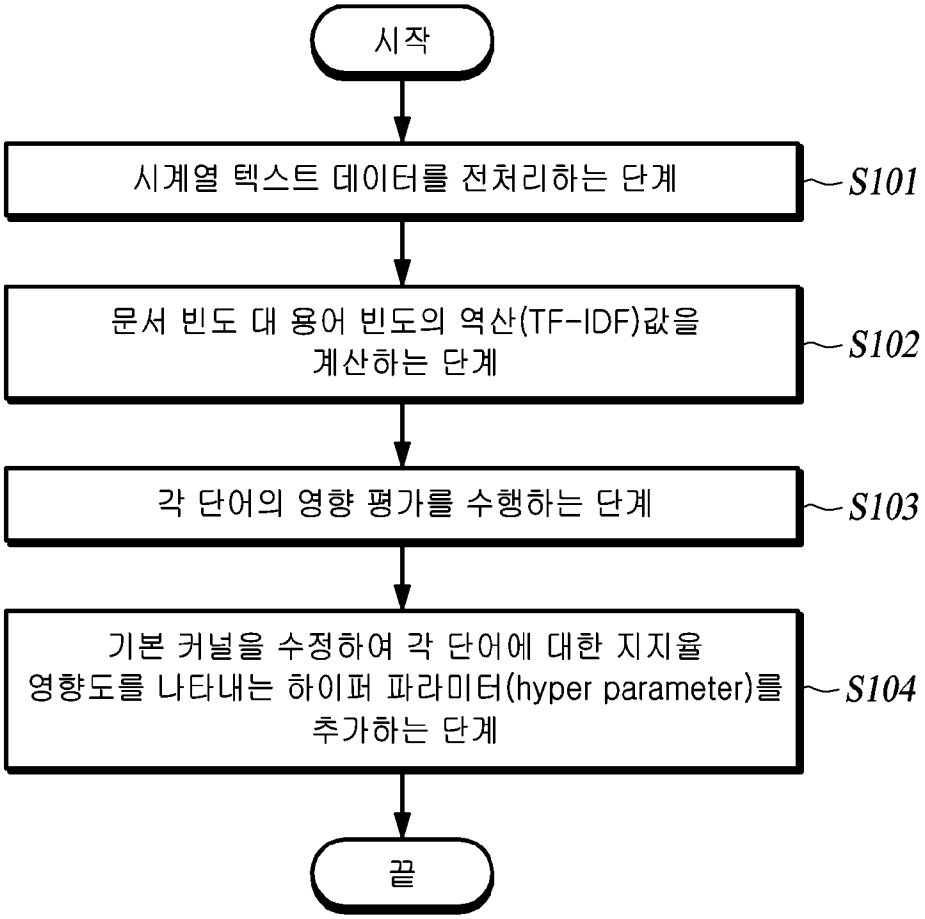
프로세스를 적용하는 방법으로서,
시계열 텍스트 데이터를 전처리하는 단계와,
상기 전처리된 시계열 텍스트 데이터에 다차원 가우시안 프로세스를 적용하는 단계와
베이지안 정보 기준(BIC) 측정 값을 사용하여 모델을 정규화하는 단계를 포함하는 지지율 예측 방법.

[청구항 9] 컴퓨터 장치로서, 제1항 내지 제8항 중 어느 한 항의 방법을 실행하도록 동작 가능한 컴퓨터 장치.

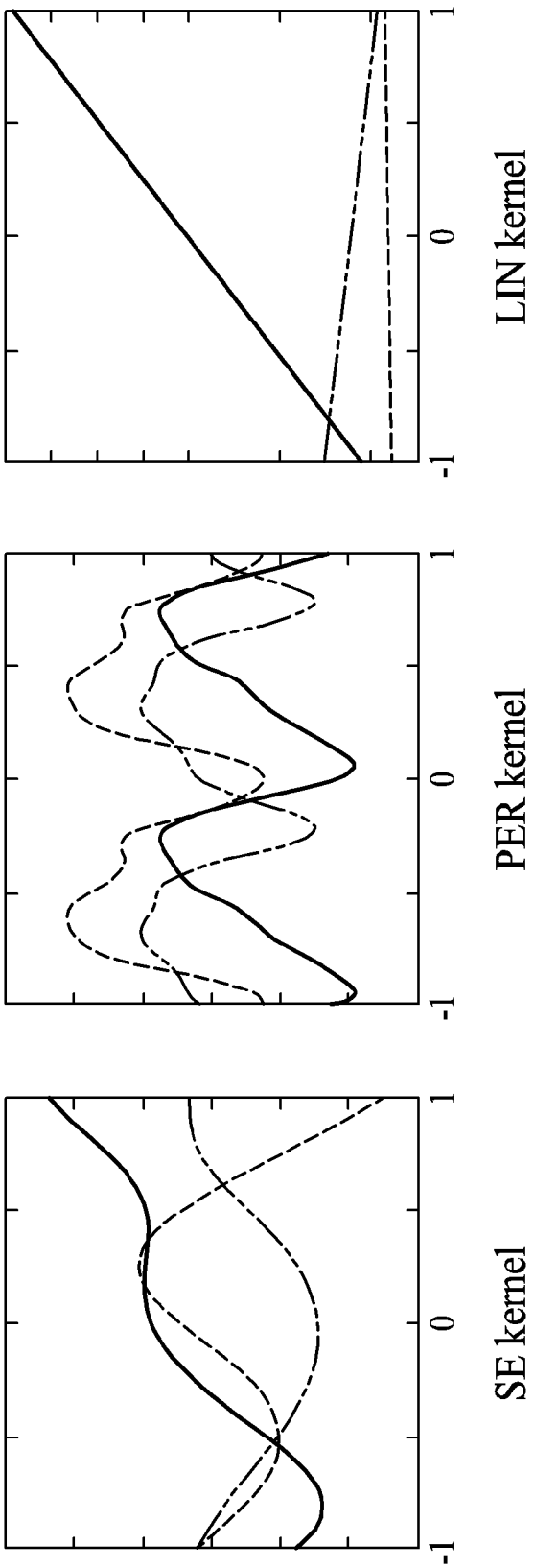
[청구항 10] 하나 이상의 명령어가 저장된 컴퓨터 판독 가능한 저장 매체로서, 상기 하나 이상의 명령어는 컴퓨터에 의해 실행되는 경우, 컴퓨터로 하여금 제1항 내지 제8항 중 어느 한 항의 방법을 실행하게 하는, 컴퓨터 판독 가능한 저장 매체.

WO 2019/013376 PCT/KR2017/007873

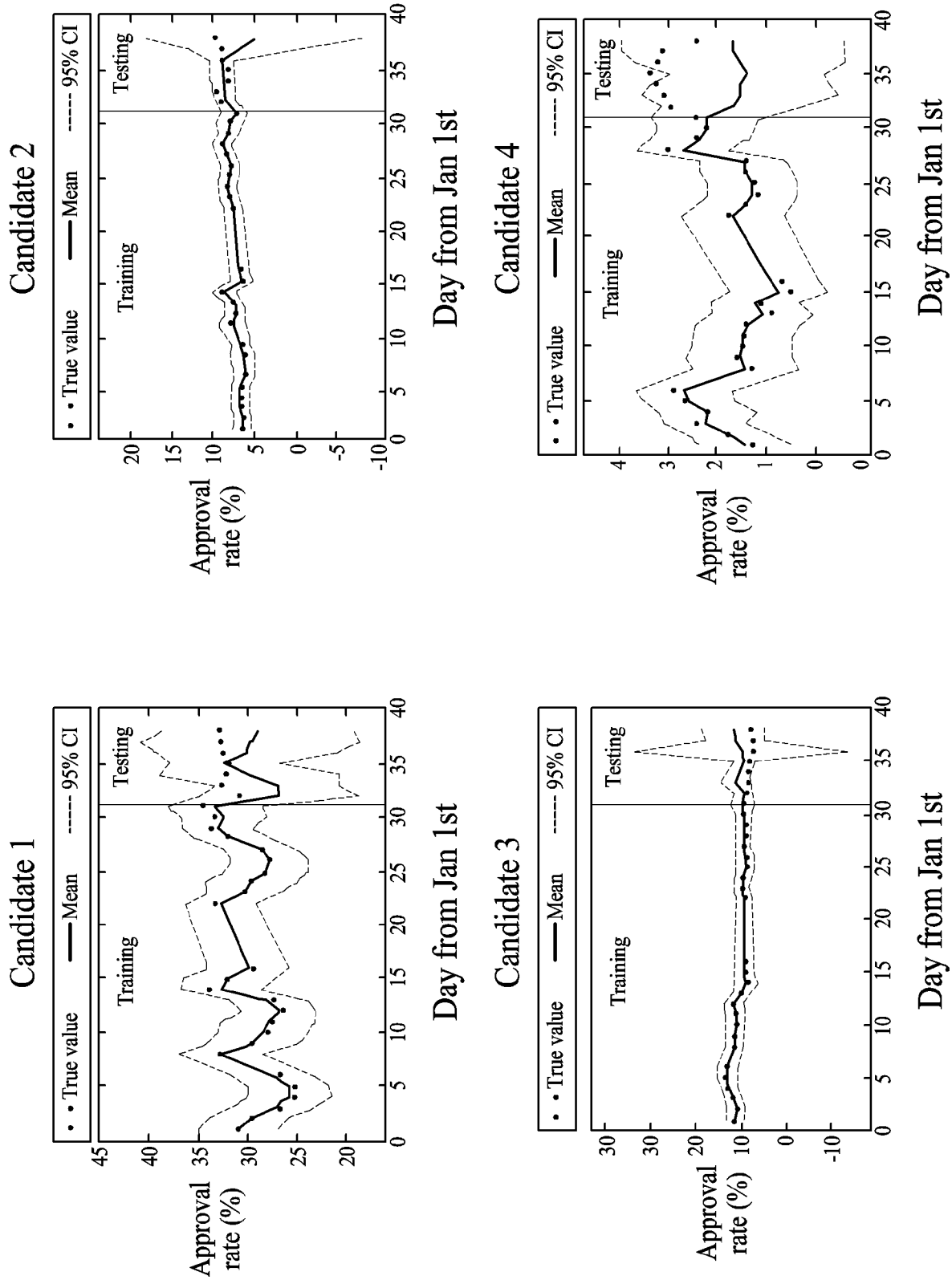
[도1]



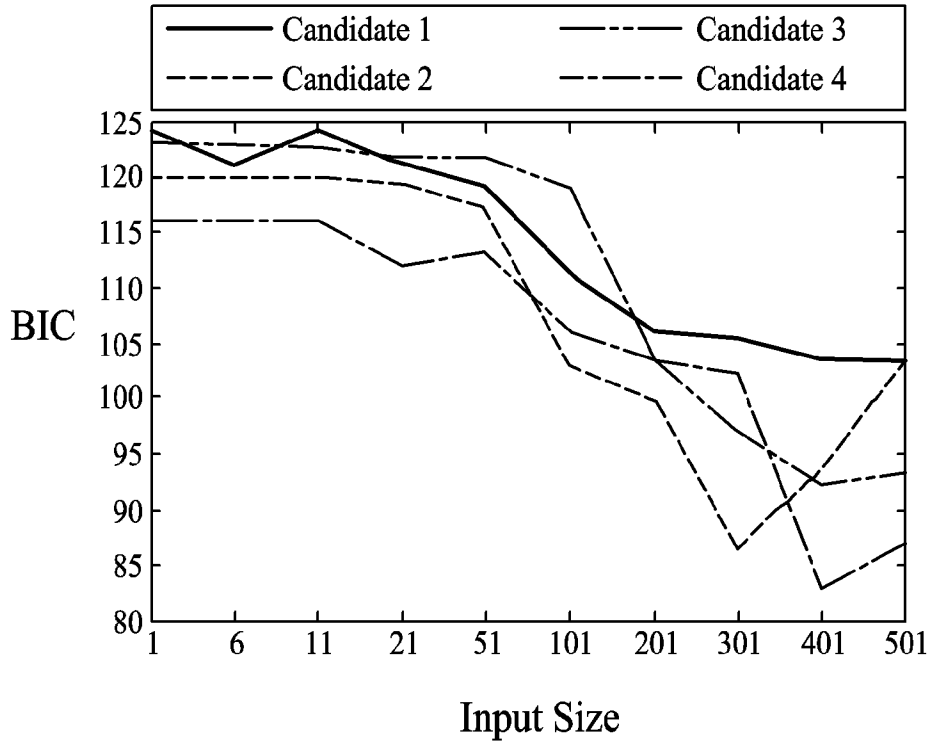
[도2]



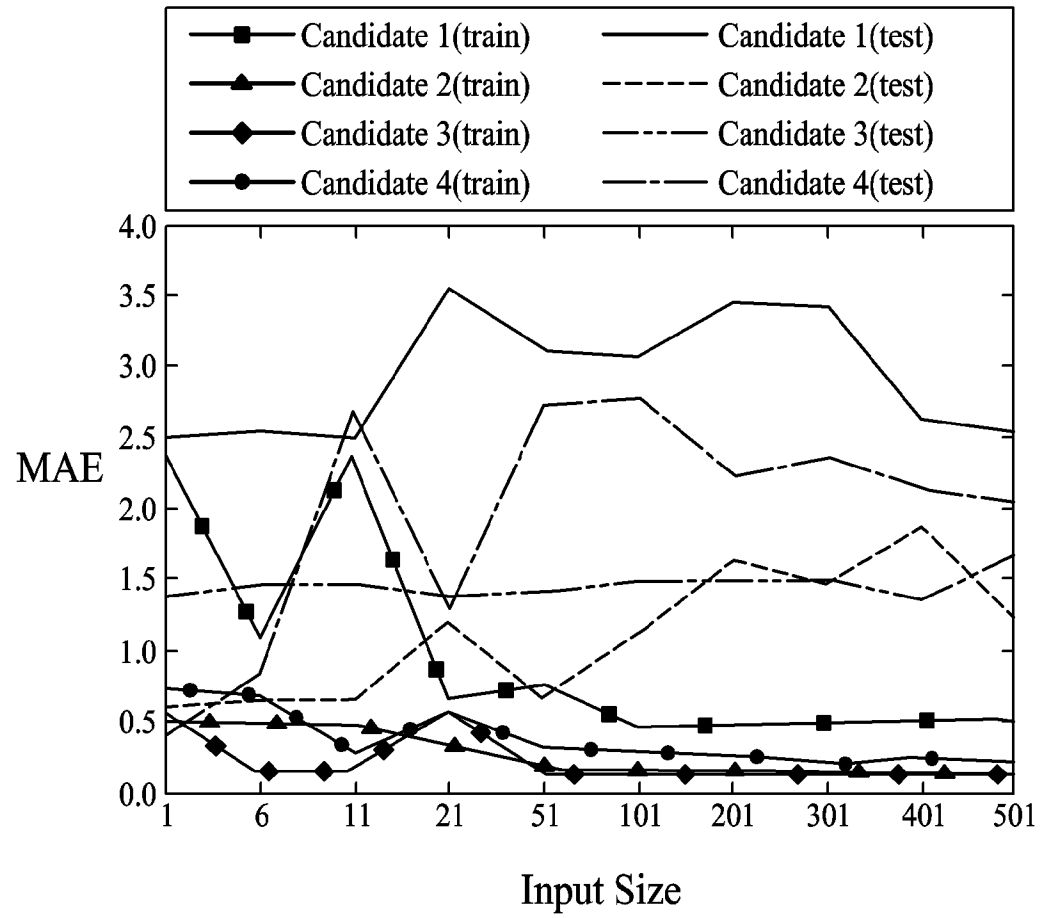
[도3]


Candidate 1
Candidate 2
Candidate 3
Candidate 4
True value
Mean
95% CI
Training
Testing
Approval rate (%)
Day from Jan 1st

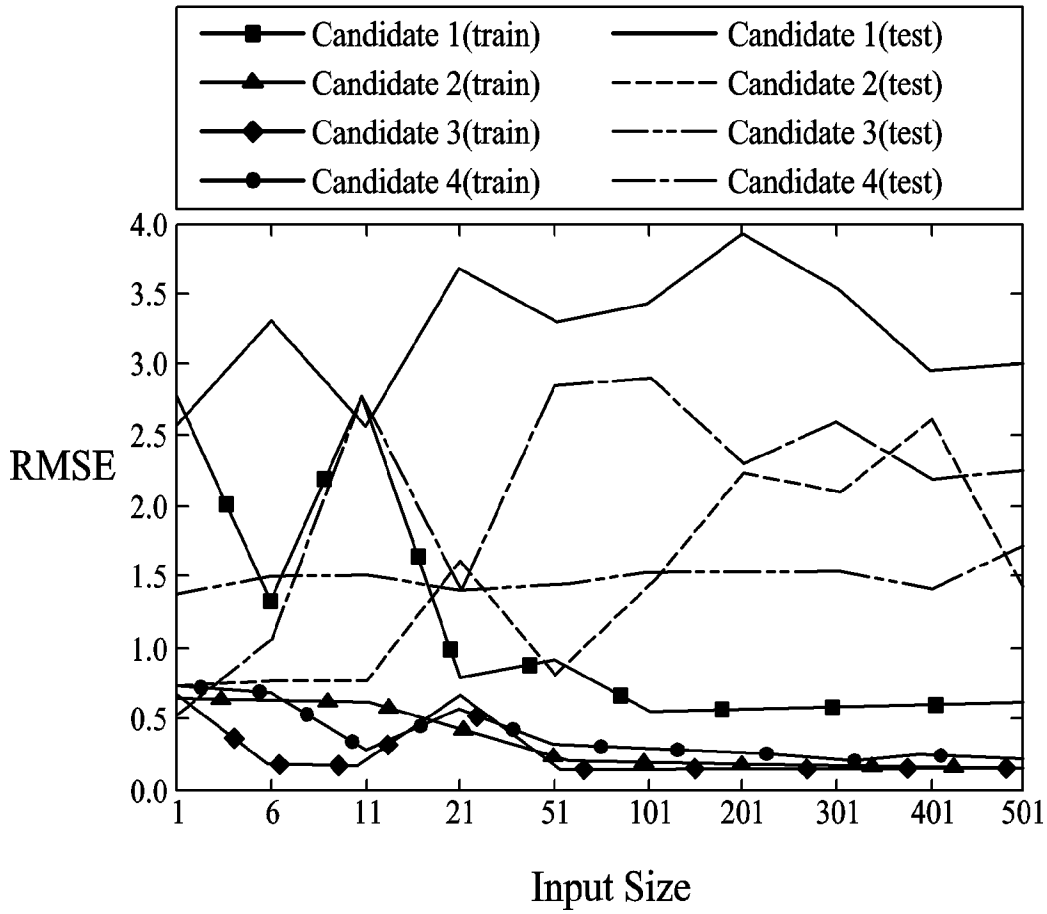
[도4a]


Candidate 1
Candidate 2
Candidate 3
Candidate 4
BIC
125
120
115
110
105
100
95
90
85
80
1
6
11
21
51
101
201
301
401
501
Input Size

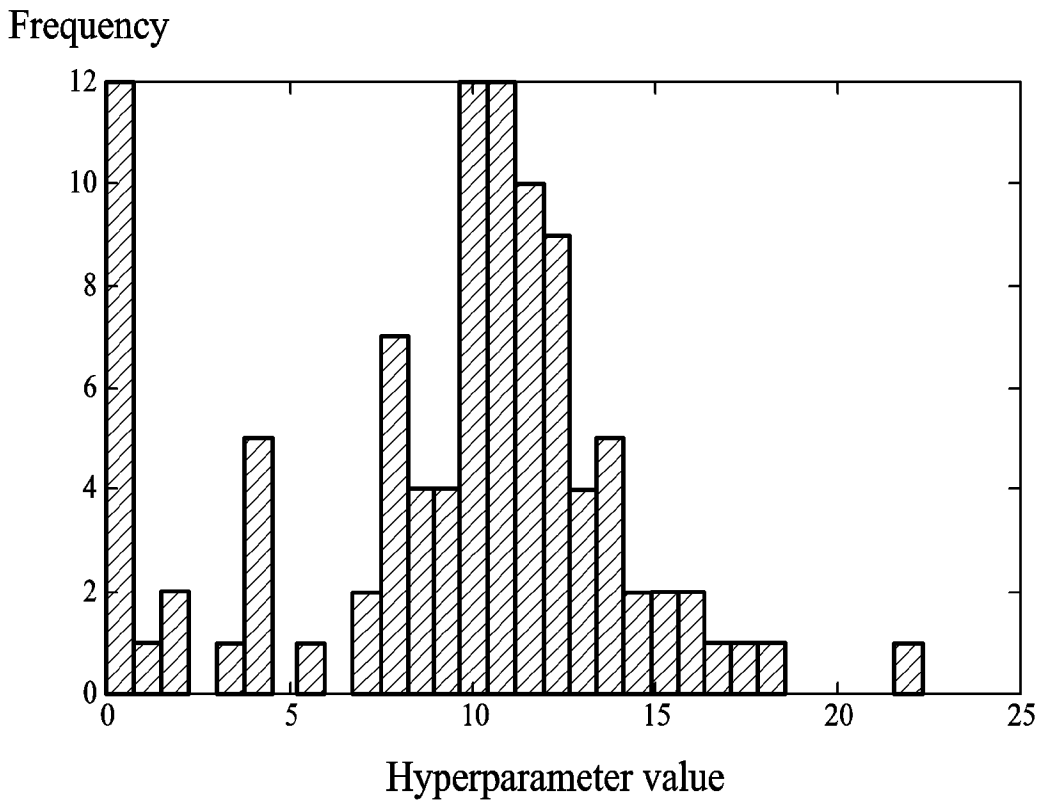
[도4b]


Candidate 1(train)
Candidate 2(train)
Candidate 3(train)
Candidate 4(train)
Candidate 1(test)
Candidate 2(test)
Candidate 3(test)
Candidate 4(test)
MAE
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
6
11
21
51
101
201
301
401
501
Input Size

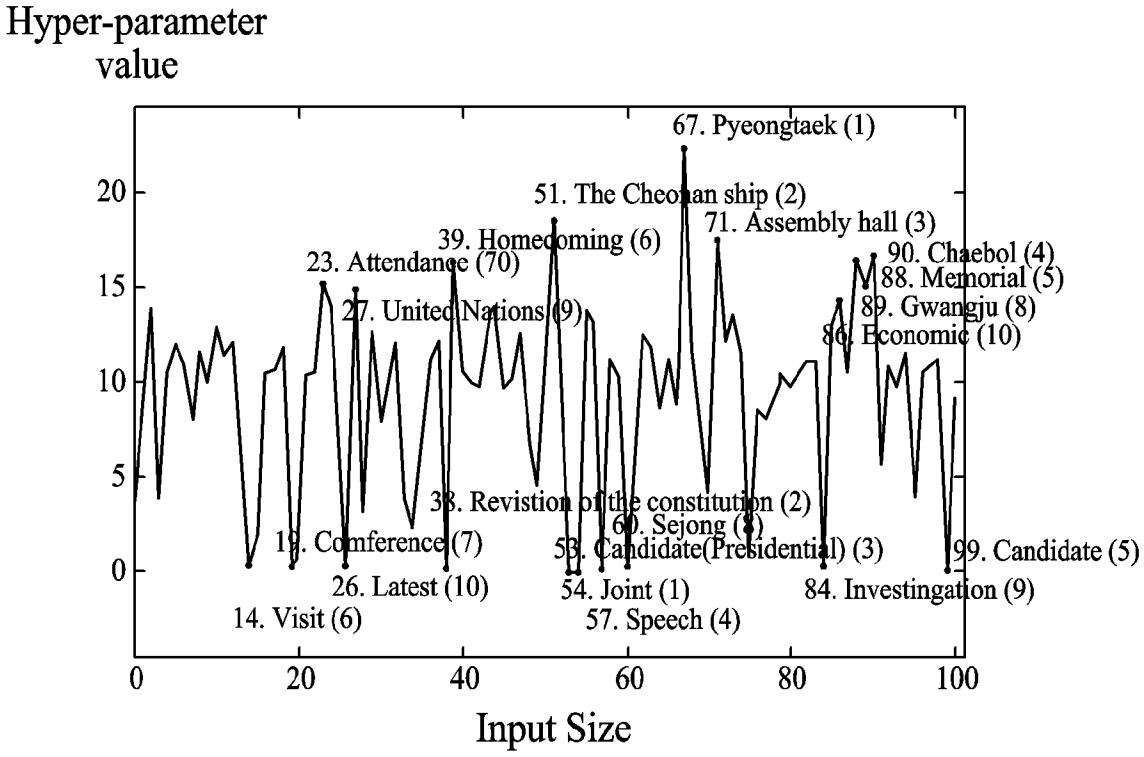
[도4c]


Candidate 1(train)
Candidate 2(train)
Candidate 3(train)
Candidate 4(train)
Candidate 1(test)
Candidate 2(test)
Candidate 3(test)
Candidate 4(test)
RMSE
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
6
11
21
51
101
201
301
401
501
Input Size

[도5a]


Frequency
12
10
8
6
4
2
0
0
5
10
15
20
25
Hyperparameter value

[도5b]


Hyper-parameter value
67. Pyeongtaek (1)
51. The Cheonan ship (2)
71. Assembly hall (3)
39. Homecoming (6)
23. Attendance (70)
90. Chaebol (4)
88. Memorial (5)
27. United Nations (9)
89. Gwangju (8)
86. Economic (10)
38. Revistion of the constitution (2)
60. Sejong (8)
19. Conference (7)
53. Candidate(Presidential) (3)
99. Candidate (5)
26. Latest (10)
54. Joint (1)
84. Investingation (9)
14. Visit (6)
57. Speech (4)
0
5
10
15
20
0
20
40
60
80
100
Input Size

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2017/007873

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/18(2006.01)i, G06F 17/30(2006.01)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)

G06F 17/18; G06F 17/00; G06F 17/27; G06F 17/30

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: time series text, multidimensional, gaussian process, approval rating prediction, word, importance

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NA, Yeong Yeon et al., "Analysis of Approval Ratings of Presidential Candidates Using Multidimensional Gaussian Process and Time Series Text Data", Proceedings of The Korean Operations Research and Management Science Society (KORMS) Conference, pages 1151-1156, April 2017 See abstract; and sections 2.2, 3.1-3.2, 4.3, 5.	1-10
A	HWANG, Yunseong et al., "Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series", International Conference on Machine Learning (ICML), vol. 48, pages 3030-3039, 19 June 2016, (https://pdfs.semanticscholar.org/4a22/143f499fe3506329f4d83fffc94a0cd88fbe.pdf?_ga=2.162429948.1974677847.1511152331-1865043545.1489130336) See sections 2-3.	1-10
A	KR 10-1613397 B1 (KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY) 18 April 2016 See paragraphs [0098]-[0118]; and figures 4-15.	1-10
A	PARK, Dae Min, "Automated Time Series Content Analysis with News Big Data Analytics", Korean Journal of Journalism & Communication Studies, vol. 60, no. 5, pages 353-407, October 2016 See section 3.	1-10
A	DVN news, "UNIST, Developing an Artificial Intelligence 'Relational Automatic Statistician System'", 27 June 2016, (http://www.dvnnews.com/news/articleView.html?idxno=14762) See pages 1-2.	1-10

☐ 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

"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
09 APRIL 2018 (09.04.2018)	**09 APRIL 2018 (09.04.2018)**

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, Republic of Korea Facsimile No. +82-42-481-8578	Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2017/007873

Patent document cited in search report	Publication date	Patent family member	Publication date
None			

국제조사보고서

국제출원번호
PCT/KR2017/007873

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06F 17/18(2006.01)i, G06F 17/30(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)
G06F 17/18; G06F 17/00; G06F 17/27; G06F 17/30

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 텍스트, 다차원, 가우시안 프로세스, 지지율 예측, 단어, 중요도

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	나영언 등, "다차원 가우시안 프로세스와 시계열 텍스트 데이터 이용한 대통령 후보자 지지율 분석", 한국경영과학회 학술대회논문집, 페이지 1151-1156, 2017.04 요약; 및 섹션 2.2, 3.1-3.2, 4.3, 5 참조.	1-10
A	YUNSEONG HWANG 등, "Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series", International Conference on Machine Learning (ICML), 제48권, 페이지 3030-3039, 2016.06.19 (https://pdfs.semanticscholar.org/4a22/143f499fe3506329f4d83fffc94a0cd88fbe.pdf?_ga=2.162429948.1974677847.1511152331-1865043545.1489130336) 섹션 2-3 참조.	1-10
A	KR 10-1613397 B1 (한국과학기술원) 2016.04.18 단락 [0098]-[0118]; 및 도면 4-15 참조.	1-10
A	박대민, "장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성", 한국언론학보, 제60권, 제5호, 페이지 353-407, 2016.10 섹션 3 참조.	1-10
A	DVN 뉴스, "UNIST,인공지능`관계형자동통계학자시스템`개발", 2016.06.27 (http://www.dvnnews.com/news/articleView.html?idxno=14762) 페이지 1-2 참조.	1-10

☐ 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

"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
2018년 04월 09일 (09.04.2018)	2018년 04월 09일 (09.04.2018)

ISA/KR의 명칭 및 우편주소
대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)
팩스 번호 +82-42-481-8578

심사관
강희곡
전화번호 +82-42-481-8264

국제조사보고서
대응특허에 관한 정보

국제출원번호
PCT/KR2017/007873

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일

없음