https://doi.org/10.5392/JKCA.2020.20.06.001

베이지안 공액 사전분포를 이용한 키워드 데이터 분석

Keyword Data Analysis Using Bayesian Conjugate Prior Distribution

전성해
청주대학교 빅데이터통계학과

Sunghae Jun(shjun@cju.ac.kr)

요약

빅데이터 분석에서 텍스트 데이터의 활용이 증가하고 있다. 따라서 텍스트 데이터의 분석 기법에 관한 많은 연구가 이루어지고 있다. 본 논문에서는 텍스트 데이터로부터 추출된 키워드 데이터의 분석을 위하여 공액 사전분포 기반의 베이지안 학습 방법이 연구된다. 베이지안 통계학은 기존의 데이터에 새로운 데이터가 추가될 때마다 모수를 갱신하는 데이터 학습을 제공하기 때문에 시간에 따라 대용량의 데이터가 생성 및 추가되는 빅데이터 환경에서 효율적인 방법을 제공한다. 제안 방법의 성능과 적용 가능성을 보이기 위하여 실제 특허 빅데이터를 전처리하여 구축된 정형화된 키워드 데이터를 분석하는 사례연구를 수행한다.

■ 중심어 : | 빅데이터 | 베이지안통계 | 베타분포 | 공액사전분포 | 키워드데이터 |

Abstract

The use of text data in big data analytics has been increased. So, much research on methods for text data analysis has been performed. In this paper, we study Bayesian learning based on conjugate prior for analyzing keyword data extracted from text big data. Bayesian statistics provides learning process for updating parameters when new data is added to existing data. This is an efficient process in big data environment, because a large amount of data is created and added over time in big data platform. In order to show the performance and applicability of proposed method, we carry out a case study by analyzing the keyword data from real patent document data.

■ keyword : | Big Data | Bayesian Statistics | Beta Distribution | Conjugate Prior Distribution | Keyword Data |

I. 서 론

키워드 데이터 분석(keyword data analysis)은 문서 기반의 빅데이터로부터 전처리 과정을 통하여 구축된 정형화된 텍스트 데이터(structured text data)를 분석하여 의미 있는 패턴을 추출하는 과정이다[1-5]. 다양한 분야에서 키워드 데이터 분석을 위하여 통계학과 머신러닝에서 제공하는 데이터 분석기법을 사용한다. Kim et al. (2019)은 통계학의 시계열 분석(time series analysis)과 코플라(copula) 모형을 이용하여 특허 빅데이터에서 추출된 기술 키워드 데이터를 분석하였다[2]. Uhm et al. (2017)은 지속 가능한 기술(sustainable technology) 예측을 위하여 특허문서에서 추출된 각 키워드의 빈도수에 대하여 구간 추정(interval estimation)을 수행하였다[3]. 지금까지 다양한 분야에서 키워드 데이터 분석에 대한 활발한 연구가 이루어지고 있다[4-6].

키워드 데이터 분석을 위한 기존의 연구들은 대부분

접수일자 : 2020년 03월 04일
수정일자 : 2020년 05월 06일
심사완료일 : 2020년 05월 25일
교신저자 : 전성해, e-mail : shjun@cju.ac.kr

수집된 데이터 전체를 한 번의 모형 구축을 위하여 모두 사용하였다. 이와 같은 키워드 분석은 새로운 데이터가 추가될 때 기존의 데이터와 새로운 데이터를 합쳐서 다시 분석해야 하는 어려움이 있다. 특히 빅데이터 환경에서는 매일 막대한 양의 새로운 데이터가 생성되고 저장된다. 따라서 이전의 데이터를 새로운 데이터와 통합하여 분석하는 것은 어려울 뿐만 아니라 향후 불가능할 수 있다. 왜냐하면 데이터 분석을 지원하는 소프트웨어와 하드웨어는 분명 데이터의 크기에 영향을 받기 때문이다. 따라서 기존 데이터에 의한 분석모형과 새로운 데이터에 의한 분석모형을 합치는 데이터 분석 전략이 필요하게 된다. 하지만 지금까지 연구되어진 키워드 데이터 분석은 새로운 텍스트 데이터가 추가될 때마다 기존의 데이터와 새로운 데이터를 모두 합친 후 전체 데이터를 다시 분석하여 갱신된 모형을 구축하였다. 이와 같은 문제점을 해결하기 위하여 본 논문에서는 기존의 데이터로부터 구축된 모형은 그대로 유지하고 새로운 데이터만을 반영한 모형을 만들어 기존의 모형과 합쳐서 갱신된 모형을 구축하는 키워드 데이터 분석전략을 제안한다. 본 연구에서는 제안하는 공액(conjugate) 분포에 기반한 베이지안 키워드 분석 방법은 이전 데이터에 대한 모형의 정보는 사전분포(prior distribution)로 새롭게 관측된 데이터에 대한 정보는 우도함수(likelihood function)로 나타낸다. 사전분포와 우도함수를 곱하여 최종 갱신된 모형인 사후분포(posterior distribution)를 구한다. 최종적으로 사후분포를 이용하여 모수의 기댓값(expectation)을 계산하고 이 결과를 이용하여 전체 데이터에 대한 정보를 파악한다.

따라서 본 연구의 목적은 텍스트 빅데이터 환경에서 새로운 데이터가 지속적으로 추가될 때 기존의 데이터에 대한 정보와 새로운 데이터에 대한 정보를 분리하여 모형화함으로써 효율적인 키워드 분석을 수행하기 위한 방법을 개발하는 것이다. 2장에서는 키워드 데이터 분석과 관련된 기존 연구 및 제안 연구의 차별성을 알아본다. 제안하는 공액 사전분포를 이용한 베이지안 분석 방법은 3장에서 설명하고 4장에서는 제안 방법의 실제 적용을 보이기 위한 사례분석이 이루어진다. 본 논문의 결과와 향후 연구과제는 마지막 장에서 다룬다.

II. 관련 연구

빅데이터 환경에서 텍스트 데이터가 차지하는 비중은 매우 크다[1]. 페이스북과 같은 사회 네트워크 서비스 (social network service, SNS)에서 다루어지는 대부분의 데이터 형태는 텍스트이다. 구글 검색을 통하여 얻게 되는 웹 문서도 마찬가지이다. 따라서 빅데이터의 효율적인 분석을 위하여 텍스트 데이터의 처리와 분석을 위한 연구는 반드시 필요하게 된다. 특히 텍스트 기반의 빅데이터로부터 추출된 키워드 데이터의 분석에 대한 연구는 다양한 분야에서 이루어지고 있다. Park and Jun (2020)은 특허문서로 이루어진 빅데이터로부터 특허 키워드를 추출하여 베이지안 네트워크와 인자분석(factor analysis)을 이용하여 재난 인공지능(disaster artificial intelligence) 관련 세부 기술들 간의 기술 연관성을 파악하였다[4]. Kim et al. (2017)은 기술 키워드 분석을 위하여 벌점 회귀분석 모형(penalized regression models)을 이용한 통계적 기술예측 모형을 제안하였다[5]. Huh (2018)는 개인 건강 활동과 관련된 의료 빅데이터의 분석을 위하여 자동으로 키워드를 추출하고 분석할 수 있는 머신러닝 절차에 대한 연구를 수행하였다[6].

기존의 키워드 데이터 분석을 위한 연구는 새로운 데이터 분석 기법에 대한 연구들이 대부분이었다. 공통적으로 키워드 데이터 분석모형을 구축하는 시점에 그때까지 수집된 빅데이터를 한꺼번에 이용하였다. 이와 같은 분석 전략은 새로운 데이터가 추가될 때마다 매번 전체 데이터를 분석해야 하기 때문에 매일 막대한 양의 새로운 데이터가 생성되는 빅데이터 환경에서는 효율적이지 않게 된다. 그러므로 기존의 빅데이터에 기반한 분석모형을 유지한 상태에서 새롭게 추가된 데이터만을 분석하는 모형과 결합하여 갱신된 모형을 구축할 수 있는 키워드 데이터 분석전략이 필요하게 된다. 3장에서 설명할 본 논문의 제안 방법은 기존의 데이터로부터 구축된 모형을 기반으로 새로운 데이터에 의한 분석모형을 결합하여 최종 모형을 구축하게 된다. 이를 위하여 제안 방법에는 베이지안 공액 사전분포와 베이지안 학습이 사용된다.

III. 공액 사전분포를 이용한 베이지안 키워드 분석

본 논문에서는 텍스트 빅데이터로부터 추출된 키워드 데이터의 효율적인 분석을 위하여 공액 사전분포 기반의 베이지안 학습을 제안한다. 제안 방법은 새로운 데이터가 추가되기 전까지의 데이터를 분석한 모형의 결과를 유지한 상태에서 새로운 데이터에 기반한 모형 결과와 결합하여 갱신된 최종 모형을 구축한다. 새로운 데이터가 추가될 때마다 기존의 데이터와 통합하여 전체 데이터를 다시 분석해야 하는 기존의 키워드 데이터 분석에 비하여 효율적인 빅데이터 분석이 가능하게 된다. 제안 방법은 베이지안 학습과 공액 사전분포와 우도함수를 이용한 사후분포의 생성 및 이를 이용한 키워드 데이텅의 분석 과정을 갖는다.

1. 베이지안 추론 및 학습

주어진 데이터 D와 가설(hypothesis) H에 대하여 베이즈 정리(Bayes' theorem)는 다음과 같이 정의된다[7].

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \quad (1)$$

데이터에 주어졌을 때 특정 가설일 확률을 조건부 확률을 이용하여 계산한다. 식 (1)에서 모수 θ와 데이터 변수 x를 고려할 때 θ에 대한 사전분포를 $f(\theta)$로 나타내고 θ가 주어진 상태에서 x가 관측될 우도함수를 $f(x|\theta)$라 하면 x가 관측되었다는 조건 하에서 θ의 사후분포는 베이즈 정리에 의해 다음과 같이 정의된다[8].

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} \quad (2)$$

식 (2)에서 $f(x)$는 데이터 x가 나타날 모든 가능도의 합을 나타낸다. $f(x)$에는 θ가 포함되어 있지 않기 때문에 식 (2)는 다음과 같이 나타낼 수 있다[7][8].

$$f(\theta|x) \propto f(\theta)f(x|\theta) \quad (3)$$

$f(\theta)$는 데이터의 특성을 나타내는 모수 θ에 대한 사전 믿음(belief)이고 $f(x|\theta)$는 모수 θ에 대한 현재 관측된 데이터 x의 가능도를 나타낸다. $f(\theta|x)$는 관측된 데이터 x에 의해서 초기 믿음 $f(\theta)$의 갱신된 사후 믿음을 나타낸다.

본 논문에서 주어진 데이터 x는 특정 키워드의 발생여부를 나타내고 θ는 특정 키워드의 발생확률을 나타낸다. 새로운 데이터가 추가되면 $f(\theta|x)$는 새로운 우도함수에 대한 사전분포가 되고 갱신된 사후분포를 계산한다. 매번 새로운 데이터가 관측될 때마다 이전 데이터는 분석에 사용되지 않고 사전분포로 대표된다. 이와 같은 반복된 사후분포의 갱신이 베이지안 학습(Bayesian learning)이다. 본 연구에서는 공액 사전분포에 기반한 베이지안 학습을 이용하여 키워드 데이터를 분석한다.

2. 베이지안 공액 사전분포를 이용한 키워드 분석

키워드 데이터 분석을 위하여 빅데이터 환경에서 텍스트 문서를 수집하고 전처리하여 [그림 1]과 같이 정형화된 데이터인 문서-키워드 행렬(document-keyword matrix)을 구축한다.

Document	Year	Keyword_1	⋯	Keyword_p
1	1983	Binary data $x_{ij} = \begin{cases} 1 & \text{occurred} \\ 0 & \text{not occurred} \end{cases}$		
2	⋮			
⋮				
n	⋮			
n+1				
⋮				
n+m	2019			

그림 1. 정형화된 문서-키워드 행렬

정형화된 행렬의 각 행은 문서이고 열은 연도(year)와 문서로부터 추출된 키워드이다. 연도를 나타내는 열을 제외하고 행렬의 각 셀은 특정 키워드가 각 문서에 발생했는지(occurred) 여부를 나타낸다. i번째 문서에서 j번째 키워드가 1번 이상 발생하면 x_{ij}=1이고 그렇지 않으면 x_{ij}=0이다. 따라서 본 연구에서는 각 키워드의 발생확률 정보를 가지고 있는 모수 θ를 추정하기 위

하여 베이지안 학습을 사용한다. 키워드의 발생 여부를 나타내는 이진 데이터(binary data)에 적합한 확률분포는 다음의 이항분포(binomial distribution)이다[9].

$$f(x)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x},\ x=0,1,\ldots,n \quad (4)$$

n과 θ는 이항분포의 모수(parameters)로 각각 시행 횟수와 성공확률을 나타낸다. 따라서 이항분포는 주어진 데이터의 우도함수가 된다. 베이지안이 아닌 빈도론자(frequentist)의 통계적 추론에서 θ는 고정된 모수이지만 베이지안 학습에서는 확률분포를 갖는 확률변수(random variable)가 된다. 이때 확률변수 θ가 따르게 되는 분포가 사전분포이다. 사전분포는 우도함수와 곱해지면서 사후분포를 계산하기 때문에 우도함수와의 연산에 편리한 분포인 공액 사전분포로 선택된다[10]. 본 연구에서는 θ는 사전분포로서 다음과 같은 베타분포(beta distribution)를 사용한다[9].

$$f(\theta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},0<\theta<1 \quad (5)$$

α와 β는 베타분포의 모수이고 $B(\alpha,\beta)$는 다음과 같이 정의되는 베타함수(beta function)이다.

$$B(\alpha,\beta)=\int_{0}^{1}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \quad (6)$$

사후분포 계산의 편리성을 위하여 이항분포에 대한 공액사전분포는 베타분포가 된다. 따라서 모수 θ의 사후분포는 다음과 같이 계산된다.

$$f(\theta|x)=\frac{1}{B(\alpha+x,\beta+n-x)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \quad (7)$$

우도함수가 이항분포이고 사전분포가 공액분포인 베타분포이기 때문에 계산된 사후분포도 베타분포가 된다. 사후분포인 식 (7)의 베타분포로부터 모수 θ의 기댓값은 다음과 같이 구할 수 있다[7].

$$E(\theta)=\frac{\alpha+x}{\alpha+\beta+n} \quad (8)$$

$E(\theta)$ 값을 이용하여 키워드의 발생 가능성을 예측할 수 있다. 본 논문에서 제안하는 베이지안 공액 사전분포를 이용한 키워드 데이터 분석은 [그림 2]와 같은 절차에 의해 수행된다.

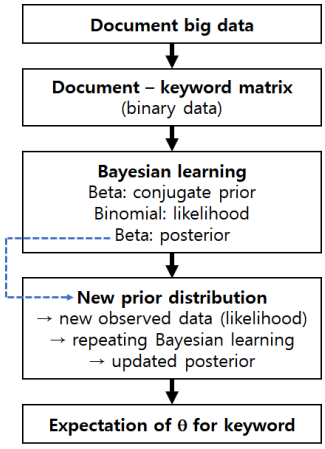

그림 2. 키워드 데이터 분석 절차

주어진 도메인에 대한 문서 빅데이터를 수집하게 되면 먼저 텍스트 마이닝(text mining)의 전처리 과정을 통하여 이진 데이터 형태로 이루어진 문서-키워드 행렬을 구축한다. 본 논문에서는 데이터의 전처리와 이후의 베이지안 학습을 위한 계산을 위하여 R 데이터 언어와 R이 제공하는 패키지를 사용한다[11-13]. 문서-키워드 행렬에서 분석에 사용될 관심 키워드를 선택하여 각 키워드에 대하여 시간에 따라 단계별로 데이터를 분할한다. 분할된 단계별 데이터를 이용하여 키워드의 발생여부를 설명하는 모수(θ)에 대한 베이지안 학습을 적용하여 매 단계별 갱신된 사후분포를 구한다. 모든 단계의 학습을 마친 후 최종 사후분포로부터 θ의 기댓값을 구하고 이 결과를 이용하여 각 키워드의 향후 발생 가능성을 예측하게 된다. 본 연구는 공액 사전분포 기반의 베이지안 학습을 이용한 키워드 데이터 분석을 제안한다. 현재 빅데이터의 상당 부분은 텍스트 데이터로 이루어지기 때문에 제안 방법은 빅데이터 분석을 위한 다양한 분야에서 사용이 가능할 것이다. 본 연구에서 제

안하는 키워드 데이터 분석 방법을 사용하게 되면 이전의 데이터에 대한 정보는 별도의 분석 절차 없이 추가된 새로운 데이터의 분석 결과에 반영되어 즉각적인 빅데이터 분석 결과를 얻을 수 있게 된다. 제안 방법의 실제 적용을 보이기 위하여 다음 장에서는 특허 빅데이터를 이용한 사례 분석 결과를 제시한다.

IV. 사례 분석

제안 방법의 실제 적용을 보이기 위하여 본 논문에서는 AI 기술 관련 특허문서를 수집하여 분석하였다[14][15]. 컴퓨터가 지능을 갖도록 하는 방법을 연구하는 AI 기술에 대한 특허 출원이 최근 활발히 이루어지고 있다[16]. 실험에 사용된 유효한 AI 기술특허는 1983년부터 2019년까지 출원, 등록된 16,874건 이었다. 특허문서에 대한 전처리 과정을 거쳐서 최종적으로 구축된 정형화된 많은 키워드를 포함하고 있다. 본 연구에서는 이들 중에서 AI에 많은 영향을 미치는 'data'와 'learning' 키워드를 사용하였다. 각 특허문서에 키워드 data와 learning의 출현 여부를 확인하여 해당 특허에 이 키워드들이 출현했으면 1의 값을, 출현하지 않았으면 0의 값을 할당하였다. 1983년부터 2019년까지 16,875 건의 특허를 시간에 따라 3단계로 나누어 각 단계별 키워드를 포함한 특허수를 구하면 [표 1]과 같다.

표 1. 키워드와 출현 빈도수

단계	전체 특허 수	키워드를 포함한 특허수	
		data	learning
1	5,625	4,386	432
2	5,625	4,653	304
3	5,625	4,804	486
합계	16,875	13,843	1,222

[표 1]의 결과를 이용하여 먼저 data와 learning 키워드에 대한 사전분포와 단계1의 사후분포를 계산하였다. θ에 대한 사전분포는 모수 α와 β가 각각 1인 베타분포를 사용하였다. θ에 대한 사전정보를 0부터 1사이의 값으로 균등하게 하였다. 키워드 data에 대한 사전 및 사후분포는 [그림 3]과 같다.

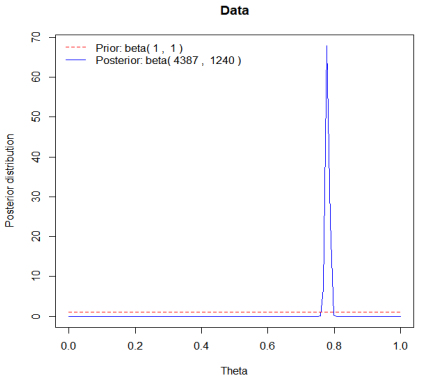

그림 3. 키워드 data의 사전 및 사후분포

[그림 3]에서 θ의 사전분포는 베타분포이고 데이터의 정보를 나타내는 우도함수는 이항분포이다. 데이터의 크기 n은 5,625이고 성공횟수 x는 4,386이다. 이를 통하여 단계 1의 최종 사후분포는 모수 α와 β가 각각 4,387과 1,240인 베타분포가 되었다. 사전분포를 0과 1사이의 어느 값이나 발생확률이 동일한 beta(1, 1)으로 지정하였기 때문에 사후분포는 x를 n으로 나눈 값인 0.7797에서 가장 큰 값을 갖는다. [그림 4]는 키워드 learning의 사전 및 사후분포를 나타낸다.

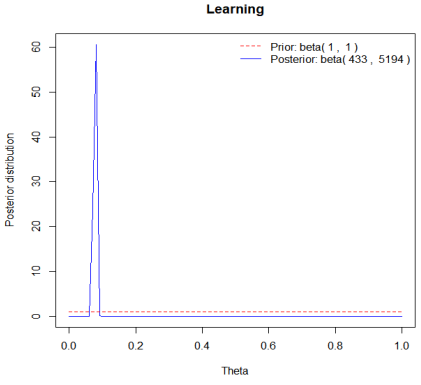

그림 4. 키워드 learning의 사전 및 사후분포

키워드 data와 마찬가지로 beta(1,1)의 사전분포를 갖고 단계 1에서 learning의 발생빈도는 432건 이었다. learning의 우도함수는 n=5,625이고 성공횟수 x=432인 이항분포이다. 이를 통하여 learning의 사후분포는 모수 α와 β가 각각 433과 5,194인 베타분포를 따른다. 이와 같은 단계 1의 결과를 이용하여 다음

단계 학습을 위한 θ의 사전분포로 [그림 5]의 베타분포를 이용하였다.

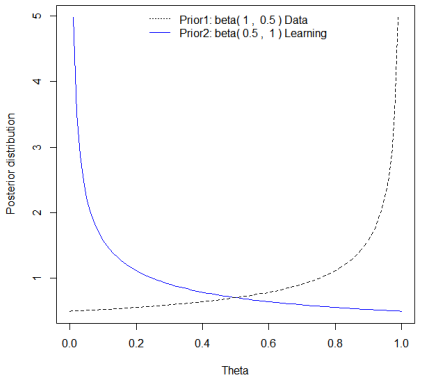

그림 5. 키워드 data와 learning의 사전분포

단계 1에서는 θ에 대한 사전정보가 없었기 때문에 사전분포로 beta(1,1)의 균등분포(uniform distribution)를 사용하였지만, 단계 2에서는 단계 1의 사후분포를 고려하여 각 키워드에 대한 θ의 사후분포를 지정하였다. [그림 5]에서 키워드 data의 θ에 대한 사전분포 beta(1.0, 0.5)를 지정하였다. θ의 값이 클수록 큰 pdf 값을 갖게 된다. 또한, learning에 대한 사전분포는 beta(0.5, 1.0)으로 지정하였고 이는 data와는 반대로 θ의 값이 클수록 작은 pdf 값을 갖는다. 각 키워드에 대한 베타 사전분포를 결정하고 단계 2와 3에 대한 사후분포를 구하면 다음과 같다. [그림 6]는 키워드 data에 대한 θ의 단계별 사후분포를 보여준다.

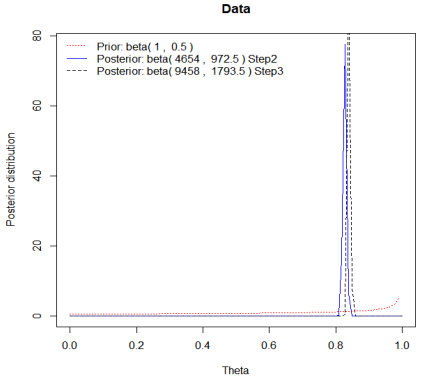

그림 6. 키워드 data의 단계별 사후분포

단계별 사후분포 결과를 통하여 data의 θ값이 증가하는 것을 알 수 있다. 시간이 지남에 따라 data의 발생 빈도는 증가하고 있음을 알 수 있다. [그림 7]는 키워드 learning의 θ값에 대한 단계별 사후분포를 보여준다.

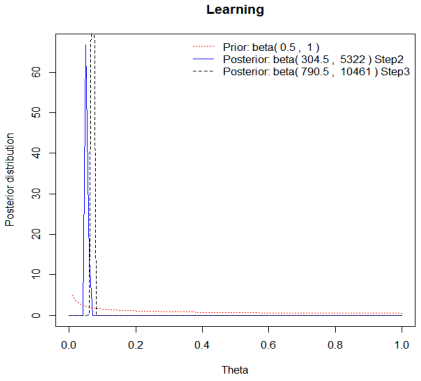

그림 7. 키워드 learning의 단계별 사후분포

키워드 data와 마찬가지로 learning의 θ값도 단계가 지날수록 증가함을 알 수 있다. [그림 6]와 [그림 7]의 비교를 통하여 θ의 증가속도는 data에 비하여 learning이 더 빠름을 알 수 있다. 3단계의 베이지안 학습을 통하여 최종적으로 갱신된 베타 사후분포의 모수와 θ의 기댓값은 [표 2]와 같다.

표 2. Beta 사후분포와 θ의 기댓값

키워드	모수 (α,β)	E(θ)
data	(9458, 1793.5)	0.8406
learning	(790.5, 10461)	0.0703

키워드 data와 learning의 발생확률에 대한 기댓값은 각각 0.8406과 0.0703임을 알 수 있다. 본 실험에서 최종적으로 구한 사후분포는 새로운 데이터가 추가되면 다시 새로운 베이지안 학습을 위한 사전분포로 사용된다. 이 사전분포는 새로운 데이터의 우도함수와 곱해져서 갱신된 사후분포를 계산한다. 갱신된 사후분포의 모수를 이용하여 θ의 기댓값도 바뀌게 된다. 이와 같이 새로운 데이터가 추가될 때마다 베이지안 학습이 이루어지고 학습 결과로부터 θ의 갱신된 정보를 얻게 된다. [표 3]은 기존의 키워드 데이터 분석 방법과 제안

방법을 비교한 결과를 나타낸다.

표 3. 기존방법과 제안방법의 비교

시점	추가된 데이터	데이터 분석모형	
		기존방법	제안방법
Past	5,625	5,625 분석	5,625 분석
Current	5,625	11,250 분석	5,625 분석 + Past 분석결과
New	5,625	16,875 분석	5,625 분석 + Current 분석결과
Total	16,875	매번 해당시간의 전체 데이터 분석	시간에 따라 추가된 데이터만 분석

[표 1]의 단계 1을 [표 3]에서는 'Past' 시점으로 하고, 단계 2는 'Current' 시점으로 표시하였다. 단계 3은 새롭게 추가된 'New' 시점으로 지정하였다. 최초의 'Past' 시점에서는 기존방법이나 제안방법이 모두 5,625 개의 관측된 데이터만을 분석하였다. 하지만 시점이 'Current'와 'New'로 지나갈수록 기존방법은 분석해야 할 데이터의 크기가 커지게 된다. 이에 비하여 제안방법은 해당시점에 새롭게 추가된 데이터만을 분석하고 나머지 이전 데이터에 대한 정보는 이전 분석결과를 그대로 사용하게 된다. 본 사례 분석에서는 데이터의 관측 시점을 3개 단계로 구분하였지만, 실제 빅데이터 환경에서는 분석 시점이 무한히 커질수 있기 때문에 매번 이전 데이터를 포함한 전체 데이터를 분석하는 기존방법은 한계가 있게 된다. 이와 같은 한계를 극복하기 위하여 본 논문은 공액 사전분포 기반의 베이지안 학습을 이용한 키워드 데이터 분석을 제안하였다.

V. 결 론

본 연구에서는 베이지안 학습을 이용한 키워드 데이터 분석을 제안하였다. 특허문서에 포함된 특정 키워드의 시간에 따른 단계별 발생 결과를 분석하기 위하여 베이지안 공액 사전분포를 이용하였다. 특허 데이터베이스로부터 검색된 AI 기술 관련 특허문서를 전처리하여 AI 기술 키워드를 추출하였다. 추출된 전체 키워드 중에서 data와 learning을 선정하였다. 왜냐하면 데이터 학습(learning from data)은 AI의 핵심 기술 중 하나이기 때문이다. 키워드 빈도 데이터를 발생 여부의 이진 데이터로 바꾸고 키워드 발생확률 θ를 추정하였다. 베타 공액 사전분포와 이항 우도함수를 이용하여 사후분포를 구하고 사후분포의 모수를 이용하여 θ의 기댓값을 계산하였다.

특히 본 논문에서는 1983년부터 2019년까지의 출원, 등록된 AI 기술특허 문서를 시간에 따라 3개의 구간으로 나누어 각 단계별로 사후분포를 구하였다. 이전 단계의 사후분포가 현 단계에서 사전분포로 사용되는 베이지안 학습을 적용하였다. 1단계를 기존의 데이터로 결정하고 2단계, 3단계의 데이터를 새롭게 추가되는 데이터로 사용하였다. 1단계의 데이터 분석 결과는 그대로 유지된 상태에서 2단계 데이터를 반영한 결과를 통하여 모형이 갱신되고 다시 2단계까지의 갱신된 모형이 유지된 상태에 3단계의 새로운 데이터를 반영하여 최종 분석모형을 구축하였다. 이 결과를 기반으로 θ의 기댓값을 이용하여 각 키워드의 발생확률을 예측할 수 있었다. 따라서 특허문서에서 세부 기술을 나타내는 각 키워드의 발생확률과 시간에 따른 발생확률의 변동을 추정하게 되면 이 결과를 이용하여 특정 기술의 예측이 가능하게 되었다.

제안 방법은 공액 사전분포를 이용한 베이지안 학습을 이용하였지만 통계학과 머신러닝에서 제공하는 더 다양한 분석 기법들을 적용한 키워드 데이터 분석에 대한 추가적인 연구가 요구된다. 먼저 베이지안 통계학을 이용한 키워드 데이터 분석을 위한 향후 연구에서는 특정 공액 사전분포뿐만 아니라 마코프체인 몬테칼로(Markov Chain Monte Carlo, MCMC)기법을 이용한 좀 더 일반화된 베이지안 학습 모형을 개발해야 하고 또한 베이지안 학습과 신경망 모형을 결합한 베이지안 신경망(Bayesian neural networks) 모형까지 확장이 가능하다. 향후 딥러닝까지 활용할 수 있는 키워드 데이터 분석 모형전략이 개발된다면 텍스트 빅데이터 분석 분야에 대하여 의미 있는 기여가 가능할 것으로 기대된다.

참 고 문 헌

[1] C. Lesmeister, *Mastering Machine Learning with R, second edition*, Birmingham, UK, Packt, 2017.

[2] J. M. Kim, J. Yoon, S. Y. Hwang, and S. Jun, "Patent Keyword Analysis Using Time Series and Copula Models," Applied Science, Vol.9, No.19, p.4071, 2019.
[3] D. Uhm, J. Ryu, and S. Jun, "An Interval Estimation Method of Patent Keyword Data for Sustainable Technology Forecasting," Sustainability, Vol.9, No.11, p.2025, 2017.
[4] S. Park and S. Jun, "Patent Keyword Analysis of Disaster Artificial Intelligence Using Bayesian Network Modeling and Factor Analysis," Sustainability, Vol.12, p.505, 2020.
[5] J. Kim, J. Ryu, S. Lee, and S. Jun, "Penalized Regression Models for Patent Keyword Analysis," Model Assisted Statistics and Applications-International Journal, Vol.12, pp.239-244, 2017.
[6] J. Huh, "Big Data Analysis for Personalized Health Activities: Machine Learning Processing for Automatic Keyword Extraction Approach," Symmetry, Vol.10, No.4, p.93, 2018.
[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis, Third Edition*, Boca Raton, FL, Chapman & Hall/CRC Press, 2013.
[8] J. K. Kruschke, *Doing Bayesian Data Analysis, Second edition*, Waltham, MA, Elsevier, 2015.
[9] R. V. Hogg, J. M. McKean, and A. T. Craig, *Introduction to Mathematical Statistics, 8th edition*, Upper Saddle River, NJ, Pearson, 2018.
[10] T. M. Donovan and R. M. Mickey, *Bayesian Staistics for Beginners*, Oxford, UK: Oxford University Press, 2019.
[11] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019.
[12] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in R," Journal of Statistical Software, Vol.25, No.5, pp.1-54, 2008.
[13] I. Feinerer and K. Hornik, *Package 'tm' Ver. 0.7-4, Text Mining Package*, CRAN of R project, 2019.
[14] USPTO, "The United States Patent and Trademark Office," http://www.uspto.gov, 2019.
[15] WIPSON, "WIPS Corporation," http://www.wipson.com, 2019.
[16] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach, Third Edition*, Essex, UK: Pearson, 2014.

저 자 소 개

전 성 해(Sunghae Jun) 정회원

- 1993년 2월 : 인하대학교 통계학과(이학사)
- 1996년 2월 : 인하대학교 통계학과(이학석사)
- 2001년 8월 : 인하대학교 통계학과(이학박사)
- 2007년 2월 : 서강대학교 컴퓨터공학과(공학박사)
- 2013년 2월 : 고려대학교 정보경영공학과(공학박사)
- 2003년 3월 ~ 현재 : 청주대학교 빅데이터통계학과 교수

〈관심분야〉 : 데이터사이언스, 인공지능, 베이지안통계학