prov-template size analysis

Luc Moreau

Nov 17, 2016

Loading data

```
library(reshape)

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

#pdf("outputs/box-template.pdf")


smart_w <- read.table("smartshare/archive_28483/outputs/toscatter.csv", sep=",")
smart_w2 <- smart_w[order(smart_w$V4),]
smart_w2$V7 <- 1:nrow(smart_w2)
smart_w2$V8 <- 1
smart_w2$V9 <- "smart"



smart_ww <- read.table("smartshare/archive_28483/outputs/toscatter2.csv", sep=",")
smart_ww2 <- smart_ww[order(smart_ww$V4),]
smart_ww2$V7 <- 1:nrow(smart_ww2)
smart_ww2$V8 <- 2
smart_ww2$V9 <- "smart"


food_w <- read.table("foodprovenance/archive_20160303/outputs/toscatter.csv", sep=",")
food_w2 <- food_w[order(food_w$V4),]
food_w2$V7 <- 1:nrow(food_w2)
food_w2$V8 <- 1
food_w2$V9 <- "food"



food_ww <- read.table("foodprovenance/archive_20160303/outputs/toscatter2.csv", sep=",")
food_ww2 <- food_ww[order(food_ww$V4),]
food_ww2$V7 <- 1:nrow(food_ww2)
food_ww2$V8 <- 2
food_ww2$V9 <- "food"


ebook_w <- read.table("ebook/20160308_big/outputs/toscatter.csv", sep=",")
ebook_w2 <- ebook_w[order(ebook_w$V4),]
ebook_w2$V7 <- 1:nrow(ebook_w2)
ebook_w2$V8 <- 1
ebook_w2$V9 <- "ebook"
```

```
ebook_ww <- read.table("ebook/20160308_big/outputs/toscatter2.csv", sep=",")
ebook_ww2 <- ebook_ww[order(ebook_ww$V4),]
ebook_ww2$V7 <- 1:nrow(ebook_ww2)
ebook_ww2$V8 <- 2
ebook_ww2$V9 <- "ebook"


picaso_w <- read.table("picaso/20160211-reexpanded/outputs/toscatter.csv", sep=",")
picaso_w2 <- picaso_w[order(picaso_w$V4),]
picaso_w2$V7 <- 1:nrow(picaso_w2)
picaso_w2$V8 <- 1
picaso_w2$V9 <- "picaso"

picaso_w2$V1 <- as.character(picaso_w2$V1)
picaso_w2$V3 <- as.character(picaso_w2$V3)
picaso_w2$V5 <- as.character(picaso_w2$V5)


picaso_ww <- read.table("picaso/20160211-reexpanded/outputs/toscatter2.csv", sep=",")
picaso_ww2 <- picaso_ww[order(picaso_ww$V4),]
picaso_ww2$V7 <- 1:nrow(picaso_ww2)
picaso_ww2$V8 <- 2
picaso_ww2$V9 <- "picaso"

picaso_ww2$V1 <- as.character(picaso_ww2$V1)
picaso_ww2$V3 <- as.character(picaso_ww2$V3)
picaso_ww2$V5 <- as.character(picaso_ww2$V5)



picaso_names <- read.table("picaso/20160211-reexpanded/raw/template-names.txt",sep=",")
names(picaso_names)<-c("name","pretty")

for_picaso_name <- function(n) {
            return(picaso_names[picaso_names$name == n, "pretty"])
          }
```

Compaction Ratio with Repect to Bindings Size

```
#########################################################################
###
###  Version 2 (bindings)
###

xx<-rbind(smart_ww2,
          food_ww2,
          ebook_ww2,
          picaso_ww2)
```

```
xx$V10 <- xx$V4 / xx$V6
xx$V11 <- xx$V2 / xx$V6

yy <- xx[order(xx$V1),]
#yy$V11 <- 1:nrow(yy)


smart_mean =mean(smart_ww2$V4 / smart_ww2$V6)
food_mean  =mean(food_ww2$V4  / food_ww2$V6)
ebook_mean =mean(ebook_ww2$V4 / ebook_ww2$V6)
picaso_mean=mean(picaso_ww2$V4/ picaso_ww2$V6)
total_mean =mean(yy$V10)

smart_sd =sd(smart_ww2$V4 / smart_ww2$V6)
food_sd  =sd(food_ww2$V4  / food_ww2$V6)
ebook_sd =sd(ebook_ww2$V4 / ebook_ww2$V6)
picaso_sd=sd(picaso_ww2$V4/ picaso_ww2$V6)
total_sd =sd(yy$V10)


smart_median =median(smart_ww2$V4 / smart_ww2$V6)
food_median  =median(food_ww2$V4  / food_ww2$V6)
ebook_median =median(ebook_ww2$V4 / ebook_ww2$V6)
picaso_median=median(picaso_ww2$V4/ picaso_ww2$V6)
total_median =median(yy$V10)
```

Ratio Between Templates Size and Expanded Provenance

```
smart_mean2 =mean(smart_ww2$V2 / smart_ww2$V6)
food_mean2  =mean(food_ww2$V2  / food_ww2$V6)
ebook_mean2 =mean(ebook_ww2$V2 / ebook_ww2$V6)
picaso_mean2=mean(picaso_ww2$V2/ picaso_ww2$V6)
total_mean2 =mean(yy$V11)

smart_sd2 =sd(smart_ww2$V2 / smart_ww2$V6)
food_sd2  =sd(food_ww2$V2  / food_ww2$V6)
ebook_sd2 =sd(ebook_ww2$V2 / ebook_ww2$V6)
picaso_sd2=sd(picaso_ww2$V2/ picaso_ww2$V6)
total_sd2 =sd(yy$V11)


smart_median2 =median(smart_ww2$V2 / smart_ww2$V6)
food_median2  =median(food_ww2$V2  / food_ww2$V6)
ebook_median2 =median(ebook_ww2$V2 / ebook_ww2$V6)
picaso_median2=median(picaso_ww2$V2/ picaso_ww2$V6)
total_median2 =median(yy$V11)

#############################################################################
# box plot (bindings)
```

Box Plot (compaction ratio bindings/expanded provenance)

```
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$V1)

print(length(tmpl))
```

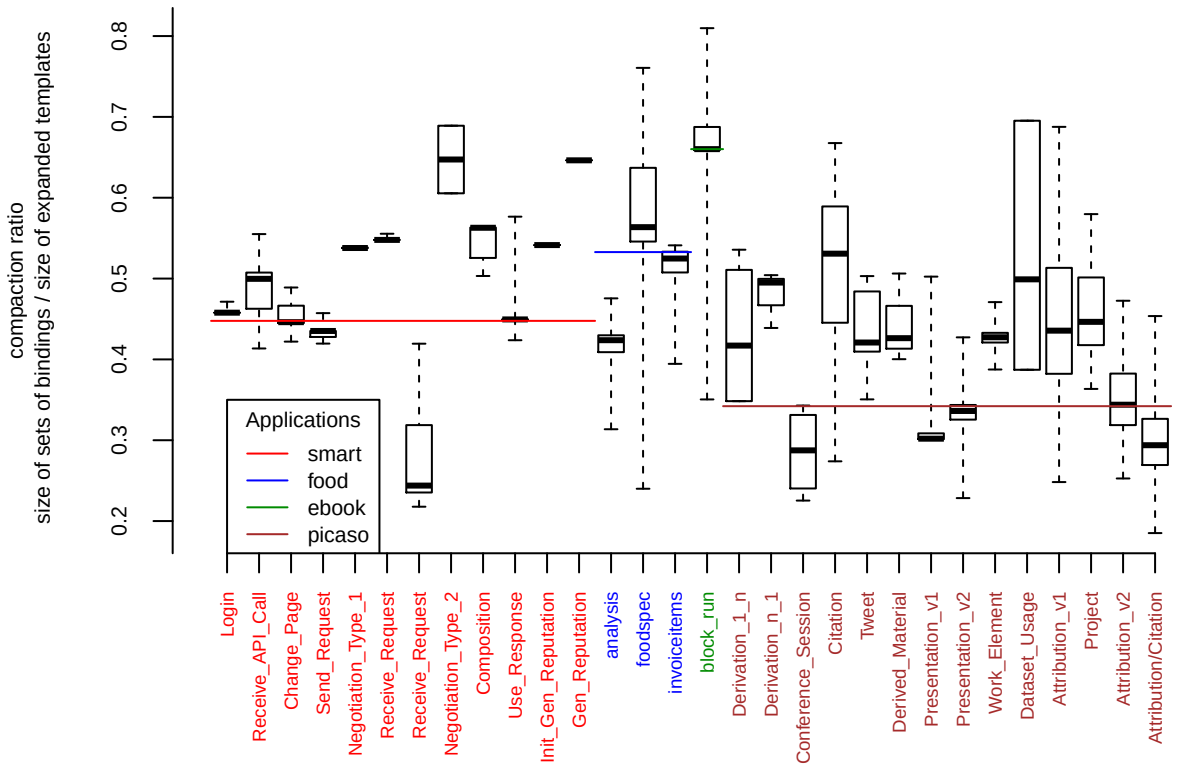
```
## [1] 30
```

```
#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))


pretty_templates <- c(
"Login",               #"template_2",
"Receive_API_Call",    #"template_25878",
"Change_Page",         #"template_3",
"Send_Request",        #"template_35836",
"Negotiation_Type_1", #"template_35838",
"Receive_Request",     #"template_35839",
"Receive_Request",     #"template_35840",
"Negotiation_Type_2", #"template_35843",
"Composition",         #"template_35844",
"Use_Response",        #"template_4",
"Init_Gen_Reputation", #"template_6",
"Gen_Reputation", #"template_9",
"analysis",
"foodspec",
"invoiceitems",
"block_run",
"Derivation_1_n",        #"tmpl_10",
"Derivation_n_1",        #"tmpl_11",
"Conference_Session",        #"tmpl_3",
"Citation",      #"tmpl_2",
"Tweet",         #"tmpl_7",
"Derived_Material",      #"tmpl_5",
"Presentation_v1",       #"tmpl_8",
"Presentation_v2",       #"tmpl_9",
"Work_Element",      #"tmpl_6",
"Dataset_Usage",         #"tmpl_4",
"Attribution_v1",        #"tmpl_1",
"Project",       #"tmpl_13",
"Attribution_v2",        #"tmpl_12",
"Attribution/Citation"       #"tmpl_1+2"
)



names(pretty_templates)=tmpl
```

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


boxplot(yy$V10 ~ yy$V1, data=tmpl, axes=FALSE, range=0, ylab="compaction ratio\n size of sets of bindin

# Make y axis
axis(2, c(0,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1,1.1,1.2), 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
```

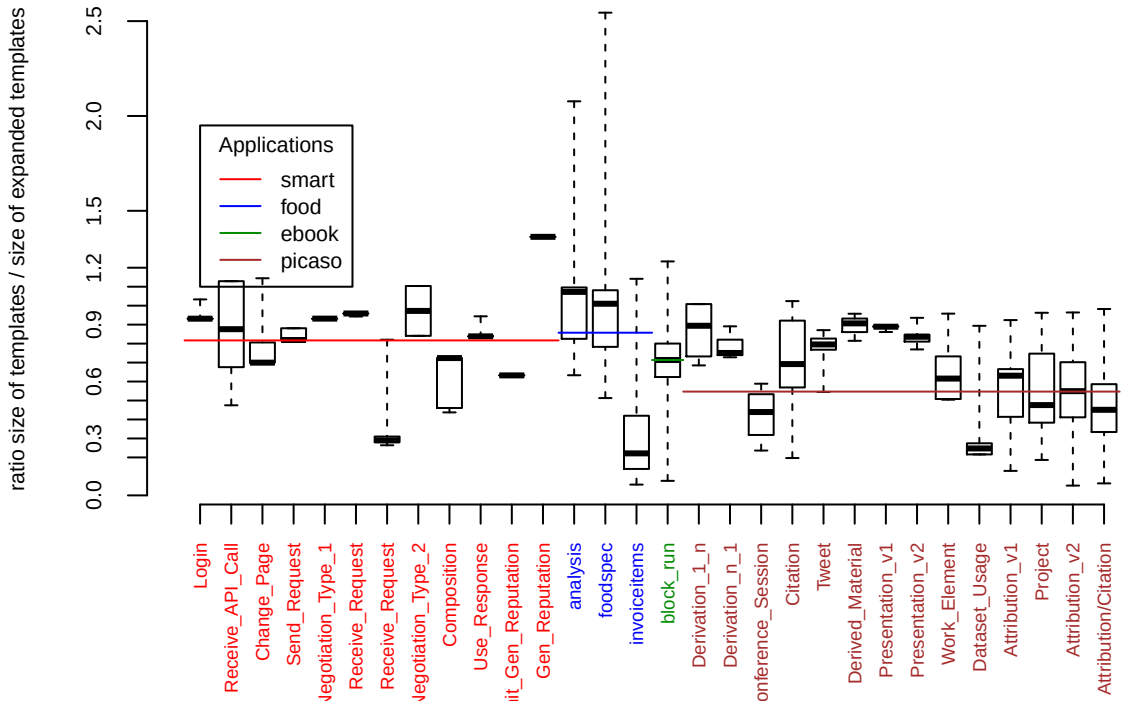
```r
# add a legend
legend(1,0.35, names(applications) , cex=0.7, col=colors[applications],lty=c(1,1),title="Applications")


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=30.5, y0=picaso_median,y1=picaso_median,col=colors[4])
```




```
##########################################################################
# box plot (templates)
```

Box Plot (compaction ratio templates/expanded provenance)

```
boxplot(yy$V11 ~ yy$V1, data=tmpl, axes=FALSE, range=0, ylab="ratio size of templates / size of expande

# Make y axis
axis(2, c(0,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1,1.1,1.2,1.5,2.0,2.5), 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(1,1.95, names(applications) , cex=0.7, col=colors[applications],lty=c(1,1),title="Applications")


segments(x0=0.5, x1=12.5, y0=smart_median2, y1=smart_median2, col=colors[1])
segments(x0=12.5,x1=15.5, y0=food_median2,  y1=food_median2,  col=colors[2])
segments(x0=15.5,x1=16.5, y0=ebook_median2, y1=ebook_median2, col=colors[3])
segments(x0=16.5,x1=30.5, y0=picaso_median2,y1=picaso_median2,col=colors[4])
```




```
###############################################################################
###
###  Version 1
###
###
###xx<-rbind(smart_w2,
###   food_w2,
###   ebook_w2,
###   picaso_w2
###          )
###
###xx$V10 <- xx$V4 / xx$V6
###yy <- xx[order(xx$V1),]
####yy$V11 <- 1:nrow(yy)
###
###
###############################################################################
#### box plot
###
###
###tmpl <- unique(yy$V1)
###
###
###
###boxplot(yy$V10 ~ yy$V1, axes=FALSE, range=0)
###
#### Make y axis
###axis(2, c(0,0.2,0.4,0.5,0.6,0.7,0.8,0.9,1,1.1,1.2,1.3), cex.axis=0.7)
###
####axis(1,tmpl,cex.axis=0.7)
###
```

```
#### how to color axis labels?
####col.axis=colors[2*applications[tmpl]+1-2]
###axis(1, at=1:length(tmpl), cex.axis=0.7, las=3, labels=FALSE)
###
###mtext(text=tmpl, side=1,at=1:length(tmpl),adj=1,col=colors[c(1,1,1,1,1,1,1,1,1,1,2,2,2,3,4,4,4,4,4,4
###
###
###legend(1,0.5, names(applications) , cex=0.7, col=colors[applications],lty=c(1,1),pch=mydots,title="Ap
###
###
###
###
print(smart_median)
```

```
## [1] 0.4476684
```

```
print(food_median)
```

```
## [1] 0.5327581
```

```
print(ebook_median)
```

```
## [1] 0.6601266
```

```
print(picaso_median)
```

```
## [1] 0.342117
```

```
print(total_median)
```

```
## [1] 0.3972071
```

```
print(smart_mean)
```

```
## [1] 0.430907
```

```
print(food_mean)
```

```
## [1] 0.5260223
```

```
print(ebook_mean)
```

```
## [1] 0.6704058
```

```
print(picaso_mean)
```

```
## [1] 0.350379
```

```
print(total_mean)
```

```
## [1] 0.4063708
```

```
print(smart_sd)
```

```
## [1] 0.07320251
```

```
print(food_sd)
```

```
## [1] 0.08038991
```

```
print(ebook_sd)
```

```
## [1] 0.06227174
```

```
print(picaso_sd)
```

```
## [1] 0.0533793
```

```
print(total_sd)
```

```
## [1] 0.1024571
```

Summary (bindings)

```
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_median,food_median,ebook_median,picaso_median,total_median))

summary <- t(summary)
colnames(summary) <- c("mean","sd", "median")
print(summary)
```

```
##             mean         sd    median
## smart  0.4309070 0.07320251 0.4476684
## food   0.5260223 0.08038991 0.5327581
## ebook  0.6704058 0.06227174 0.6601266
## picaso 0.3503790 0.05337930 0.3421170
## total  0.4063708 0.10245710 0.3972071
```

Summary (templates)

```
summary2 = c(1,2,3,4)
names(summary2)= names(applications)
summary2["smart"]=smart_mean2
summary2["food"]=food_mean2
summary2["ebook"]=ebook_mean2
summary2["picaso"]=picaso_mean2
summary2["total"]=total_mean2


summary2 = rbind(summary2,c(smart_sd2,food_sd2,ebook_sd2,picaso_sd2,total_sd2))
summary2 = rbind(summary2,c(smart_median2,food_median2,ebook_median2,picaso_median2,total_median2))



summary2 <- t(summary2)
```

```r
colnames(summary2) <- c("mean","sd", "median")
print(round(summary2,3))
```

```
##         mean    sd median
## smart  0.751 0.227  0.817
## food   0.797 0.419  0.857
## ebook  0.706 0.150  0.714
## picaso 0.580 0.218  0.547
## total  0.657 0.274  0.680
```

```r
print(picaso_names)
```

```
##    name             pretty
## 1     1     Attribution v1
## 2     2           Citation
## 3     3 Conference-Session
## 4     4      Dataset-Usage
## 5     5   Derived-Material
## 6     6       Work-Element
## 7     7              Tweet
## 8     8    Presentation v1
## 9     9    Presentation v2
## 10   10     Derivation-1-n
## 11   11     Derivation-n-1
## 12   12     Attribution v2
## 13   13            Project
```

```r
print(for_picaso_name("3"))
```

```
## [1] Conference-Session
## 13 Levels: Attribution v1 Attribution v2 Citation ... Work-Element
```

```r
print(for_picaso_name("10"))
```

```
## [1] Derivation-1-n
## 13 Levels: Attribution v1 Attribution v2 Citation ... Work-Element
```