DBLP Analysis

using NoSQL (MongoDB)
& R

By:
Muhammad Sohaib Ali

Introduction

- Performing Analytics on DBLP Data is done by different Techniques.
  - Statistical Analysis Techniques
  - Sentiment Analysis
  - Aggregations
  - Grouping and Joins
  - OLAP queries
  - No-SQL(MongoDB) projection

Details About Data and Tool

- In my Data Analytics and performing different OLAP queries are done by Following Tools
  - No-SQL
    - MongoDB, Studio 3T (DBMS for MongoDB)
  - Xml Reader for Big Data
    - EmEditor
  - MS-Excel
    - Conversion of XML to CSV
  - R Studio
    - R Language, Data Visualization

*Dataset is compressed and 10K records are used for analysis

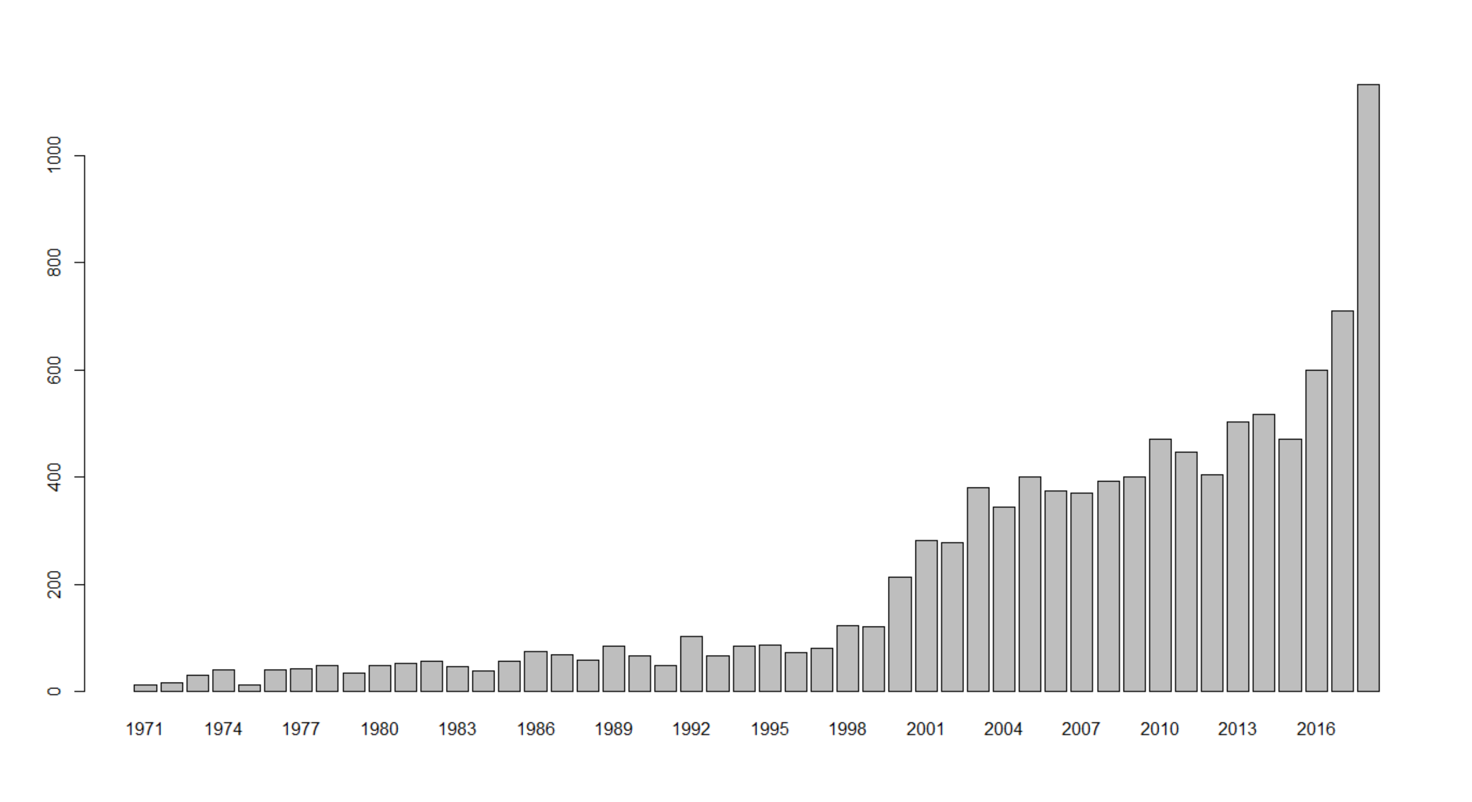
First Query

- In the very first analysis I was able to analyze the paper publishing behavior in last 3-4 decades.
- In 70s there were very number of rare papers published by the people.
- With the passage of time it started increasing.
- And Until 2016, the number of papers published is increased by a huge growth rate and still increasing.

Query:

- years= table(dataset$year)
- barplot(years)

Bar Graph of Years



Analysis

- As seen in the previous graph, We come to know number of publications are increasing every year with a huge rate.
- From 1971, to until 2016 there is almost a 1000 times increase in publication.
- Also shows interest of people are increasing in doing research and to publish their work in different publications.

Second Query

```
db.getCollection("dataset").aggregate([
{
    $group: { _id: { author: '$author' }, publtype: {
$addToSet: '$publtype'} }
},
{
    $unwind:"$publtype"
},
{
    $group: { _id: "$author", TotalAuthors : { $sum:1} }
}
]);
```

```
print(table(dataset$author))
```

Output List of Authors and their Number of Publications

Output

TotalAuthors

22406

Analysis

- Whenever we are doing analysis we do have a look at the data patterns to understand it to start the analysis most of the time.
- So in this query, we just came to know about the total number of authors and list of publications of each author.
- There are 2 parts of query
  - 1st is MongoDB Query which counts and shows number of Authors
  - 2nd is R language query showing list of authors and their number of publications

Third Query

```
all_years <- c(1971, 1974, 1977, 1980, 1983, 1986, 1989,
1992, 1995, 1998, 2001, 2004, 2007, 2010, 2013, 2016)
for (year in all_years)
{
data_year = grep(year, dataset$year)
print(table(dataset$year[data_year])) }
```

Output:

```
1971: 11     1974: 41     1977: 42     1980: 49     1983: 47
1986: 74     1989: 84     1992: 102    1995: 87     1998: 122
2001: 282    2004: 343    2007: 371    2010: 471    2013: 502
2016: 599
```

Analysis

- Analyzing Number of Publication each year. Tells us how many numbers are increased in publications per year.
- So that we can predict about number of increase in coming year according to previous increase rate.
- That is shown in the previous Query. Where a vector is made for total years and then a counter for number of publications is applied on publications according to each year.

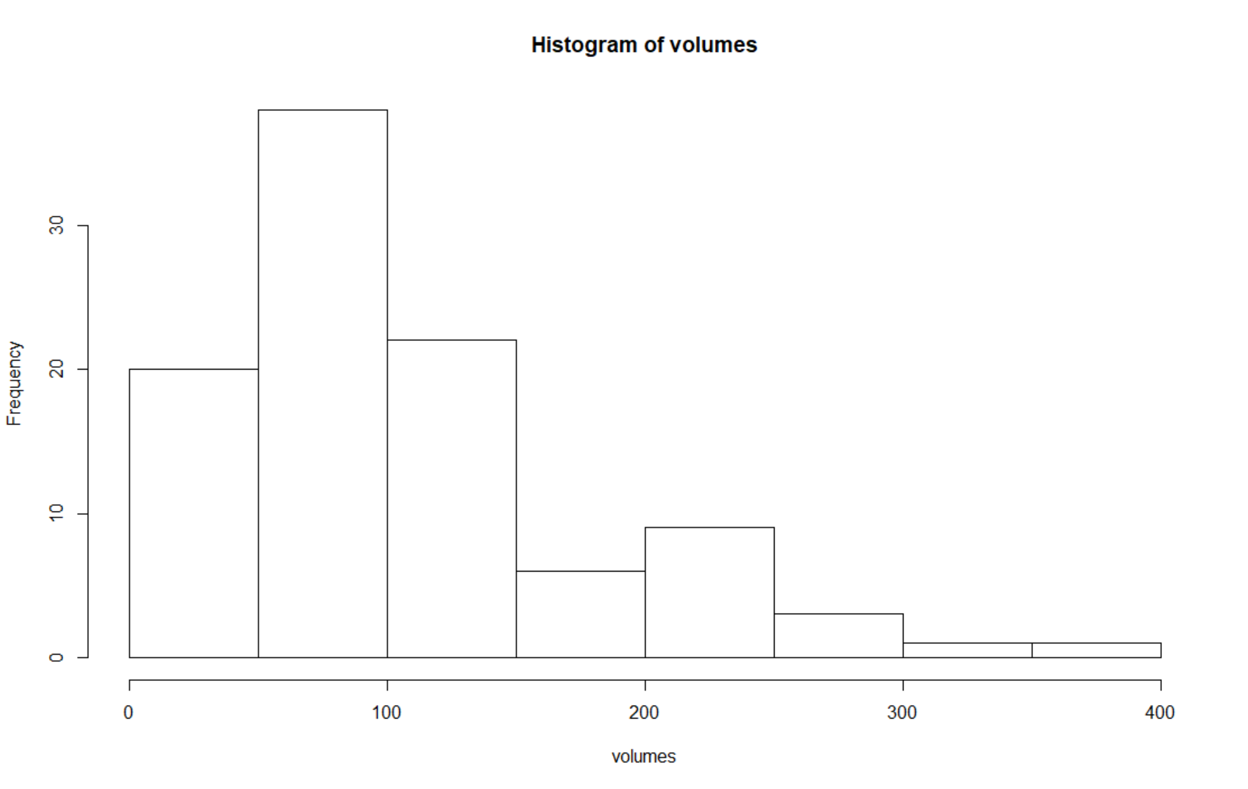
Fourth Query

- To view graphically and analyzing in the increase rate of volumes published within last 30 years.

Query:

```
volumes <- table(dataset$volume)
hist(volumes)
```

Histogram of Volumes of Last 30 years



Analysis

- Again for the purpose of prediction, we need to analyze the previous data.
- Query makes a histogram for the number of volumes published within last 30 years.
- Helps in predicting number of volumes going to be published within next few years.

Fifth Query

Queries	Outputs
➢ volumes <- table(dataset$volume) ➢ print(mean(volumes))	104.05
➢ volumes <- table(dataset$volume) ➢ print(mean(volumes))	88
➢ volumes <- table(dataset$volume) ➢ print(sum(volumes))	10405

Analysis

- There are different behaviors in data insights, we can understand them by applying different Statistical Analysis Techniques.
- Mean is found to be 104 which gives rough idea of almost 104 volumes are published by authors.
- Median give Mid value of Volumes as 88.
- Where as the mean is greater then the median which means that there are some outliers values which pulled the mean towards them.