

데이터분석을 위한 프로그래밍 언어들의 수요추이에 대한 탐색적 연구

이우창¹, 문수현¹, 정세민², 이충권^{2*}¹ 계명대학교 통계학과² 계명대학교 경영정보학과

e-mail : bigdata_analyst@naver.com, ag8617@naver.com, hiemily@naver.com, cklee@kmu.ac.kr

An Exploratory Study of the Demanding Trends of Programming Languages for Data Analysis

Woo Chang Lee¹, Su Hyeon Moon¹, Se Min Jung², Choong Kwon Lee^{2*}¹Dept. of Statistics, Keimyung University²Dept. of MIS, Keimyung University

요 약

프로그래밍 언어는 컴퓨터와 함께 지속적으로 발전해 왔다. 수천 개의 프로그래밍 언어들이 소프트웨어 개발자들의 손을 거쳐 갔고, C 또는 Java 처럼 오랫동안 인기를 누린 대표적인 언어도 있다. 그러나 등장과 함께 많은 주목을 받았지만 일찍 사라지거나 개발자들로 부터 주목조차 받지 못하는 언어도 있었다. 최근에 빅데이터의 폭발적인 수요가 발생하면서 R, Python 그리고 SAS와 같은 언어들이 인기를 얻고 있다. 이에 본 연구는 C, Java와 같은 전통적인 언어를 R, SAS 및 Python과 비교하여 시계열 분석을 이용하여 추세를 탐색하였다. 분석을 위해 취업 포털 사이트인 'CareerBuilder'에 게재된 1072개의 구인광고를 수집하였다. 시계열 분석의 결과에 따르면, 구인광고에서 주간 별로 문서 당 평균빈도가 증가하는 추세를 가지는 프로그래밍 언어는 Python과 SAS로 나타났고 Java는 감소하는 추세를 보였다. 그리고 C와 R은 중립적인 추세를 가지는 것으로 나타났다.

1. 서 론

인공지능, 로봇, 빅데이터 등과 같은 기술 분야의 급속한 발전으로 기업과 공공 조직들은 컴퓨터 프로그래밍의 중요성을 다시 한 번 인지하게 되었다. 우수한 프로그래밍 능력을 보유한 인력을 확보하기 위하여 기업들은 다양한 인력확보 전략을 활용하고 있다. 소프트웨어 개발인력 양성을 위하여 교육기관들과 협약을 체결하기도 하고, 자사 내의 인력을 선발하여 별도의 교육을 실시하기도 한다. 이것은 다가오는 4차 산업혁명에서 기업들의 경쟁력이 하드웨어 보다는 소프트웨어에서 결정될 것이라고 믿기 때문이다. 특히 IT 산업에서 소프트웨어를 만드는 프로그래밍을 수행하는 인력의 보유 및 역량수준은 기업의 핵심성공요인이자 다른 기업들과 경쟁에서 우위를 선점할 수 있는 근원으로 인식되고 있다 [1].

소프트웨어를 개발하기 위하여 발명되는 프로그래밍 언어는 목적과 용도에 따라 다양하게 발전해왔고, 지금도 새로운 언어들이 만들어지고 있다. 따라서 프로그래밍을 학습하고자 하는 사람에게 어떤 언어를 배울 것인가는 중요한 의사결정이다. 자신의 취업이나 이직에 유리한 선택을 하려면, 실제 기업에서 인력수요가 많은 프로그래밍 언어를 학습하여야 한다. 그러나 어떤 프로그래밍 언어들이 현재의 시점에서 꾸준한 수요가 있고 미래에 중요해질 것

인지를 예측하는 연구는 거의 없었다. 특히 4차 산업혁명과 관련하여 최근에 등장한 인공지능, 빅데이터 등과 같은 데이터 분석 분야에서 활용도가 높은 R, Python 그리고 SAS와 같은 언어들이 기존의 전통적인 언어들과 비교하여 인력시장에서 어느 정도의 수요가 있는 지를 파악할 필요가 있다. 이에 본 연구는 구인광고 분석을 통하여 기존의 전통적인 프로그래밍 언어인 C, Java 등과 같은 언어들과 최근 데이터 분석으로 주목받고 있는 R, Python 그리고 SAS의 인력수요 추이를 탐색하고자 한다.

2. 본 론

2.1 데이터 수집

본 연구에서는 실제 잡포털 사이트에서 게재되는 구인광고에 나타나는 프로그래밍 언어들에 대한 주별 문서 당 단어의 평균 빈도를 계산하여 개별 언어들의 추이를 알아보기 위하여 시계열 분석을 실시하고자 한다. 따라서 미국 내 최대 규모의 온라인 취업사이트이고 게시글의 게재날짜가 명확하게 나타나는 'Careerbuilder' 에서 2016년 10월 12일에서 2017년 8월 22일까지 구인광고들 중에서 제목에 "프로그래머"라는 단어를 가진 1,072개의 데이터를 수집하였다.

2.2 데이터 전처리

본 연구는 구인광고 텍스트데이터 1072개에서 나타나는 프로

* 교신저자

그램 언어를 통해 주간 문서 당 평균 단어빈도를 구한 후 시계열 분석을 이용하여 프로그래밍 언어에 대한 기간별 추이를 파악하고자 한다. 따라서 각 프로그램 언어에 대한 주별 문서 당 단어의 평균빈도를 도출해내기 위하여 Python을 활용하였다. 형태소 분리와 불용어 제거를 하는 과정을 수행하였고 프로그램 언어에 대한 단어의 빈도들을 추출하였다. 그 후 같은 의미를 가지는 단어를 통합하는 데이터 전처리를 하였다.

2.3 데이터 분석

본 연구의 분석기간 단위는 45주이다. 특정 주에는 구인광고가 존재하지 않는 데이터가 상당히 많이 존재하는 경우 모델을 적합 시켜도 정확도가 매우 떨어지는 특징이 있다[2]. 따라서 구인광고에서 최소 16주 이상 단어가 출현하였고 시계열 모델을 적합 시켜 추이를 분석하기에 적절하다고 판단하여 본 연구의 분석 단위로 설정하였다. 표 1은 프로그래밍 언어가 출현한 주의 빈도와 해당 프로그램 언어를 나타낸 결과이다.

(표 1) 프로그램 언어가 출현한 주의 빈도

단어가 나타난 주의 빈도	프로그래밍 언어
31주 이상	Java
26~30주	C, SAS
21~25주	Python
16~20주	R

다섯 가지 프로그래밍 언어들의 추이를 예측하기 위하여 주별 문서 당 단어의 평균빈도를 구하여 2016년 10월 12일에서 2017년 6월 20일까지 총 36주 간 자료에 대해 시계열 분석을 위한 기술 통계량(평균, 표준편차, 최댓값, 최솟값)을 계산하였다. 다음으로 분석기간 동안 모형의 모수를 추정하고 잔차의 독립 여부를 확인하기 위하여 Ljung-Box 검정을 실시하였다. 그 후 주어진 데이터에 적합도가 가장 높은 모델을 선정하였다. 마지막으로 예측 기간인 2017년 6월 21일에서 2017년 8월 22일까지 총 9주간 실제 값과 예측 값의 비교를 통하여 모델에 대한 최종 평가를 실시하였다.

시계열분석은 과거 시계열의 형태가 미래에도 영향을 미친다는 가정 하에 과거에 관측된 값들의 상호관계로 모형을 구축하여 미래에 대한 예측을 하는 것이며 대표적인 방법론은 ARIMA 모형이다[3]. ARIMA 모형은 시계열 데이터에서 과거의 관측값들이 설명변수인 AR(Auto Regression)모형과 과거의 오차항들이 설명변수인 MA(Moving Average)모형을 모두 고려한 모형이다. 여기서 ARMA(Autoregressive Moving Average model)모형의 식은 수식 1과 같다. 여기서 Z_t 는 시계열 데이터, β_p 는 자기회귀 계수, p 는 자기회귀 시차, θ_q 는 이동평균 계수, q 는 이동평균 시차, ϵ_t 는 오차항 또는 백색잡음(White Noise)을 의미한다.

$$Z_t = \beta_0 + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \dots + \beta_p Z_{t-p} + \epsilon_t - \theta_0 - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q} \quad (1)$$

오차의 추정은 시계열 모형을 선택하고 모형을 평가하기에 매우 중요한 판단기준이다. 본 연구에서는 RMSE(Root Mean Squared Error), AIC(Akaike's Information Criterion)와 SBC(Schwarz-Bayesian Information Criterion)를 오차추정의 척도로 지정하였다. RMSE는 MSE(Mean Squared Error)의 제곱근인 표준오차로 RMSE가 작을수록 모형의 성능이 뛰어나는 의미를 가진다[4]. 이 방식을 오차의 추정의 척도로 적용할 경우 오차 값의 크기에 따라 상대적으로 각 시점에서의 영향력이 다소 과장하게 나타날 수 있으며, 이러한 특성은 오차의 크기와 효과에 따라 모형 선정에서 중요하게 기능할 수 있다[5]. RMSE는 수식 2와 같이 나타난다.

$$RMSE = \sqrt{\frac{1}{n} \sum (F_i - Z_i)^2} = \sqrt{MSE} \quad (2)$$

AIC는 ARIMA 모형이 관측된 계열에 얼마나 적합한지를 고려하고자 할 때 유용하게 사용되며, 같은 데이터에 다른 모형들을 적합시켰을 때 AIC가 작을수록 좋은 모형이 된다[5]. AIC의 식은 수식 3과 같다.

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2(p + q + P + Q) \quad (3)$$

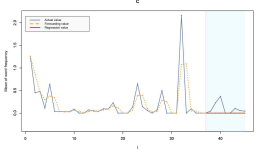
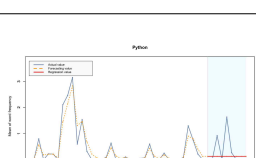
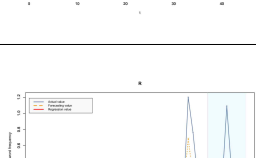
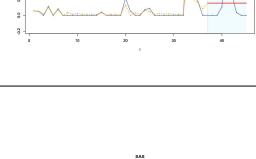
SBC는 AIC와 함께 모수추정에 있어 객관적 기준으로 사용되고 있는데, 시계열 모형을 선택할 때 SBC의 수치가 작을수록 정확한 차수로 보고, 최적의 모형이라고 해석한다[3]. SBC는 수식 4와 같다.

$$SBC = n \ln\left(\frac{SSE}{n}\right) + (p + q + P + Q) \ln(n) \quad (4)$$

2.4 분석 결과

본 연구에서는 2016년 10월 12일에서 2017년 6월 20일까지 각 프로그래밍 언어의 주간 문서 당 단어의 평균빈도를 토대로 ARIMA모형을 구축하여 2017년 6월 21일에서 2017년 8월 22일까지의 데이터를 예측하였다. 다섯 개의 프로그래밍 언어들에 대하여 적합 시킨 ARIMA모형의 경우 추정된 모수들의 유의확률이 모두 0.05이하로 나타났다. 모형들의 Ljung-Box 검정의 결과, 5개의 프로그램 언어 모두 유의확률이 모두 0.05보다 크게 나타나 백색잡음 항이 독립인 것으로 나타났다. 또한 앞으로의 프로그램 언어 추이를 알아보기 위해 예측기간안의 데이터에 회귀직선을 그어본 결과 증가하는 추이를 가지는 언어는 Python과 SAS로 나타났다. 이에 반해 감소하는 추이를 가지는 언어는 Java로 나타났으며, C와 R은 중립적인 추세를 가지는 것으로 나타났다. 표 2는 데이터에 대한 분석결과를 종합적으로 보여준다.

(표 2) 기술통계 및 시계열 분석 결과

프로그래밍 언어	시계열 도표	기술통계량		오차추정 및 잔차 검정		최적모형 모수추정	회귀계수
C		최소	0.000	RMSE	0.152	ARIMA(1,1,0) AR(1) : -0.502(0.001) ***	$\hat{\beta} : 0.000$
		최대	0.368	Ljung Box Test	45.944 (0.083)		
		평균	0.094	AIC	53.751		
		표준편차	0.126	SBC	55.306		
Java		최소	0.000	RMSE	0.193	ARIMA(1,0,0) AR(1) : 0.400(0.009) **	$\hat{\beta} : -3.833E-06$
		최대	0.423	Ljung Box Test	37.796 (0.300)		
		평균	0.144	AIC	52.194		
		표준편차	0.138	SBC	53.778		
Python		최소	0.000	RMSE	0.578	ARIMA(0,1,1) MA(1) : -0.422(0.006) **	$\hat{\beta} : 5.551E-18$
		최대	1.636	Ljung Box Test	20.940 (0.961)		
		평균	0.319	AIC	83.266		
		표준편차	0.580	SBC	84.850		
R		최소	0.000	RMSE	0.338	ARIMA(0,1,2) MA(1) : 0.435(0.008) ** MA(2) : 0.442(0.018) *	$\hat{\beta} : 0.000$
		최대	1.091	Ljung Box Test	15.867 (0.997)		
		평균	0.166	AIC	-4.851		
		표준편차	0.358	SBC	-1.741		
SAS		최소	0.000	RMSE	0.269	ARIMA(0,0,2) MA(1) : -0.671(0.000) *** MA(2) : -0.323(0.049) *	$\hat{\beta} : 0.010$
		최대	0.759	Ljung Box Test	22.640 (0.931)		
		평균	0.119	AIC	132.150		
		표준편차	0.249	SBC	135.317		

3. 결 론

본 연구는 시계열 분석을 이용하여 Python, SAS, Java, C, R 프로그램의 추이를 예측하였다. 본 연구에서 도출해낸 연구결과를 통해 다음과 같은 시사점을 제시하고자 한다.

첫째, 프로그램 언어 중 Python과 SAS는 증가하는 추이로 나타났다. Python은 C언어를 기반으로 한 오픈소스 프로그래밍 언어지만 C언어와는 달리 인터프리터(Interpreter)식 동적 타이핑(Dynamically typing) 대화형 언어라는 강점을 가지고 있다. 이러한 형식은 사용자가 컴파일(Compile)을 하지 않고서 작성한 프로그램을 간편히 실행할 수 있어 사용자가 C언어보다 쉽게 사용할 수 있다. 또한 Python은 웹 개발이 가능하며, 데이터 분석, 머신러닝, 그래픽, 학술 연구 등의 여러 분야에서 활용되는 범위가 넓다는 장점을 가지고 있어 최근 다양한 업계에서 사용할 것으로 판단된다.

SAS는 통계 분석에 특화된 프로그램 언어로 데이터를 올바른 정보로 변환하여 정보를 필요로 하는 사람에게 가장 적절한 시기

에 정보를 제공할 수 있는 통합 어플리케이션 소프트웨어라고 할 수 있다. SAS는 자료 관리와 자료 처리에 있어 효율성을 보이며 COBOL과 같은 컴퓨터 언어들을 이용하면 며칠씩이나 걸리는 프로그램 작업도 SAS를 이용하면 몇 줄의 프로그램으로 간단하게 끝낼 수 있으므로 사용의 용이성이 있다. 또한 다른 언어를 배우지 않아도 간편하게 통계적 처리를 쉽게 할 수 있고 보고서 작성에 용이하다는 장점도 있다. 이러한 장점들과 편리성에 의해 프로그래머들이 SAS프로그램을 앞으로 더 많이 사용할 것이라고 예상된다.

둘째, 프로그램 언어 중 Java는 감소하는 추이로 나타났다. Java는 이용자의 요구를 받아들여 응답하는 텔레비전을 뜻하는 인터랙티브 텔레비전(interactive television)전용으로 개발된 객체 지향 언어이며 현재 가장 널리 사용되는 프로그래밍 언어 중 하나이다. 하지만 Java코드는 구문이 너무 장황하여 함수를 전달하기가 쉽지 않고 작업 하나를 완료하는 데 너무 많은 확장이 필요하다는 점이 있어 구문이 비교적 간단하고 함수 전달이 용이한

Python과 같은 언어가 Java를 대체할 것으로 보인다.

셋째, 프로그램 언어 중 C와 R은 중립적인 추이로 나타났다. C는 세계적으로 가장 많이 쓰이는 프로그래밍 언어 중 하나이며 절차 지향적 특성을 지닌 언어이다. C로 작성된 프로그램은 CPU의 종류에 상관없이 실행이 가능하고, 운영체제의 차이에도 민감하지 않다는 장점이 있어 이식성이 좋다. 또한 다른 언어들에 비해서 사용하는 메모리의 양이 상대적으로 적고, 속도를 저하시키는 요소들을 최소화시킬 수 있다는 장점이 있다[10]. 또한 Unix 운영체제가 C로 작성되었으며 Perl, Python, Pascal, LISP 등과 같은 언어들의 컴파일러와 인터프리터도 C로 만들어져 강력하고 유연하다는 장점이 있다는 것을 알 수 있다. 따라서 C는 지속적으로 수요가 있는 프로그래밍 언어가 될 것으로 기대된다.

R은 다양한 통계 계산의 지원뿐만 아니라 그래픽 부분에서도 고품질로 자유롭게 나타낼 수 있는 프로그래밍 언어이다. 특히 패키지 개발이 용이하다는 장점이 있어 통계학자들 사이에서 통계 소프트웨어를 개발하는데 주로 사용되고 있다[6]. 또한 R은 자유 소프트웨어로 사용자가 제작한 패키지를 추가하고 사용할 수 있어 통계학자 외의 사용자들도 쉽게 사용할 수 있다. 따라서 최근 빅데이터 기술이 새로운 산업 구조의 변화를 이끌어 나가는 추세에서 데이터 분석가 사이 지속적인 수요와 공급이 있을 것으로 기대된다.

현재 빠르게 변화하는 정보기술에 따라 프로그램 언어의 추이도 빠르게 변화하고 있다. 따라서 데이터 분석가를 준비하는 사람들은 IT와 관련한 빠른 변화를 확인하고, 정보기술분야의 구인광고를 분석하여 어떤 프로그래밍 역량이 필요해지는지 지속적으로 확인할 필요가 있다. 본 연구의 결과를 바탕으로 구직자들이 기업에서 요구하는 프로그램 언어의 추이를 확인하고, 향후 지속적인 연구를 통하여 컴퓨터 언어들의 추이를 더 정확히 예측한다면, 더욱 의미 있는 결과를 도출할 수 있을 것으로 예상된다.

참 고 문 헌

- [1] 김범성, “중소 IT기업의 인적 역량과 구성이 재무성에 미치는 영향”, 『대한경영학회지』 제24권 제2호, 2011. pp.1071-1093.
- [2] 장영순, 서종현, “다수의 결측치가 존재하는 가전업 고객 데이터 활용을 위한 고객분류기법의 개발”, 『대한산업공학회』 제19권, 제1호, 2006. pp.86-96.
- [3] 김시연, 정현우, 박정도, 백승목, 김우선, 전경희, 송건빈, “계절 ARIMA 모형을 이용한 104주 주간 최대 전력수요예측”, 『조명·전기설비학회논문지』 제28권, 제1호, 2014. pp.50-56.
- [4] 김은미, 이배호, “모멘트의 동적 변환에 의한 Kernel Relaxation의 성능과 RMSE”, 『멀티미디어학회논문지』 제6권, 제5호, 2003. pp.788-796.
- [5] 임상범, 박선형, “시계열 분석을 통한 시도별 고등학교 학생 수 예측”, 『한국콘텐츠학회논문지』 제16권, 제12호, 2016. pp.735-748.
- [6] 이종기, “회계자료 빅데이터 분석을 위한 R프로그래밍 실무 처리 사례”, 『전산회계연구』 제13권, 제1호, 2015. pp.1-22.