R을 이용한 분석 시스템 구축 방안 및 설계

[1]권영오, [2]진병삼, [*3]배영철
[1, 전남대학교 iahalu@naver.com
[2, 전남대학교 bsjin@misotech.net
[*3] 전남대학교, ycbae@jnu.ac.kr

Design and Construction of Analysis System using R

[1]Youngoh Kwon, [2]Byoungsam Jin, [*3]Youngchul Bae
[r Chonnam National University, iahalu@naver.com
[2] Chonnam National University, bsjin@misotech.net
[*3] Chonnam National University, ycbae@jnu.ac.kr

요 약

최근에 이르러 많은 분석자들은 파이썬과 R 등을 이용하여 다양한 분석 및 통계를 효율적으로 사용하여 데이터를 분석하고 있다. 하지만 웹 시스템을 통하여 다양한 분석을 실시간적으로 표현 하고자 하는 요구 사항을 두가지 언어만을 이용하여 해결하기는 어렵다. 그리고 R 에서 P Value 등을 추출할 때 적합한 알고리즘을 적용하지만 Java 에서는 해당하는 알고리즘이 없거나 또는 해당 소스가 검증이 되지 아니하여 추출된 P-Value 값이 R 에서의 값과 다른 것을 확인 할 수 있다. 따라서 본 논문에서는 다른 언어에서 R 을 이용하여 분석자의 분석 알고리즘 내용을 바로 자신의 시스템에 적용 할 수 있는 방안을 제안하고자 한다.

Abstract

In recent years, many analysts have been using Python and R to analyze data using efficient analysis and statistics. However, it is difficult to solve the requirement to express various analyzes in real time through the web system using only two languages. In addition, it is possible to confirm that the extracted P-Value is different from the value in R because there is no corresponding algorithm in Java or the corresponding source is not verified. Therefore, in this paper, we propose a method to apply analyst 's analysis algorithm contents directly to his system using R in other languages.

***Keywords**: R language, R linkage, System Integration, R Serve, P-value*

1. 서론

최근에 이르러 많은 분석자들은 R 과 파이썬을 이용하여 다양한 분석 결과를 보고서에 제공하며 활용하고 있다. 특히 R 은 오픈소스로서 GNU General Public License 를 가지면서 오픈 소스라는 강점과 함께 사용자가 폭발적인 증가하고 확산하여 많은 최신 통계 알고리즘이 보급화 되었고, 대학 교육의 표준 툴로 자리를 잡았다.

Google 같은 경우에는 많은 사이트들이 연동하여 사용하고 있는 Google Analytics 도 R 을 이용하여 제공하고 있는 것으로 알고 있다. SAS 나 SPSS 등의 분석자들 사이에서 많이 사용되는 있는 상용 통계 패키지도 매우 희귀한 분석이나 다양한 최신의 알고리즘을 적용 할 경우에는 R 의 패키지를 활용 함으로서 최신의 분석 알고리즘을 제공하고 있다. R 은

* Corresponding Author
Received: Dec. 2, 2017, Revised: Dec. 27, 2018, Accepted: Dec. 21, 2017

오픈소스로 인해 최신의 알고리즘과 희귀성의 분석 기능이 빠르게 적용되기 때문이다[1]. 그래서인지 R 은 국내에서 매우 빠르게 확산되었고 구글에서 검색 되어지는 관심도의 속도가 증가되고 있다. 그 내용은 그림 1 에서와 같이 구글 트랜드에서 증명 되어지고 있다. 사용자가 많은 만큼 R 의 모듈들도 안정화를 거치고 많은 사용자들에게 검증이 된 올바른 연산을 처리하고 있다.

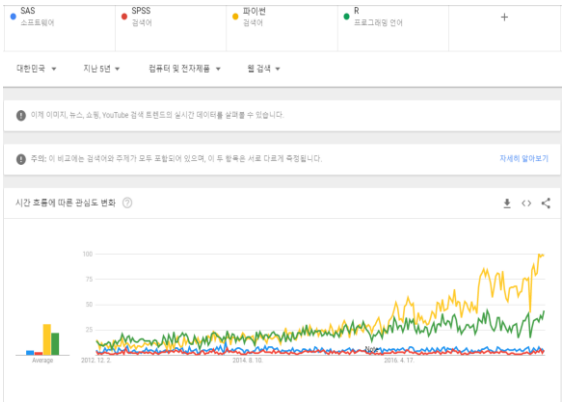

Figure 1. Change of interest in analysis system

Hadoop 은 컴퓨터 클러스터에서 분산저장 및 분산처리를 위해 개발된 오픈 소스의 프레임워크이다. 하둡 기반에서 분산처리가 가능해진 R 은 기존에 가지고 있던 문제점인 In-Memory 연산의 특징인 메모리 한계 문제를 해결하게 되었다. R 은 RHIPE(R and Hadoop Integrated Processing)와 RHadoop 같은 R 과 하둡의 연계 모듈의 등장으로 R 은 하둡에서 사용하는 맵리듀스의 분산처리 프로세스를 연계 사용함으로서 대량의 분석을 처리 및 속도의 향상을 가지는 것이 가능하게 되어 더욱 많이 관심을 가지게 되었고 이기종의 언어에서 연계 하여 사용하고자 하는 요구사항은 커지게 되었다.

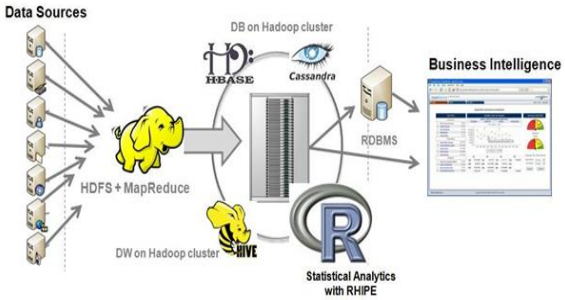

Figure 2. Using Hadoop-based R [1]

Rserve 는 GPL License 를 가지고 있는 R 의 라이브러리로서 Java, C, C++, PHP 와 같은 다른 프로그램 언어에서 TCP/IP 로 R 에 원격 접속, 인증, 파일 전송을 가능하게 해준다. Rserve 는 최근까지도 최신버젼이 지속적으로 업데이트가 되고 있고, R 연계의 요구사항을 해결해 주고 있다.

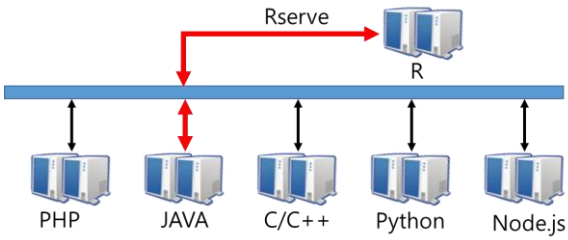

Figure 3. Interworking between R and Java [12]

엔터프라이즈 웹 시스템에서 다양한 분석을 실시간적으로 표현 하고자 하는 요구 사항을 R 언어만을 이용하여 해결 하기는 어렵고 각각의 언어에서 제공하는 라이브러리를 이용하여 P Value 등을 추출할 때 같은 알고리즘을 이용하여 계산할 때에도 서로 다른 값들이 추출 되는 것을 확인 할 수 있었다. 본 논문에서는 Java 에서 RServe 를 이용하여 R Project 의 분석 모듈을 사용 하여 분석자가 분석하고 검증한 알고리즘을 실시간적으로 웹 시스템에 적용 할 수 있는 방안을 제안하고자 한다.

2. R 의 자료구조와 시스템 연계

Table 1 에서 보듯이 R 은 사용자의 편의성을 제공하기 위하여 다양한 자료구조를 제공한다. REngine 이 제공하는 REXP 클래스는 R 이 사용 하는 자료구조를 Java 에서도 사용이 가능하도록 기능을 제공하여 준다.

REXP 는 수치를 일반적으로 나타내는 수치형과 같은 데이터 타입의 경우에는 double 과 같은 단일형, double[]같은 일차원 배열, double[][]같은 이차원 배열등의 데이터 타입으로 반환하여 사용할 수가 있다. 그밖에 다른 데이터 타입의 경우에는 단일형이나 일차원 배열로 데이터를 가져올 수 있다.

Table 1. R data structure

자료구조	구성차원	데이터타입	복수 데이터 타입 적용여부
벡터(vector)	1 dimensional	Numerical / Character / Complex / Logic	impossible
행렬(matrix)	2 dimensional	Numerical / Character / Complex / Logic	impossible
데이터 프레임 (data frame)	2 dimensional	Numerical / Character / Complex / Logic	possible
배열(array)	2 dimensional or higher	Numerical / Character / Complex / Logic	impossible
요인(factor)	1 dimensional	Numerical / Character	impossible
시계열 (time series)	2 dimensional	Numerical / Character / Complex / Logic	impossible
리스트(lit)	2 dimensional or higher	Numeric / character / complex number / logic / function / expression / call	possible

하지만 R 에서는 일반적인 데이터 구조는 데이터 프레임과 같이 복수 데이터 타입을 제공하는 자료 구조일 것이다. 데이터 프레임과 같은 경우에는 REXP 만 이용하여 Java 에서 사용하기에는 무리가 있다. 이럴 경우 Java 에서는 REngine 에서 제공하는 RList 를 이용하여 Java 에서 데이터를 사용할 수 있게 지원하여 준다. 하지만 데이터 프레임과 같은 형식으로 사용 할 수 있으므로 데이터 프레임에 대한 기본 지식을 가져야 사용 할 수가 있을 것이다. Table2 에 R 의 데이터 프레임을 나타내었다.

Table 2. Data frame of R

	Kids	Age	Height	Stats
1	Lillian	12	152.3	CA
2	Jack	9	130.2	MA
3	Jill	13	158.3	CA
4	Jone	8	123.8	HI
5	Jillian	12	151.7	CA

데이터 프레임은 관찰치수가 같은 행 별로 있는 이루어져 있는 데이터 형이다. 행을 나타내는 필드명이 곧 변수명이라고 생각 할 수 있고 레코드의 값들은 변수에 할당된 배열구조 라고 생각 할 수 있다. 위 예제와 같이 Kids, Ages, Height, Stats 등의 리스트들이 나열된 형태의 자료구조를 데이터 프레임이라고 한다.

Age 의 레코드수가 5 개이라면 Height 의 레코드수도 5 개가 존재해야 한다. 각 리스트의 개수가 다르면 데이터 프레임은 생성될 수가 없다. Java 에서 데이터 프레임을 받아들인 RList 객체는 필드의 Index 또는 필드명에 따라 REXP 객체를 반환 해준다. Java 는 이를 이용해 데이터 프레임의 데이터를 사용 할 수가 있다.

Row 별로 데이터를 가져오면 좋겠지만 아직 Java 에서는 Row 에 대한 데이터만 가져오는 방식은 제공하고 있지 않다.

Rserve 에서 제공하는 Java client API 는 원격에 있는 R 에 접속하여 데이터를 송수신 하며 파일을 생성 및 읽어 들일 수 있다. Rserve 에서 기본적으로 사용하는 기본 API 중 본문에서 테스트로 사용하는 API 메소드 몇 가지만 알아보도록 하겠다. 자세한 API 는 http://rforge.net/org/doc/ 를 통하여 확인 할 수 있다.

```
RConnection c = new RConnection();
REXP x = c.eval("R.version.string");
System.out.println("R version : " + x.asString());
```

Figure 4. R connection and function call

로컬 호스트에 있는 R 에 접속을 하고 R 서버의 버전을 가져와서 출력하는 소스이다.

```
double[] dataX = { 10, 1, 2, 3, 4, 5, 6, 7, 8, 9 };
double[] dataY = { 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 };
c.assign("x", dataX);
c.assign("y", dataY);
```

Figure 5. Data transmission

RConnection 의 assign 메소드로 R 서버에 x 와 y 라는 변수에 데이터를 지정 하여 데이터를 전송하였다.

```
REXP x = c.eval("rnorm(100)");
double[] d = x.asDoubles();
.....
RList l = c.eval("lowess(x,y)").asList();

double[] lx = l.at("x").asDoubles();
String[] ly = l.at("y").asStrings();
....
RList l =
c.eval("{d=data.frame(₩"TestData₩",c(11:20));
lapply(d,as.character)}").asList();

int cols = l.size();
int rows = l.at(0).length();

String[][] s = new String[cols][];

for (int i = 0; i < cols; i++) {
      s[i] = l.at(i).asStrings();
}
```

Figure 6. Sample source using data object in R and Java

R 에서 rnorm() 함수를 이용하여 100 까지의 랜덤으로 생성된 숫자를 double[] 형으로 받았다. R 에서 lowess 를 통하여 처리된 데이터프레임 형태의 데이터를 받아서 double[] lx 와 String[] ly 로 받아 들여 Java 에서 사용할 수 있다. 위에서 예제로 적용한 방법과 자료구조등을 이용하여 R 의 자료구조 및 알고리즘을 사용할 수 있다.

3. R 과 Java 에서의 P-Value Test

Log Rank Test 는 약물 치료 및 수술에 대한 표준치료군과 새 치료군사이에서 생존율에 대한 생존 분석을 할 때에 가장 많이 사용하는 알고리즘중에 하나이다. Java 를 사용하여 개발한 웹 사이트에서 실시간으로 적용하고자 하고자 검증된 마땅한 라이브러리를 찾을 수가 없을 것이다. Apache 프로젝트중에 Apache Common math 에 대한 라이브러리가 있지만 Log Rank Test 알고리즘은 찾을 수가 없다. 그래서 본 연구에서는 javastat 라는 프로젝트의 라이브러리를 찾아 Log Rank Test 를 i5-6200 CPU 의 8G 메모리의 PC 에서 테스트 해 보았다.

```
double[] time1 = [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15];
double[] time2 = [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15];

double[] censor1 = [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0];
double[] censor2 = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1];

LogRankTest testclass1 = new LogRankTest(time1, censor1, time2, censor2);
double pValue = testclass1.pValue;
```

Figure 7. Log Rank Test using javastat in Java

해당 값을 출력한 결과는 P-Value 는 0.157 이었다. Figure 9 은 해당 P-Value 에 대한 서바이벌 차트이다.

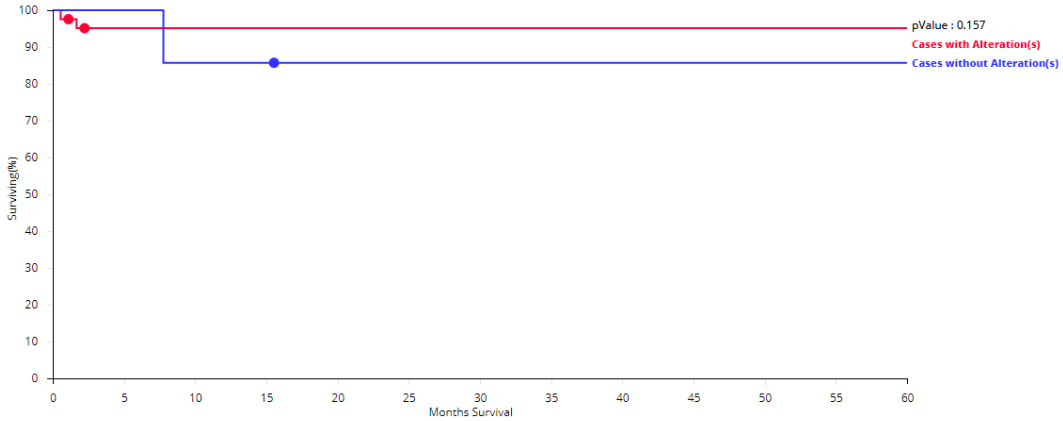

Figure 8. Log Rank Test Value using javastat in Java

그래서 R 에서 해당 값이 정확한 것인가를 Figure 10 과 같이 테스트 해보게 되었고 추출된 값은 javastat 와 다른 값인 0.564 가 추출되는 것을 확인 할 수 있었다.

```
> LR_time <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
> LR_status <- c(0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)
> LR_treatment <- c(rep(1,16), rep(2,16))
> survdiff(Surv(LR_time, LR_status) ~ LR_treatment)
Call:
survdiff(formula = Surv(LR_time, LR_status) ~ LR_treatment)

                N Observed Expected (O-E)^2/E (O-E)^2/V
LR_treatment=1 16        2      1.5     0.167     0.333
LR_treatment=2 16        1      1.5     0.167     0.333

 Chisq= 0.3  on 1 degrees of freedom, p= 0.564
```

Figure 9. Log Rank Test using R

J Log Rank Test 에 대한 알고리즘은 이미 R 에서 오래전에 개발되었고 이미 전세계적으로 사용되어 알고리즘에 대하여 검증되어 있기 때문에 R 의 알고리즘을 사용하는 것이 데이터에 대한 혼란이 없을 것이라 판단 되었다. 그래서 본 논문에서는 아래와 같이 RServe 를 이용하여 R 의 모듈을 이용하여 Log Rank Test 의 P-Value 를 추출 하였다.

```
RConnection c = new RConnection();
c.eval("library(survival)");

double[] LR_time = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
                    0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15];
double[] LR_status = [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                      0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1];

c.assign("LR_time", LR_time);
c.assign("LR_status", LR_status);
c.eval("LR_treatment <- c(rep(1,"+time1.length+"), rep(2,"+time2.length+"))");

c.eval("LR_diff <- survdiff(Surv(LR_time, LR_status) ~ LR_treatment)");

double pValue = c.eval("1 - pchisq(LR_diff$chisq, length(LR_diff$n)
                        - 1)").asDouble();
```

Figure 10. Log Rank Test using R in Java

Figure 12 은 추출된 P-Value 를 적용한 웹 사이트의 서바이벌 차트이다. R 로 연산된 P-Value 인 0,564 가 적용된 것을 확인 할 수 있다.

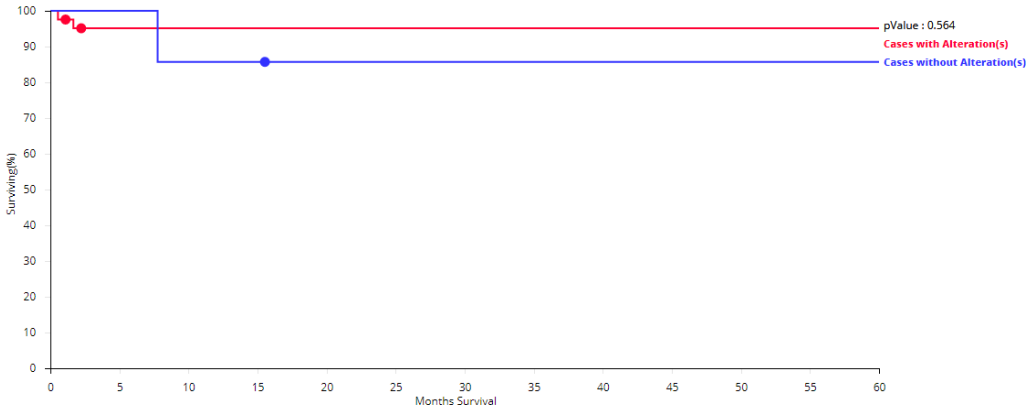

Figure 11. Log Rank Test using R in Java

4. 결론

최신의 통계 기술은 R 과 파이썬으로 좁혀져 있다고 할 수 있다. P Value 같은 경우에도 각각의 언어에서 제공하는 것을 손쉽게 사용하는 것을 확인 할 수 있는데 언어마다 알고리즘의 구현이 다르게 되어 있어서 값이 다르게 나오는 것을 확인 할 수 있다. Java 에서는 Log Rank Test 같은 알고리즘은 Java 애서 구현되어 있는 오픈 모듈이 거의 없는 것도 확인 할 수 있다. R 은 다른 언어 에서 제공하지 않은 다수의 분석 모듈이 이미 사용하기 쉽게 제공되어지고 있고 이미 검증이 되어있는 많은 라이브러리를 사용하고 있다. 그리고 분산 처리를 이용할 경우에 있어 대용량 분석은 파이썬보다는 R 에서 아직은 최적화 되어 있고 강점이 있다고 다수의 분석자들은 평가를 내리고 있다. Lab 에서의 분석자들이 공통 환경의 값을 나타내고자 할 때는 한가지 언어를 사용 하는 것이 좋지만 기존에 사용 했던

프로젝트가 있어서 그 마저도 쉽지가 않다. 하지만 R 의 중요 라이브러리를 다른 언어들에서 사용 할 수 있다면 많은 문제가 해결 될 것으로 보여진다. 본 논문에서는 그러한 문제를 해결하는 방법으로서 해당 내용을 기재한다.

5. 참고문헌

[1] Paul Gerrad, Radia M. Johnson, “Mastering Scientific Computing with R”, in Packt publishing, Birmingham, pp.19-74, 2016
[2] Pfeiffer, Andreas; Pia, Maria Grazia, “Data analysis with R in an experimental physics environment”, Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2013 IEEE, pp.1-2, 2013
[3] K. Sailaja Kumar, D. Evangelin Geetha, T. V. Sai Manoj, N Nagesh, “Identify the influential user in online social networks using R, Hadoop and Python”, 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), IEEE, pp.273-279, 2016
[4] Travis E. Oliphant, “Python for Scientific Computing”, Computing in Science & Engineering, vol. 9, pp. 10-20, 2007.
[5] Vignesh Prajapati, “‘Big Data’ Analytics with R and Hadoop”, Packt publishing, pp.35-62, 2013,
[6] Lee E. Edlefsen, Ph.D., “The Coming Revolution in Statistics”, Revolution Analytics, 2011, http://blog.revolutionanalytics.com/2011/03/index.html
[7] Simon Urbanek, “FastRWeb: Fast Interactive Web Framework for Data Mining Using R”, in Proceedings of the IASC 2008 World Congress, pp.1-6, 2008
[8] John M. Chambers , “Facets of R”, The R Journal: article published in 2009, vol 1, No. 1, pp.5-8, 2009