prov-template time analysis

Luc Moreau

Nov 17, 2016

Loading data

```
library(reshape)

setwd("/Users/lavm/luc-git/papers/prov-template/data/")

#pdf("outputs/time.pdf")

smart_w <- read.table("smartshare/archive_28483/outputs/time.csv", sep=",")
smart_w2 <- smart_w[order(smart_w$V3),]
smart_w2$count <- 1:nrow(smart_w2)
smart_w2$app <- "smart"

smart_l <- read.table("smartshare/archive_28483/outputs/bindings2_output.txt", sep=",")
smart_l$V1 <- sub("normalized/bindings2/","bindings/",smart_l$V1)
smart_l$V1 <- sub(".json","",smart_l$V1)
names(smart_l)=c("V1","size")

smart_z<-merge(smart_w2,smart_l)

food_w <- read.table("foodprovenance/archive_20160303/outputs/time.csv", sep=",")
food_w2 <- food_w[order(food_w$V3),]
food_w2$count <- 1:nrow(food_w2)
food_w2$app <- "food"

food_l <- read.table("foodprovenance/archive_20160303/outputs/bindings2_output.txt", sep=",")
food_l$V1 <- sub("normalized/bindings2/","bindings/",food_l$V1)
food_l$V1 <- sub(".json","",food_l$V1)
names(food_l)=c("V1","size")

food_z<-merge(food_w2,food_l)


ebook_w <- read.table("ebook/20160308_big/outputs/time.csv", sep=",")
ebook_w2 <- ebook_w[order(ebook_w$V3),]
ebook_w2$count <- 1:nrow(ebook_w2)
ebook_w2$app <- "ebook"

ebook_l <- read.table("ebook/20160308_big/outputs/bindings2_output.txt", sep=",")
ebook_l$V1 <- sub("normalized/bindings2/","bindings/",ebook_l$V1)
ebook_l$V1 <- sub(".json","",ebook_l$V1)
names(ebook_l)=c("V1","size")

ebook_z<-merge(ebook_w2,ebook_l)
```

```
picaso_w <- read.table("picaso/20160211-reexpanded/outputs/time.csv", sep=",")
picaso_w2 <- picaso_w[order(picaso_w$V3),]
picaso_w2$count <- 1:nrow(picaso_w2)
picaso_w2$app <- "picaso"

picaso_l <- read.table("picaso/20160211-reexpanded/outputs/bindings2_output.txt", sep=",")
picaso_l$V1 <- sub("normalized/bindings2/","bindings/",picaso_l$V1)
picaso_l$V1 <- sub(".json","",picaso_l$V1)
names(picaso_l)=c("V1","size")

picaso_z<-merge(picaso_w2,picaso_l)




################################################################################
###
###

xx<-rbind(smart_z,
          food_z,
          ebook_z,
          picaso_z)



xx$norm <- xx$V3 / xx$size * 1000
yy <- xx[order(xx$V1),]


################################################################################
# box plot
```

Box Plot

```
mydots=c(1,2,3,4,5,6,7,8)

applications=c(1,2,3,4)
names(applications)=c("smart","food","ebook","picaso")

colors=c("red","blue","green4","brown")

tmpl <- unique(yy$V2)

print(length(tmpl))
```

```
## [1] 29
```

```
#pretty_templates <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,"tmpl_28
pretty_templates <- c(1:length(tmpl))
```

```
print(tmpl)
```

```
##  [1]  templates/template_35836      templates/template_35840
##  [3]  templates/template_3          templates/template_2
##  [5]  templates/template_25878      templates/template_4
##  [7]  templates/template_35839      templates/template_35844
##  [9]  templates/template_35838      templates/template_35843
## [11]  templates/template_9          templates/template_6
## [13]  templates/foodspec            templates/invoiceitems
## [15]  templates/analysis            templates/template_block_run
## [17]  templates/3                   templates/1
## [19]  templates/2                   templates/12
## [21]  templates/10                  templates/6
## [23]  templates/4                   templates/13
## [25]  templates/11                  templates/7
## [27]  templates/8                   templates/9
## [29]  templates/5
## 29 Levels:  templates/template_2 ...  templates/9
```

```
pretty_templates <- c(
"Send_Request",       #"template_35836",
"Receive_Request",    #"template_35840"   ,
"Change_Page",        #"template_3",
"Login",              #"template_2"       ,
"Receive_API_Call",   #"template_25878",
"Use_Response",       #"template_4"       ,
"Receive_Request",    #"template_35839",
"Composition",        #"template_35844"   ,
"Negotiation_Type_1", #"template_35838",
"Negotiation_Type_2", #"template_35843"   ,
"Gen_Reputation", #"template_9" ,
"Init_Gen_Reputation", #"template_6"      ,
"foodspec" ,
"invoiceitems"    ,
"analysis" ,
"block_run",
"Conference_Session",       #"templates/3",
"Attribution_v1",       #"templates/1",
"Citation",     #"templates/2",
"Attribution_v2",       #"templates/12",
"Derivation_1_n",       #"templates/10",
"Work_Element",     #"templates/6",
"Dataset_Usage",        #"templates/4",
"Project",      #"templates/13",
"Derivation_n_1",       #"templates/11",
"Tweet",        #"templates/7",
"Presentation_v1",      #"templates/8",
"Presentation_v2",      #"templates/9",
"Derived_Material"      #"templates/5"
)



names(pretty_templates)=tmpl
```

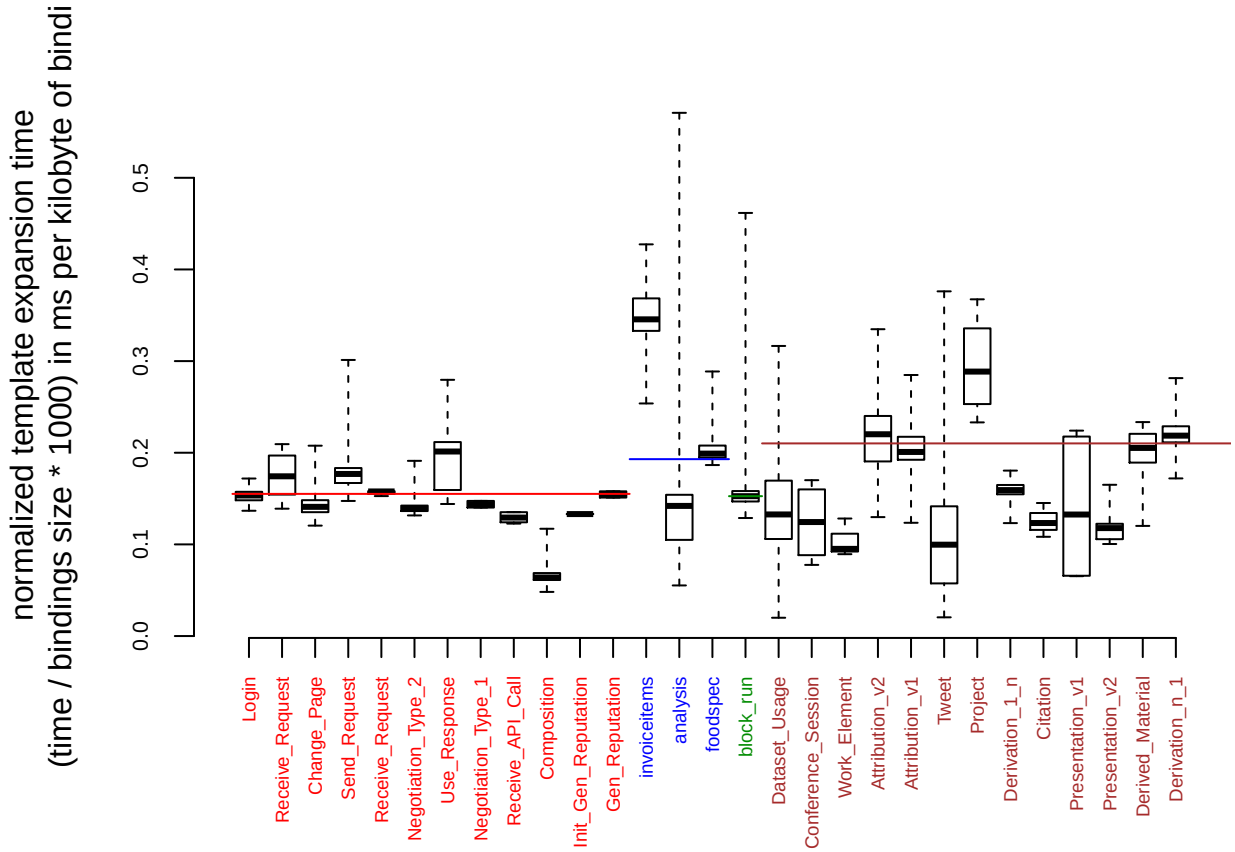
```
smart_median =median(smart_z$V3 / smart_z$size * 1000)
food_median  =median(food_z$V3  / food_z$size * 1000)
ebook_median =median(ebook_z$V3 / ebook_z$size * 1000)
picaso_median=median(picaso_z$V3 / picaso_z$size * 1000)
total_median=median(yy$V3 / yy$size * 1000)

print(food_median)
```

```
## [1] 0.1930065
```

```
par(mar = c(6,5,2.3,0.5))


boxplot(yy$norm ~ yy$V2, data=tmpl, axes=FALSE, range=0, ylab="normalized template expansion time\n (ti

# Make y axis
axis(2, cex.axis=0.7)

# Make x axis
axis(1, at=1:length(tmpl), labels=FALSE, cex.axis=0.7, las=3)

# Labels on x axis
mtext(text=pretty_templates[tmpl], side=1,at=1:length(tmpl),adj=1,col=colors[c(1,1,1,1,1,1,1,1,1,1,1,1,1
```

```
## Warning in mtext(text = pretty_templates[tmpl], side = 1, at =
## 1:length(tmpl), : "pos" is not a graphical parameter
```

```
# add a legend
legend(1,4, names(applications) , cex=0.7, col=colors[applications],lty=c(1,1),title="Applications")  #


segments(x0=0.5, x1=12.5, y0=smart_median, y1=smart_median, col=colors[1])
segments(x0=12.5,x1=15.5, y0=food_median,  y1=food_median,  col=colors[2])
segments(x0=15.5,x1=16.5, y0=ebook_median, y1=ebook_median, col=colors[3])
segments(x0=16.5,x1=31.5, y0=picaso_median,y1=picaso_median,col=colors[4])
```



Summary data and Correlation

```
print(names(pretty_templates))
```

```
##  [1] " templates/template_35836"     " templates/template_35840"
##  [3] " templates/template_3"         " templates/template_2"
##  [5] " templates/template_25878"     " templates/template_4"
##  [7] " templates/template_35839"     " templates/template_35844"
##  [9] " templates/template_35838"     " templates/template_35843"
## [11] " templates/template_9"         " templates/template_6"
## [13] " templates/foodspec"           " templates/invoiceitems"
## [15] " templates/analysis"           " templates/template_block_run"
## [17] " templates/3"                  " templates/1"
## [19] " templates/2"                  " templates/12"
## [21] " templates/10"                 " templates/6"
## [23] " templates/4"                  " templates/13"
## [25] " templates/11"                 " templates/7"
## [27] " templates/8"                  " templates/9"
## [29] " templates/5"
```

```
smart_mean =mean(smart_z$V3)
food_mean  =mean(food_z$V3)
ebook_mean =mean(ebook_z$V3)
picaso_mean=mean(picaso_z$V3)
total_mean =mean(yy$V3)
```

```
smart_median2 =median(smart_z$V3)
food_median2  =median(food_z$V3)
ebook_median2 =median(ebook_z$V3)
picaso_median2=median(picaso_z$V3)
total_median2 =median(yy$V3)

smart_sd =sd(smart_z$V3)
food_sd  =sd(food_z$V3 )
ebook_sd =sd(ebook_z$V3)
picaso_sd=sd(picaso_z$V3)
total_sd =sd(yy$V3)

smart_bindings_mean =mean(smart_z$size)
food_bindings_mean  =mean(food_z$size)
ebook_bindings_mean =mean(ebook_z$size)
picaso_bindings_mean=mean(picaso_z$size)
total_bindings_mean =mean(yy$size)


summary = c(1,2,3,4)
names(summary)= names(applications)
summary["smart"]=smart_mean
summary["food"]=food_mean
summary["ebook"]=ebook_mean
summary["picaso"]=picaso_mean
summary["total"]=total_mean


summary = rbind(summary,c(smart_sd,food_sd,ebook_sd,picaso_sd,total_sd))
summary = rbind(summary,c(smart_median2,food_median2,ebook_median2,picaso_median2,total_median2))
summary = rbind(summary,c(smart_bindings_mean,food_bindings_mean,ebook_bindings_mean,picaso_bindings_mea
summary = rbind(summary,c(smart_median,food_median,ebook_median,picaso_median,total_median))



summary <- t(summary)
colnames(summary) <- c("mean","sd", "median", "bindings sets", "norm. median")

pearson=cor.test(yy$V3, yy$size)
spearman=cor.test(yy$V3, yy$size,method = "spearman")
```

```
## Warning in cor.test.default(yy$V3, yy$size, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
print(round(summary,digit=3))
```

```
##         mean    sd median bindings sets norm. median
## smart  0.181 0.119  0.147      1106.346        0.155
## food   0.605 0.379  0.451      3367.530        0.193
## ebook  0.174 0.067  0.160      1126.923        0.153
## picaso 0.174 0.100  0.165       875.955        0.210
## total  0.234 0.225  0.167      1282.232        0.183
```

```r
print(pretty_templates)
```

```
##      templates/template_35836      templates/template_35840
##                "Send_Request"             "Receive_Request"
##          templates/template_3          templates/template_2
##                 "Change_Page"                        "Login"
##      templates/template_25878          templates/template_4
##            "Receive_API_Call"                "Use_Response"
##      templates/template_35839      templates/template_35844
##             "Receive_Request"                 "Composition"
##      templates/template_35838      templates/template_35843
##          "Negotiation_Type_1"          "Negotiation_Type_2"
##          templates/template_9          templates/template_6
##              "Gen_Reputation"         "Init_Gen_Reputation"
##            templates/foodspec        templates/invoiceitems
##                    "foodspec"                "invoiceitems"
##            templates/analysis  templates/template_block_run
##                    "analysis"                    "block_run"
##                   templates/3                   templates/1
##          "Conference_Session"               "Attribution_v1"
##                   templates/2                  templates/12
##                    "Citation"               "Attribution_v2"
##                  templates/10                   templates/6
##               "Derivation_1_n"                 "Work_Element"
##                   templates/4                   templates/13
##                "Dataset_Usage"                      "Project"
##                   templates/11                  templates/7
##               "Derivation_n_1"                      "Tweet"
##                   templates/8                   templates/9
##              "Presentation_v1"             "Presentation_v2"
##                   templates/5
##             "Derived_Material"
```

```r
print(pearson)
```

```
##
##  Pearson's product-moment correlation
##
## data:  yy$V3 and yy$size
## t = 167.46, df = 7523, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8830801 0.8926429
## sample estimates:
##       cor
## 0.8879575
```

```r
print(spearman)
```

```
##
##  Spearman's rank correlation rho
##
## data:  yy$V3 and yy$size
## S = 1.1326e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##       rho
## 0.8405236
```

Scatter plot

```
plot(yy$V3, yy$size)
```