A Study of Social Media Data Management using R Programming

Kapil Sharma¹, Nitin Sharma²

¹,²Assistant Professor, Department of Computer Science & Engineering, Institute of Engineering & Technology, Alwar.

Abstract

Since a regularly expanding piece of the populace influences utilization of web-based social networking in their everyday to lives, web-based Social Media Data Management (SMDM) is being broke down in various orders. The online networking Data Management process includes four particular advances, collection, data discovery, preparation, and analysis. While there is a lot of literature on the difficulties and challenges including particular data analysis strategies, there scarcely exists inquire about on the phases of information revelation, accumulation, and planning. We propose a system for online social media data management is R. R is an open-source data analysis environment and programming language, allows user to conduct a number of tasks that are essential for the effective processing and analysis of more than 1000 TB of data. R includes different packages which are useful in analysis of data. The process of converting data into knowledge, insight and understanding is Data analysis, which is a critical part of statistics. For the effective processing and analysis of Social media Data, it allows users to conduct a number of tasks that are essential. Because R is a high level language, a function can have a deep hierarchy of operation. Although online social networking data processing might be proficient with different devices too, it is the point at which one stages on to the information examination that R truly stands interesting, outstanding to the huge amount of third-party algorithms and built-in statistical formulae available. 

General Terms: R Programming, Social Media Data Management.

Keywords: Social Media Data Management, R Programming

Introduction

Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as Social networking, business [3], entertainment [4], science, crisis management and politics. According to the survey conducted by the pew research center in September, 2016. Fig.1 shows that Facebook remains at the top among many social networking sites. The use of social networking sites like LinkedIn, Instagram, Facebook, and Twitter has seen a rapid growth in past year [1]. Though uploading of digital data to these sites is not a Difficult task but the storage, processing and managing of these data is difficult.
Case study by Kim, Choi, & Natali,[2] that collect a large data set during a specific time frame on a specific subject and analyses it quantitatively. Despite the variety of disciplines such projects can be found in, they have much in common. The steps necessary to gain useful information or even knowledge out of social media are often similar. Therefore, the field of “Social Media Data Management” aims to combine, extend, and adapt methods for the analysis of social media data. It has gained considerable attention and subsequently acceptance in academic research, but there is still a lack of comprehensive discussions of Social Media Data analytics, and of general models and approaches. In the data analytic process, several operators are needed to compose a workflow for analyzing the data. A suitable platform is required in order to facilitate the users and provide an efficient and accurate result. There are various platforms available for data analytics. Some of them are KNIME, WEKA, Rapid Miner, and R [6]. R is an open source and statistical programming language which was created after language S. R was developed by Ross Ihaka and Robert Gentleman at University of Auckland, New Zealand. R language provides an Integrated Development environment (IDE), RStudio. It also works as a tool for plotting, debugging and workspace management [7]. As R is a statistical language that works by formulating mathematical model, when applying on stochastic model, R provides better advantage [6].

The requirement of a tool for computational, statistics, visualization and data science has increased in both academic and industrial sector within the last decade. This has lifted the R programming language as one of the most important language. Across the globe, millions of statistician and data scientist use R to solve their most difficult problems in fields ranging from computational biology to quantitative marketing.

R has become one of the most popularized language for data management and a requisite tool for finance and analytics-driven companies such as Amazon, IBM, Google, Facebook etc.

A domain-specific, dynamically typed programming language having functional features. It should come as no surprise then that those who cherish R on the desktop want to bring it with them when they move to the cluster. Of course, R has a reputation for being slow, lacking scalability, and being inappropriate for HPC environments. However, the Programming with Big Data in R (pbdR) project [8] and other similar efforts from R developers are changing this perception. Like R itself, pbdR too was built for the convenience of the programmer with big data and large distributed computing resources.

Data scientists have a number of options to analyze data using statistical methods. One of the most convenient and powerful methods is to use the free R programming language. R is one of the best ways to create reproducible, high-quality analysis, since unlike a spreadsheet, R scripts can be audited and re-run easily [9]. The R language and its package repositories provide a wide range of statistical techniques, data manipulation and plotting, to the point that if a technique exists, it is probably implemented in an R package. R is almost as strong in its support for machine learning, although it may not be the first choice for deep neural networks, which require higher-performance computing than R currently delivers.

The rest of the paper is as follow: Section 2 describe the theoretical background about the social media data management and phases involve in social media analysis it also sharp us about the use of data. Section 3 describe the challenges of social media development it direct us about how to take care of investment and how to maximize the return, how to manage data, how to create strategy and finally how to choose interesting stuff for the user. Section 4 describe that how R play vital role in the field of data analysis and its growth over the years, it also cover
schemes for statically analysis using R. Finally, section 5 provides the conclusion of our review.

**Theoretical Background**

**Social Media data management**

Since the rise of social media usage in the last decade, people have been seeking to gain information from the crowd as an additional source to traditional media. Common too many of them. Because of the amount of the content produced daily and the number of active users on the platforms, organizations are motivated to understand which issues and trends evolve to recognize risks and chances in the communication and derive useful implications. In addition the amount of content, it is also relevant for organizations to understand who creates the content and which actors are the most significant drivers in the communication. Both businesses and non-profit organizations seek to collect the data produced by the crowd in order to gain insights into mass communication. The data is often collected with tools which communicate with the respective API of the social media platform, if one exists, and crawl the data. Among others, three domains in which social media is important and generates visible benefits are 1) in businesses, in 2) crisis communication, mainly in disaster management, and in 3) journalism and political communication.

Crisis communication research is an example of a field where social media data has had an impact. Social media is often used as a channel for emergency management agencies to inform people in an affected area on the current status of the respective crisis or how to behave [5].

Finally, social media platforms have been recognized in recent years as sources of data on political communication and for journalism and also uses for freedom of speech. People debate on current issues and further actions of politicians and discuss the consequences.

Overall, it can be stated that social media analytics is a highly complex process with different aspects regarding the respective application domain and the use of different methods. It is therefore useful and necessary to standardize this phenomenon to a process model, considering each step.

**Phases of social media data analysis**

**Preparation phase**: While in preparation phase the data is being planned, collected and selected for further processing or analysis. The preparation phase undergoes following steps viz data planning, data collection, feature generation, data selection. Here collection of data can be done either by the survey of data or using on hand data and thus the data is selected for second stage processing.

**Pre-Processing Phase**: The pre-processing phase deals mainly with syntax analysis and correction of data. After the preparation of data, the data is ready for initial processing. This phase includes the data cleaning process, filtering of data, and completion of data, data correction, standardization and transformation.

**Analysis Phase**: After the per-processing phase there is the analysis phase where the classification and grouping of data is done in a manner that data with similar patterns are brought together. The analysis phases undergoes the following steps viz visualization, correlation, regression, forecasting, classification and clustering.

**Decision making phase**: The post processing phase is the last stage of data analytics. The post processing phase includes the data interpretation, documentation of data and evaluation of data [4].

**Challenges of social media data management**

The continuing decline of organic reach

On Facebook and YouTube, the number of monthly users is over one billion, while Twitter, Instagram, and Pinterest are in the hundreds of millions. Each day, a staggering amount of content is produced by each of these channels. To keep users on the channel and engaged, social channels use algorithms to weed out junk, and only display the content that’s the most relevant and interesting to the user.

While consumers welcome a curated feed, the algorithm can be frustrating for marketers. Over the last couple of years, we’ve all watched our reach drop significantly, despite healthy, growing audiences.

In the early days of social media, when someone followed your page or account, they would see your posts — but this has changed. Because of the large amount of traffic on social channels, your posts can be suppressed, deprioritized, or simply fail to gain traction during a test push by the channel.
Solution?

How do you work with this? Right now, social channels are concern about keeping users on-channel and feeding them elevated quality, relevant content. It observe that Facebook and Twitter prefer marketer who post less often and post high quality content. Video is a priority, because a longer video will keep eyes on a channel, and live video offers an up-to-date view of the world. A smaller amount of posts with greater engagement and a mix of video will drive your reach. Instagram Stories is particularly potent here.

Creating an effective cross-channel strategy

If you’re working hard to grow your organic reach, you’re already optimizing your content by channel. Translating your content across many channels and making it work within the channel is a great skill for an SMM.

Multichannel strategies can be challenging because each channel has unique format and tone that direct how your content should fit. Additionally, your target audience has their own preferences and habits, and will often consult different channels while getting to know your company.

Having each of these channels work together in synergy creates a cohesive brand impression in response to your audience’s social behavior.

Solution?

Creating an effective cross-channel strategy involves an in-depth knowledge of the characteristics of each channel, and good organization.

Data management

Collecting data from social media or measuring your brand status all require sifting through social data. The information that you gather can nourish your business, and update your work as a social media marketer.

The challenges of data management that we face are in the quantity of data existing, as well as the dissemination of information to the right people within your business.

Solution?

Data management has grown from the tedious, manual process it used to be. If we can use a centralized, unified platform, we’ll be able to stream the barrage of information into functional blocks of data, and reveal opportunities you may have otherwise missed. Connecting your social media management tool with other data sources such as your CRM will help you collate and share key customer insights and interaction history.

Proving ROI

This is the never-ending challenge of social media managers in particular. How do you prove the value of your social campaign?

Solution?

Reporting on performance shows the effectiveness of your work, and keeps you on track week after week. By attaching a goal to your social media campaign, showing the investment, and then showing how the performance affected your bottom line, you clearly outline the return on investment in social media marketing.

R programming for data Analysis

R Growth

In 2016, IEEE had listed R at 6th position in the top 10 languages of 2016. In addition to this as the hunger for intensive data work increases, demand for tools like R for data-mining, processing and visualization will also increase.

Different Statically analysis using R Programming

Time Series Analysis

Time series is a series of data points in which each data point is linked with a timestamp. A simple example is the price of a stock in the stock market at different points of time on a given day. An added example is the amount of rainfall in a region at different months of the year. In R language we can uses several functions to create, manipulate and plot the time series data. The data for the time series is stored in an R object called time-series object.

The function that is use for creation of time series object is ts() function.

Syntax: timeseries.object.name<ts(data,start,end,frequency)

Data: is a vector or matrix containing the values used in the time series.

Start: specifies the start time for the first observation in

Figure: Programming language Ranking
Survival Analysis

Survival analysis is responsible to predict special event. It is also known as failure time analysis or analysis of time to death. For example predicting the number of days a person with cancer will survive or predicting the time when a mechanical system is going to fail.

The R package which named survival is used to carry out survival analysis. This package contains the function Surv() which takes the input data as a R formula and creates a survival object among the chosen variables for analysis. Then the survfit() function is used to create a plot for the analysis.

Examples:

Consider the annual rainfall details at a place starting from January 2012. We create an R time series object for a period of 12 months and plot it.

**Code:**

```r
rainfall <- c(799, 1174.8, 865.1, 1334.6, 635.4, 918.5, 685.5, 998.6, 784.2, 985, 882.8, 1071)

rainfall.timeseries <- ts(rainfall, start = c(2012, 1), frequency = 12)

print(rainfall.timeseries)

png(file = "rainfall.png")

plot(rainfall.timeseries)

dev.off()

Output:

![Figure Time Series of Rainfall time.](image)

**Survival Analysis**

Event indicates the status of occurrence of the expected event.

Formula is the relationship between the predictor variables.

Examples:

We will consider the data set marked “pbc” present in the survival packages installed above. It describes the survival data points about people affected with primary biliary cirrhosis (PBC) of the liver. Among the many columns present in the data set we are primarily concerned with the fields “time” and “status”. Time represents the number of days between registration of the patient and earlier of the event between the patient receiving a liver transplant or death of the patient.

**Code:**

```r
library("survival")

print(head(pbc))
```
Non linear least square

While we modeling real world data for regression analysis, we notice that it is infrequently the case that the equation of the model is a linear equation giving a linear graph. In majority of time, the equation of the model of real world data includes mathematical functions of higher degree like an exponent of 3 or a sin function. In such a state of affairs, the plot of the model gives a curve rather than a line. The ultimate aim of both linear and non-linear regression is to regulate the values of the model’s parameters to find the line or curve that comes closest to your data. On finding these values we will be able to estimate the response variable with good accuracy.

In Least Square regression, we set up a regression model in which the sum of the squares of the vertical distances of different points from the regression curve is minimized. We usually begin with a defined model and assume some values for the coefficients. Then we apply the nls() function of R to obtain the more accurate values along with the confidence intervals.

Syntax: nls(formula, data, start)

- Formula is a nonlinear model formula which includes variables and parameters.
- Data is a data frame used to evaluate the variables in the formula.
- Start is a named list.

Examples:

Let us consider a nonlinear model that consist assumption of initial values of its coefficients. Next we have to observe that which confidence intervals of these assumed values so that we can judge how well these values fir into the model.

So below equation suits for the purpose.

\[ a = b_1 x^2 + b_2 \]

Assume that the initial coefficients to be 1 and 3 and fit these values into nls() function.

Code:

```r
xvalues <- c(1.6, 2.1, 2, 2.23, 3.71, 3.25, 3.4, 3.86, 1.19, 2.21)
yvalues <- c(5.19, 7.43, 6.94, 8.11, 18.75, 14.88, 16.6, 19.12, 3.21, 7.58)

png(file = "nls.png")
plot(xvalues, yvalues)
model <- nls(yvalues ~ b1*xvalues^2+b2,start = list(b1 = 1,b2 = 3))
new.data <- data.frame(xvalues = seq(min(xvalues),max(xvalues),len = 100))
lines(new.data$xvalues,predict(model,newdata = new.data))
dev.off()
print(sum(resid(model)^2))
print(confint(model))
```

Output:

```
null device
1
[1] 1.081935
2.5% 97.5%
b1 1.137708 1.253135
b2 1.497364 2.496484
```
Conclusion

Web-based social networking investigation is as yet a generally new research territory; however it is of incredible enthusiasm to the data frameworks group and numerous specialists setting out on SMA venture in our field. In this paper we described an analytical tool for social media data analysis, named R. Our work employ on analysis of social media data using R. R provides us more than 5000 packages. These provide us facility to analyze social media data the more effectively and accurately. Numerous analytical tools are available however R has prominent advantage over them. Presently R has been used by various big companies and provides them a better performance.

References

8. Stefan Stieglitz, Milad Mirbabaie, Björn Ross, Christoph Neuberger[Social media analytics – Challenges in topic discovery, data collection, and data preparation] in International journal of information management.

Date of Submission: 2018-03-22
Date of Acceptance: 2018-04-17