Journal of The Korea Society of Computer and Information
Vol. 21 No. 1, pp. 125-129, January 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.1.125

Bayesian Regression Modeling for Patent Keyword Analysis

JunHyeog Choi*, SungHae Jun**

Abstract

In this paper, we propose an efficient dynamic workload balancing strategy which improves the performance of high-performance computing system. The key idea of this dynamic workload balancing strategy is to minimize execution time of each job and to maximize the system throughput by effectively using system resource such as CPU, memory. Also, this strategy dynamically allocates job by considering demanded memory size of executing job and workload status of each node. If an overload node occurs due to allocated job, the proposed scheme migrates job, executing in overload nodes, to another free nodes and reduces the waiting time and execution time of job by balancing workload of each node. Through simulation, we show that the proposed dynamic workload balancing strategy based on CPU, memory improves the performance of high-performance computing system compared to previous strategies.

▸ Keyword : Allocation, Workload, Migration, Load balancing, Simulation

I. Introduction

대표적인 지식재산(intellectual property)인 특허는 오랜 기간 동안 발명에 대한 의욕을 고취시키고 기술에 대한 보급에 이바지해 왔다 [1-3]. 세계지식재산권기구(World Intellectual Property Organization: WIPO)를 통하여 전세계의 특허 시스템은 유지 관리되고 있다 [4]. 연구자는 자신의 연구개발을 통하여 얻어진 기술을 문서화하여 전세계의 특허청에 출원, 등록함으로써 자신의 기술에 대한 배타적인 권리를 일정기간 동안 보장 받는다 [1-2]. 기업과 국가의 기술 경쟁력은 특허 지식재산의 경영과 밀접한 연관성이 있기 때문에 연구개발(R&D) 기획단계에서부터 특허에 대한 분석이 이루어지고 있다. 왜냐하면 출원, 등록된 특허문서에는 연구, 개발된 기술에 대한 상세한 내용과 범위가 잘 나타나 있기 때문이다. 따라서 특정 기술에 대한 분석을 위한 가장 효과적인 정보원천의 하나로서 특허문서 데이터는 가치를 가지고 있다. 기술분석을 위한 특허분석은 기술로드맵(technology road-mapping) 작성, 시나리오분석, 델파이(Delphi) 분석 등 여러 분야에서 다양하게 이용되고 있다 [5-8]. 또한 특허데이터를 위한 다양한 통계분석 방법에 대한 연구결과들이 발표되고 이를 이용한 다양한 기술경영 전략이 이루어지고 있다 [9-16]. 대표적인 통계모형인 선형회귀분석(linear regression)을 이용한 특허데이터의 분석도 활발하게 이루어지고 있다. 하지만 선형회귀분석은 현재 관측된 데이터에 분석이 집중되기 때문에 데이터에 대하여 이전에 알고 있던 사전정보(prior)를 이용하지 않는다. 본 연구에서는 사전정보와 사후정보(posterior)를 모두 포함하는 베이지안 학습(Bayesian learning)을 회귀분석에 적용한 베이지안 회귀분석(Bayesian regression)을 이용하여 특허데이터의 분석을 수행한다. 즉, 이전에 알고 있던 특정기술에 대한 정보와 검색된 특허데이터에 기반한 현재의 기술정보를 결합하여 보다 정확한 분석모형을 구축한다. 특히 제안된 방법을 적용한 사례연구를 통하여 실제문제 해결을 위한 접근방안을 제시한다. 본 논문에서는 3차원 프린팅 기술(three-dimensional printing technology)과 관련된 특허문서를 검색하고 이를 베이지안 회귀모형으로 분석하여 3차원 프린팅 기술에 대한 기술분석을 수행한다. 본 논문의 2장에서는 베이지안 통계모형에 대하여 알

• First Author: JunHyeog Choi, Corresponding Author: SungHae Jun
*JunHyeog Choi(jhchoi@kimpo.ac.kr), Dept. of Secretarial Management, Kimpo College
**SungHae Jun(shjun@cju.ac.kr), Dept. of Statistics, CheongJu University
• Received: 2015. 01. 11, Revised: 2015. 01. 20, Accepted: 2016. 01. 26.
• This research was supported by KIMPO College's Research Fund

아보고 3장에서는 베이지안 회귀모형을 이용한 기술분석에 대한 방법론을 제안한다. 실제 기술분야인 3차원 프린팅 기술관련 특허분석에 대한 사례연구는 4장에서 다룬다. 마지막 장에서는 본 연구에 대한 결론을 제시하고 향후 연구과제에 대하여 알아본다.

II. Bayesian Statistical Model

베이지안 학습에서 가장 기본이 되는 베이즈 정리(Bayes' theorem)를 사용하기 위하여 먼저 표본공간(sample space) S가 다음과 같이 n개의 사상(events) F1, F2, ..., Fn으로 분할(partition)되어야 한다 [17-22].

$$S = \bigcup_{i=1}^{n} F_i \qquad (1)$$

이 때 n개의 Fi는 서로 배반(disjoint)이어야 한다. 이 때 사상 E의 발생확률 P(E)는 Fi를 이용하여 다음과 같이 계산할 수 있다 [17-22].

$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i) \qquad (2)$$

즉 P(E)는 P(E|Fi)에 P(Fi)를 곱한 결과로 해석할 수 있다. 확률 P(Fi)는 Fi에 대한 사전경험이 된다. 이와 같은 사전경험은 주어진 데이터에 의한 가능도함수(likelihood function) P(E|Fi)를 이용하여 다음과 같이 사후확률을 계산할 수 있다 [17-22].

$$P(F_i|E) = \frac{P(E|F_i)P(F_i)}{\sum_{j=1}^{n} P(E|F_j)P(F_j)} \qquad (3)$$

식 (3)의 베이즈 정리를 이용하여 베이지안 학습을 수행할 수 있다. 그림 1은 베이지안 학습을 나타내고 있다.

Fig. 1. Bayesian Learning

사전정보는 주어진 데이터와 결합하여 사후정보가 되고 사후정보는 다시 새롭게 얻어지는 데이터를 위한 사전정보로 사용된다. 이를 통해 다시 사후정보가 얻어지고 이것은 다시 다음의 수집데이터를 위한 사전정보로 사용된다. 이와 같은 반복을 통하여 정교한 모형을 구축할 수 있게 된다. 본 연구에서는 이와 같은 베이지안 학습을 회귀모형에 적용한 베이지안 회귀분석을 이용하여 특허데이터를 분석한다. 또한 구축된 모형의 성능평가는 다음과 같이 계산되는 아카이케 정보기준(Akaike information criterion: AIC)을 사용한다 [23].

$$AIC = n\log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2p \qquad (4)$$

위 식에서 n은 데이터의 크기를 나타내고 y_i는 실제값이고 $\hat{y}_i$는 예측값을 나타낸다. 또 p는 회귀계수의 개수를 의미한다. 일반적으로 AIC 값이 작을수록 추정된 회귀모형의 성능이 우수하다고 판단한다. 다음 절에서는 본 논문에서 제안하는 방법론에 대하여 설명한다.

III. Technology Analysis Using Bayesian Regression Model

일반적으로 회귀모형은 반응변수(response variable) Y와 예측변수(predictive variable) X로 이루어진다. 본 논문에서는 X와 Y 모두 검색된 특허문서 데이터로부터 추출된 키워드가 되는 다음의 모형으로 구성된다.

$$Y_{response\,keyword} = \beta X_{predictive\,keywords} + \epsilon \qquad (5)$$

반응 키워드(response keyword)는 1개, 예측 키워드(predictive keywords)는 1개 이상으로 이루어지며 예측 키워드를 이용하여 반응 키워드의 결과를 예측할 수 있다. β는 회귀계수(regression coefficients)가 되고 ϵ는 서로 독립이고 등분산(equal variance)을 갖는 오차 항이다. 따라서 반응 키워드 Y는 다음과 같은 정규분포(Normal distribution)를 갖는다.

$$Y_{|\beta,\sigma^2,X} \sim N(X\beta, \sigma^2) \qquad (6)$$

이와 같은 일반적인 회귀모형에 베이지안 학습이 적용될 때에는 사전분포(prior distribution)와 예측분포(predictive distribution)를 이용하여 사후분포(posterior distribution)를 구한다. 본 논문에서는 β와 σ^2에 대한 사전분포로 다음과 같은 비정보(noninformative) 사전분포를 사용한다 [22].

$$P(\beta,\sigma^2)\propto \frac{1}{\sigma^2} \quad (7)$$

즉 X가 주어졌을 때 β와 σ^2의 결합확률분포(joint probability distribution)는 $\frac{1}{\sigma^2}$에 비례한다. 이를 통하여 σ^2이 주어졌을 때 β의 사후분포는 다음과 같다.

$$\beta_{|\sigma^2,Y} \sim N((X'X)^{-1}X'Y,(X'X)^{-1}\sigma^2) \quad (8)$$

이와 같은 사전분포와 사후분포를 이용하여 일반적인 회귀분석과 베이지안 학습을 결합한 베이지안 회귀분석을 수행한다.

IV. 3D Printing Technology Analysis

제안방법을 적용한 사례분석을 위하여 본 논문에서는 3D 프린팅 관련 기술분석을 수행하였다. 방대한 3D 프린팅 관련 특허들 중에서 대표적인 3D 프린팅 회사인 '3D Systems'의 출원, 등록 특허를 수집하여 분석하였다. 즉 출원인이 '3D Systems'인 특허를 특허데이터베이스로부터 검색하였다 [24]. 총 검색된 특허수는 315건이었다. 베이지안 회귀모형을 적용한 특허분석을 위하여 대표적인 데이터 분석언어인 R과 R project에서 제공하는 다양한 패키지들을 사용하였다 [25-27]. 일반적으로 검색된 특허문서 데이터는 통계적 분석에 적합하지 않은 데이터구조를 가지고 있기 때문에 텍스트 마이닝의 전처리를 통하여 통계분석이 가능한 정형화된 데이터 구조로 변환하였다 [27]. 정화된 데이터는 행렬구조이고 각 열은 키워드, 그리고 각 행은 특허로 이루어진다. 또한 행렬의 원소(element)는 특정 키워드가 각 특허에 나타난 빈도값을 나타낸다. 키워드 추출을 위하여 기존의 3D 프린팅 기술분석 관련 연구결과를 이용하였다 [28]. 선정된 키워드들 중에서 '3D'와 'Printing'을 각각 반응 키워드로 하고 나머지 전체 키워드들을 예측 키워드로 사용하여 베이지안 회귀분석을 수행하였다. 표 1은 반응 키워드로 '3D'를 선택하고 나머지 키워드를 예측 키워드로 사용한 베이지안 회귀분석의 결과를 나타내고 있다.

Table 1. Bayesian Regression: Response-3D

Predictive	Estimate	S.E.	p-value
Deposition	-0.3561	0.0998	0.0004
Modeling	0.3548	0.0487	0.0001
Object	0.6223	0.0596	0.0001
Pour	-0.1776	0.0745	0.0178
Material	0.1992	0.0928	0.0327
Imaging	-0.1394	0.0556	0.0127
System	-0.0736	0.0367	0.0459
Freeform	-0.3227	0.1366	0.0188
Printing	0.3109	0.0781	0.0001

각 예측 키워드에 대한 회귀계수의 추정치(estimate), 표준오차(standard error: S.E.), 그리고 유의확률(p-value)에 대한 계산결과를 나타내고 있다. 전체 예측 키워드들 중에서 유의수준(α) 0.05에서 통계적으로 유의한 키워드들만을 포함하고 있다. 'Modeling', 'Object', 'Printing'의 유의확률이 모두 0.0001보다 작기 때문에 '3D' 키워드에 통계적으로 매우 유의하게 영향을 미치고 있음을 알 수 있다. 표1의 베이지안 회귀분석의 AIC 값은 201.72로 나왔다. 다음은 'Printing'을 반응 키워드로 결정하고 나머지 전체 키워드들을 예측 키워드로 사용한 베이지안 회귀분석 결과를 나타내고 있다.

Table 2. Bayesian Regression Result: Response-Printing

Predictive	Estimate	S.E.	p-value
3D	0.1705	0.0428	0.0001
Modeling	-0.1141	0.0387	0.0035
Object	-0.1552	0.0512	0.0026
Material	0.1873	0.0684	0.0066
Control	-0.1371	0.0448	0.0024
Method	0.0877	0.0304	0.0042
System	0.1295	0.0263	0.0001
Manufacturing	-0.1474	0.0629	0.0198

표1의 결과와 마찬가지로 표2에서도 유의수준 0.05에서 통계적으로 유의한 예측 키워드만을 선택하여 나타내고 있다. '3D'와 'System' 키워드의 유의확률이 0.0001보다 작음을 알 수 있다. 즉 이들 예측 키워드가 'Printing' 키워드에 매우 큰 영향을 미치고 있음을 알 수 있다. 표1과 표2의 결과를 통하여 'Modeling', 'Object', 'Material', 그리고 'System'의 4개 예측 키워드들은 '3D'와 'Printing'에 동시에 통계적으로 유의한 영향을 미치고 있음을 알 수 있다. 또한 AIC 값은 12.60으로 이전의 '3D' 키워드를 반응 키워드로 사용한 베이지안 회귀모형에 비해 모형의 성능이 더 우수함을 알 수 있다. 표1과 표2의 결과를 종합하여 3D 프린팅 기술에 대한 기술연관도(technology association map)를 그리면 다음과 같다.

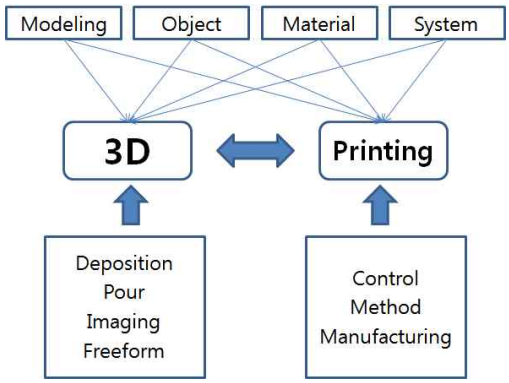

Fig. 2. Technology Association Map

'3D'와 'Printing'은 서로 영향을 미치고 있고, 'Modeling', 'Object', 'Material', 그리고 'System'의 4개 키워드들은 '3D'와 'Printing'에 동시에 영향을 미치고 있다. 즉 3D 프린팅 기술개발에서 'Modeling', 'Object', 'Material', 그리고 'System'과 관련된 세부기술 개발이 중요함을 알 수 있다. 또한 '3D' 관련 기술개발에서는 'Deposition', 'Pour', 'Imaging', 'Freeform'이 중요한 세부기술이 되고 있음을 확인할 수 있다. 또한 'Print' 기술개발에 있어서는 'Control', 'Method', 그리고 'Manufacturing' 기술 개발이 선행되어야 함을 알 수 있다. 본 연구에서는 그림2의 3D 프린팅 기술연관도를 통하여 이 기술분야의 기술분석을 수행하였다. 이 결과는 3D 프린팅 기술에 대한 연구개발 계획에 사용될 수 있다. 그림3은 3D 프린팅 기술에서 중요하게 다루어지는 4개의 키워드들에 대한 이산히스토그램(discrete histogram)을 나타내고 있다.

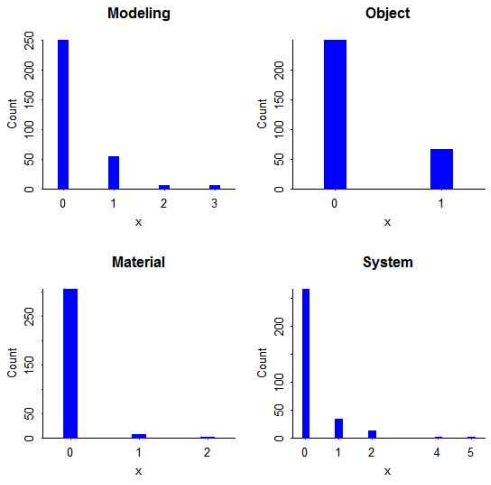

Fig. 3. Discrete Histogram of Major Keywords

'Object'나 'Material'에 비해 'Modeling'과 'System'의 빈도가 더 많이 퍼져 있음을 알 수 있다. 즉 'Modeling'과 'System'과 관련된 기술이 3D 프린팅 기술에서 더 광범위 하게 사용될 수 있음을 확인할 수 있다.

V. Conclusions

본 논문은 베이지안 회귀분석을 이용하여 특허 키워드를 분석하는 방법론을 제안하였다. 베이지안 회귀모형은 일반적인 회귀모형에 비해 회귀모수에 대한 사전정보를 모형에 추가할 수 있기 때문에 모형의 예측력 향상을 기대할 수 있다. 제안 방법에서는 비정보 사전분포를 사용하여 사전정보를 추가하였다. 물론 비정보 사전분포함수 뿐만 아니라 공액사전확률함수(conjugate prior probability function), 제프리의 사전함수(Jeffrey's prior function) 등을 이용하여 보다 정교한 베이지안 회귀모형을 구축할 수 있다. 향후 연구과제에서는 이와 같이 다양한 사전분포함수들을 사용하여 모형의 성능을 향상시킬 수 있을 것이다.

제안방법의 실제 적용을 위하여 본 연구에서는 3D 프린팅 기술관련 특허분석을 수행하였다. 베이지안 회귀분석의 최종적인 결과로 얻어지는 기술연관도를 이용하여 3D 프린팅 기술관련 연구개발 기획에 활용하여 효과적인 기술경영 전략을 수립할 수 있게 된다. 본 연구는 3D 프린팅 기술뿐만 아니라 다른 다양한 기술분야에도 적용하여 해당 기술분야의 기술경영 전략 수립에 사용될 수 있다. 제안된 방법론에 의해 구축되는 기술연관도의 해석과 최종적인 적용은 해당 기술분야의 전문가에 의해 보다 효율적으로 사용될 수 있을 것이다.

REFERENCE

[1] Patent and Information Analysis, KIPO, 2007.
[2] Roper, A. T., Cunningham, S. W., Porter, A. L., Mason, T. W., Rossini F. A., Banks J. Forecasting and Management of Technology, John Wiley & Sons, 2011.
[3] Hunt, D., Nguyen, L., Rodgers, M., Patent Searching Tools & Techniques, Wiley, 2007.
[4] WIPO, World Intellectual Property Organization, www.wipo.org, 2014.
[5] Cyert, R. M., Kumar, P., "Technology Management and the Future", IEEE Transactions on Engineering Management, Vol. 41, No. 4, pp. 333-334, 1994.
[6] McDermott, C. M., Kang, H., Walsh, S., "A Framework for Technology Management in Services", IEEE Transactions on Engineering Management, Vol. 48, No. 3, pp. 333-341, 2001.
[7] Yun, Y. C., Jeong, G. H., Kim, S. H., "A Delphi technology forecasting approach using a semi-Markov concept", Technological Forecasting and Social Change, Vol. 40, pp. 273-287, 1991.

[8] Martino, J. P., "Technology forecasting - An overview", Management Science, Vol. 26, No. 1, pp. 28-33, 1980.
[9] Jun, S., Park, S., Jang, D. "Technology Forecasting using Matrix Map and Patent Clustering", Industrial Management & Data Systems, Vol. 112, Iss. 5, pp. 786-807, 2012.
[10] Daim, T. U., Rueda, G., Martin, H., Gerdsri, P. "Forecasting emerging technologies: Use of bibliometrics and patent analysis", Technological Forecasting and Social Change, Vol. 73, Iss. 8, pp. 981 - 1012, 2006.
[11] Jun, S. "IPC Code Analysis of Patent Documents Using Association Rules and Maps-Patent Analysis of Database Technology", Communications in Computer and Information Science, Vol. 258, pp. 21-30, 2011.
[12] Hwang, J., Kim, B. "Analysis on the multi technology capabilities of Korea and Taiwan using patent bibliometrics," Asian Journal of Technology Innovation, Vol. 14, No. 2, pp. 183 - 199, 2006.
[13] Mishra, S., Deshmukh, S. G., Vrat, P., "Matching of technological forecasting technique to a technology", Technological Forecasting and Social Change, Vol. 69, pp. 1-27, 2002.
[14] Lee, S., Jun, S., "Key IPC Codes Extraction Using Classification and Regression Tree Structure," Advances in Intelligent Systems and ComputingVolume 271, pp 101-109, 2014.
[15] J. Choi, S. Jun, "A Technology Analysis Model using Dynamic Time Warping", Journal of the Korea Society of Computer and Information, Vol. 20, No. 2, pp. 113-120, 2015.
[16] Jun, S. "A clustering method of highly dimensional patent data using Bayesian approach", International Journal of Computer Science Issues, Vol. 9, No. 1, pp. 7-11, 2012.
[17] Korb, K. B., Nicholson, A. E., Bayesian Artificial Intelligence, London, UK, Chapman & Hall/CRC, 2004.
[18] Ross, S. M. Introductory Statistics, AP Elsevier, 2010
[19] Ross, S. M. Introduction to Probability and Statistics for Engineers and Scientists, Elsevier, 2012.
[20] W. C. Kim et al., Statistics, 4th edition, YoungJi, 2003
[21] Jerak, A., & Wagner, S. "Modeling probabilities of patent oppositions in a Bayesian semiparametric regression framework", Empirical Economics, Vol. 31, No. 2, pp. 513-533, 2006.
[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian Data Analysis, Third Edition, Boca Raton, FL, Chapman & Hall/CRC Press, 2013.
[23] Akritas, M., Probability and Statistics with R for Engineers and Scientists, Boston, Pearson, 2016.
[24] WIPSON, WIPS Corporation. http://www.wipson.com, 2015.
[25] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org, 2015.
[26] A. Gelman, Y.-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, T. Zheng, V. Dorie, Data Analysis Using Regression and Multilevel/Hierarchical Models, Package 'arm', CRANS, R-Project, 2015
[27] Feinerer, I., Hornik, K., Meyer, D. "Text mining infrastructure in R", Journal of Statistical Software, Vol. 25, No. 5, pp. 1-54, 2008.
[28] Park, S., Kim, J., Jang, D., Lee, H., Jun, S., "Methodology of Technological Evolution for Three-dimensional Printing," Industrial Management & Data Systems, Vol. 116, No. 1, pp. 122-146, 2016.

Authors

JunHyeog Choi received a B.S. degree in Computer Science from Kyunggi University, Korea in 1990, a M.S. and a Ph.D. degree in Computer Science from Inha University, Korea in 1995 and 2000 respectively, He also received a MBA and Ph.D. degree in Management of Technology from Yonsei University, Korea, in 2003 and 2013 respectively. He worked as a invited scholar in Software research center, at ETRI. He is currently a professor in the Department of Secretarial Management, Kimpo College. His Research interests include Patent Analysis, Management of Technology, and Technology forecasting and IoT.

Sunghae Jun is professor at department of statistics Cheongju University, Korea. He recieved Phd department of management engineering at Korea University in 2013. He was visiting professor in department of statistics at Oklahoma State University, USA during 2009 to 2010. He has studied on big data analysis and technogy forecasting.