NIPS Conference Analysis

Paper publishing patterns by various Organizations

Dinesh Daultani
Illinois State University

21 April, 2016

Abstract

NIPS is an international conference for machine learning and computational neuroscience. In the study, all published papers were gathered since the starting of the conference i.e., from 1987 to 2015. Subsequently, email ids of the authors has been extracted from all published papers. After the gathering, cleaning and the transformation of the data, I have performed pattern analysis of various Organization that have published the papers in the conference and the numbers of papers that have been published every year.

Introduction:

NIPS is an acronym for Neural Information Processing Systems (NIPS). NIPS is a conference on machine learning and computational neuroscience that is held every December. NIPS is also the topmost and most renowned conference for machine learning. NIPS conference was started in 1987 and has been continued since then. Other than machine learning and computational neuroscience, NIPS conference also represents papers in various categories such as computer vision, statistical linguistics, cognitive science, artificial intelligence, psychology and information theory.

Languages Used:

The papers have been collected using Python scripts from nips website and R Programming language has been used for data cleaning, data transformation and graph plotting. Python scripts has been submitted with the other material.

Data Cleaning and Tranformation:

Loading Libraries

```
suppressMessages(library(ggplot2))
suppressMessages(library(yaml))
suppressMessages(library(dplyr))
suppressMessages(library(stats))
suppressMessages(library(graphics))
suppressMessages(library(labeling))
```

Data

There are total 3 files used to create the graphs and they are as follows:
1. Author.csv - Containing 5 columns namely 'ID', 'Author_ID', 'Author_Name', 'Author_Email', 'Year' representing unique id for each record, paper' author's ID, paper' author's name, his/her email ID and paper published year respectively.
2. Years.csv - Containing two columns namely 'Total_papers' representing total papers published in a year and 'Year' representing the year in which papers are published.
3. Organization Dictionary - Containing two columns namely 'University_name' representing the names of organizations or universities and 'code' representing unique email address service provide URLs. For example: Email id is - xyz@ilstu.edu, then 'Univeristy_name' contains 'Illinois State University' and code contains 'ilstu.edu'.

```
# Authors file contains ID and Author_ID which is associated with the NIPS conference.
# It also contains Author_Name, Author_Email and paper published year.
Authors <- read.csv("Authors.csv", header = TRUE)
# Years file contains total numbers of papers published in each year.
year_dataset <- read.csv("Years.csv", header = TRUE)
# Dictionary that contains university names and their email codes
universities_dataset <- read.csv("OrganizationDictionary.csv", header = T)
```

Here's the structure of all input files

```
str(Authors)
```

```
## 'data.frame':    12065 obs. of  5 variables:
##  $ ID          : int  7956 2649 8299 8300 575 8419 8437 8437 8366 8367 ...
##  $ Author_ID   : int  5677 5677 5941 5941 5941 6019 6035 6035 5978 5978 ...
##  $ Author_Name : Factor w/ 5919 levels "Aaditya Ramdas",..: 3831 1323 745 73 4417 2132 5123 5123 556
##  $ Author_Email: Factor w/ 6617 levels "002@eng.cam.ac.uk",..: 4124 1272 742 75 689 2342 5888 1762 6
##  $ Year        : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
```

```
str(year_dataset)
```

```
## 'data.frame':    29 obs. of  2 variables:
##  $ Total_papers: int  403 411 360 368 306 292 262 250 217 204 ...
##  $ Year        : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
```

```
str(universities_dataset)
```

```
## 'data.frame':    225 obs. of  2 variables:
##  $ University_name: Factor w/ 216 levels "Aalto University",..: 127 79 19 44 43 214 34 40 89 161 ...
##  $ code           : Factor w/ 220 levels "aalto.fi","adelaide.edu",..: 138 90 25 50 49 217 42 47 107
```

Performing data transformation on Email ids.

I am going to use specifically Email ID column to identify which organization/university have published most papers in the conference. As every email id contains some unique url after the @ symbol representing the organization/university name. Hence I have splitted the string after the @ symbol in different column to do further manipulations in below chunk of code.

```
# Splitting the email ids from @ symbol to compare the email provider organizations.
Authors_new <- data.frame(do.call(rbind, strsplit(as.vector(Authors$Author_Email), split = "@")))
Authors[ , "T_Email"] <- Authors_new[2]
# Converting the Year column from int value to categorical values for easy grouping based on 'Year'.
Authors$Year <- as.factor(Authors$Year)
```

There are some raw data in transformed Email column after splitting the email ID's. Hence filtering those records from the dataset.

Now after transformation and cleaning of the email ID's of the authors, I am going to map each email unique code to organization/university name in below chunk of code.
Mapping the email urls to the Organizations names.

```
Authors[ , "Organization"] <- ""
# Loops to compare cleaned email id codes and university names
# i iterates for each record in authors dataset.
# j iterates for each record in university dictionary.
for (i in 1:12065)
{
    for(j in 1:225)
        {
        if(length(grep(universities_dataset[j,2], Authors[i,6])))
            {
            Authors[i,7] <- as.character(universities_dataset[j,1])
               break
            }
        }
}
# Filtering any blank email addressess if found
Cleaned_Authors_new <- filter(Authors, !grepl("^$", Organization))
```

Now after getting the names of the organization, taking unique entries based on year, author and organization name. And then counting papers published by each organization every year.

Grouping and counting Organization paper by year

```
# Taking only relevant columns
Cleaned_Authors_new <- Cleaned_Authors_new[,c(1,2,5,7)]
# Filtering unique values based on Author, Year and Organization columns.
Cleaned_Authors_new1 <- unique( Cleaned_Authors_new[ , c(2,3,4) ] )
# Counting total papers published by each Organizations every year
Counter_Organization <- count_(Cleaned_Authors_new1, c('Year', 'Organization'))
# Here's the structure of Count dataset.
str(Counter_Organization)
```

```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  1671 obs. of  3 variables:
##  $ Year        : Factor w/ 29 levels "1987","1988",..: 2 2 2 2 2 2 2 3 3 3 ...
##  $ Organization: chr  "Arizona State University" "Brown University" "California Institute of Technol
##  $ n           : int  1 1 1 2 1 1 1 1 2 1 ...
##  - attr(*, "vars")=List of 1
##   ..$ : symbol Year
##  - attr(*, "drop")= logi TRUE
```

Taking sum for total papers published by each organization in the whole span of time in the NIPS conference.
And then just taking top 10 organization/university names to plot the graphs in further step.

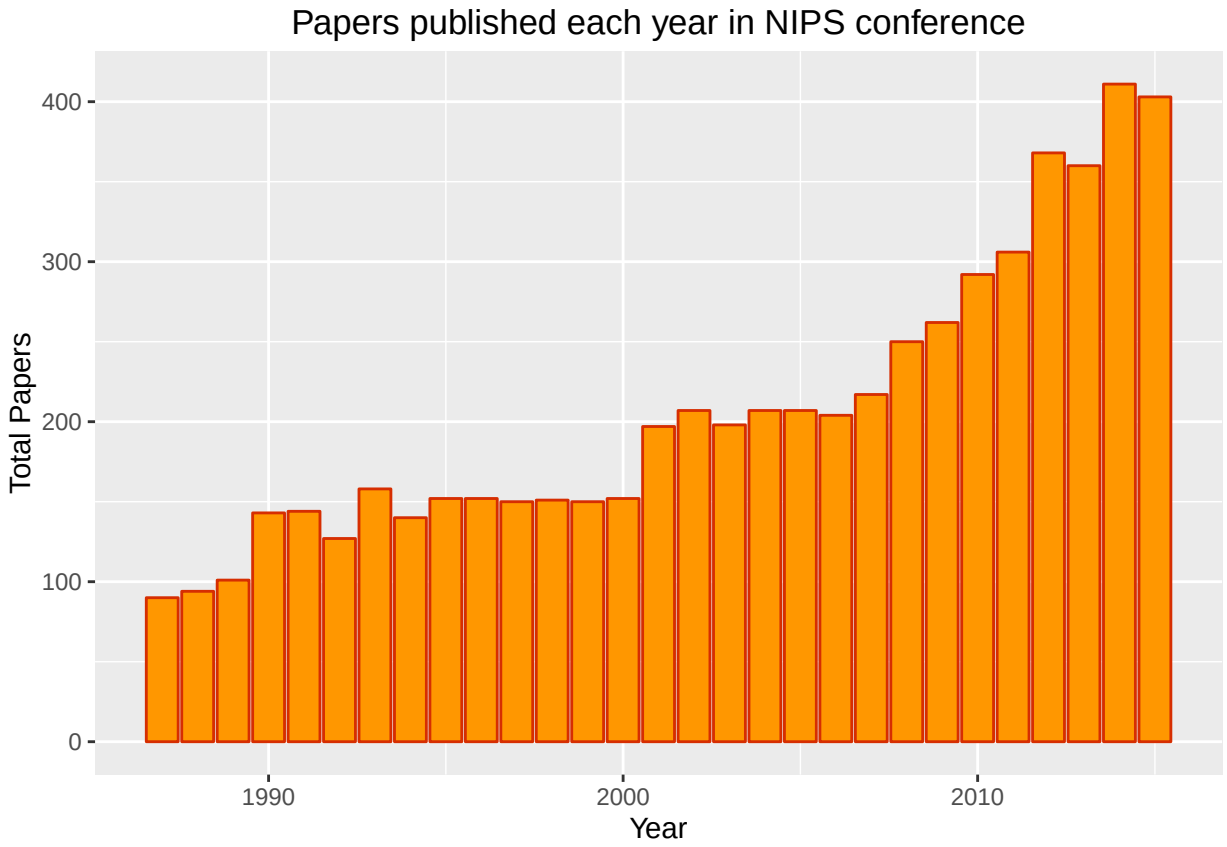
Performing final transformation in datasets.

```
## max papers by an organization
# Summing the records for each organization for the whole span of time.
Organization_Total_papers <- aggregate(Counter_Organization$n, by=list(Category=Counter_Organization$Org
# Changing the names of columns from default values
colnames(Organization_Total_papers) <- c('Organization','n')
# Taking top 10 organization from all the years
Max_papers <- Organization_Total_papers %>% top_n(10, n)
# Ordering the records in decreasing form.
Max_papers <- Max_papers[order(Max_papers$n,decreasing = T),]
# Making rownames empty just for removing random values in row numbers.
rownames(Max_papers) <- NULL
# Changing the name of the count column to Total_Papers
colnames(Max_papers)[2] <- 'Total_Papers'
```

Results

Graph plotting for papers published per year

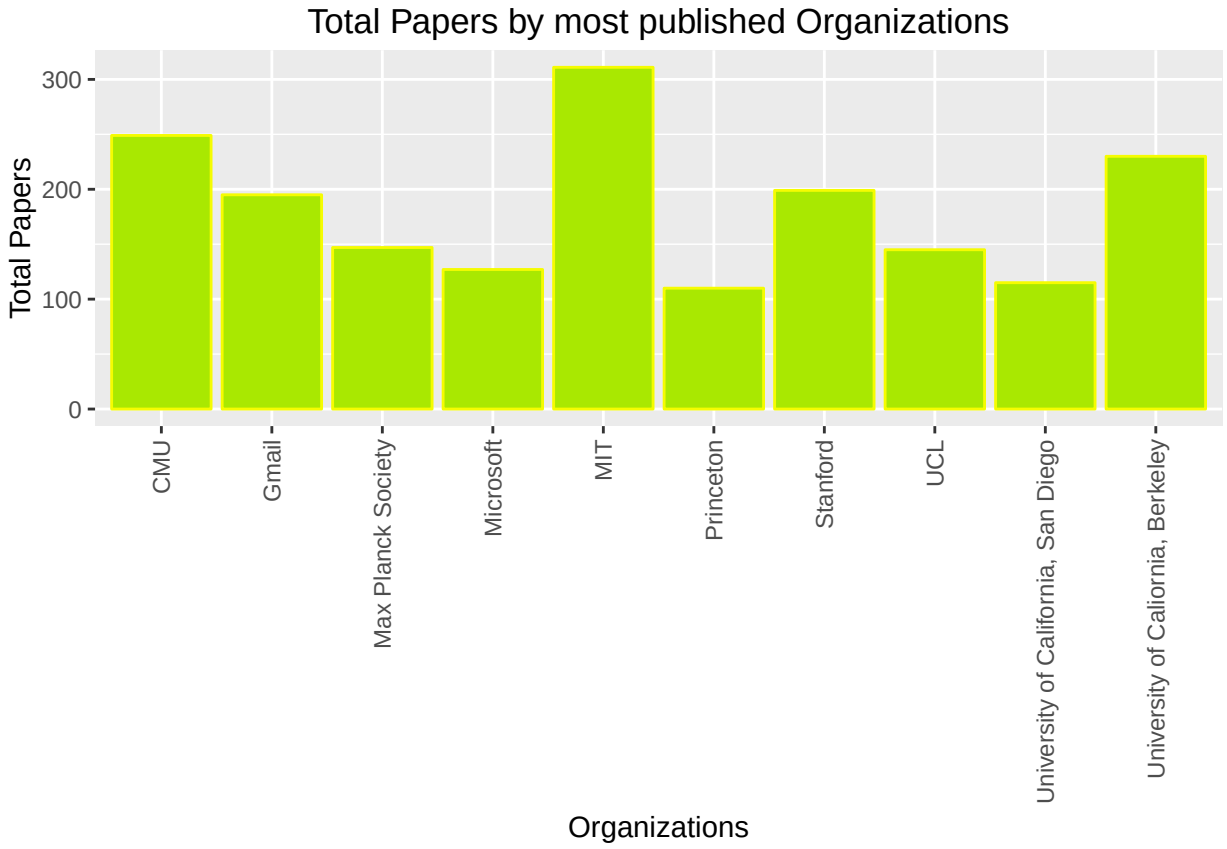
Graph is plotted between Year and Total papers published each year.
* Note: There is no need to clean the dataset "year_dataset" which contains Papers published each year because it is perfectly collected through the python scripts.



Analysis of the graph

Above graph shows how number of papers have increased over time. Specifically after year 2006-2007 the papers published each year were almost kind of exponential in nature, while before that, only in year 1990 and 2001 there was a substantial increase in number of papers publsihed.

Graph representing Organizations that published most number of papers in NIPS Conference.



Analysis of the graph

Above graph shows how some of the organizations/university are doing persistence research in machine learning and artificial intelligence over years. Also we can interpret from the above graph that the unievrsities such as “MIT”, “CMU”, “University of california, Berkeley” and “Stanford” are the most contributing universities in the field of machine learning, artificial intelligence and reltaed fields.

Limitations:

There are two limitations that were found during the study. They are as follows:
1. Conference' beginning years email id's were not properly gathered due to different format of pdfs. Hence some year email ids used to map the paper to the organization were not been proper.
2. Also the Organization/University dictionary used for mapping the names of Organizations to the appropriate email ID's contain just around 230 , although the organizations that published papers in the conference were much more than 230. Hence there were around 10% of records that were being filtered due to that.

References:

1. https://nips.cc
2. http://academic.research.microsoft.com/RankList?entitytype=3&topDomainID=2&subDomainID=6
3. https://github.com/benhamner/nips-2015-papers