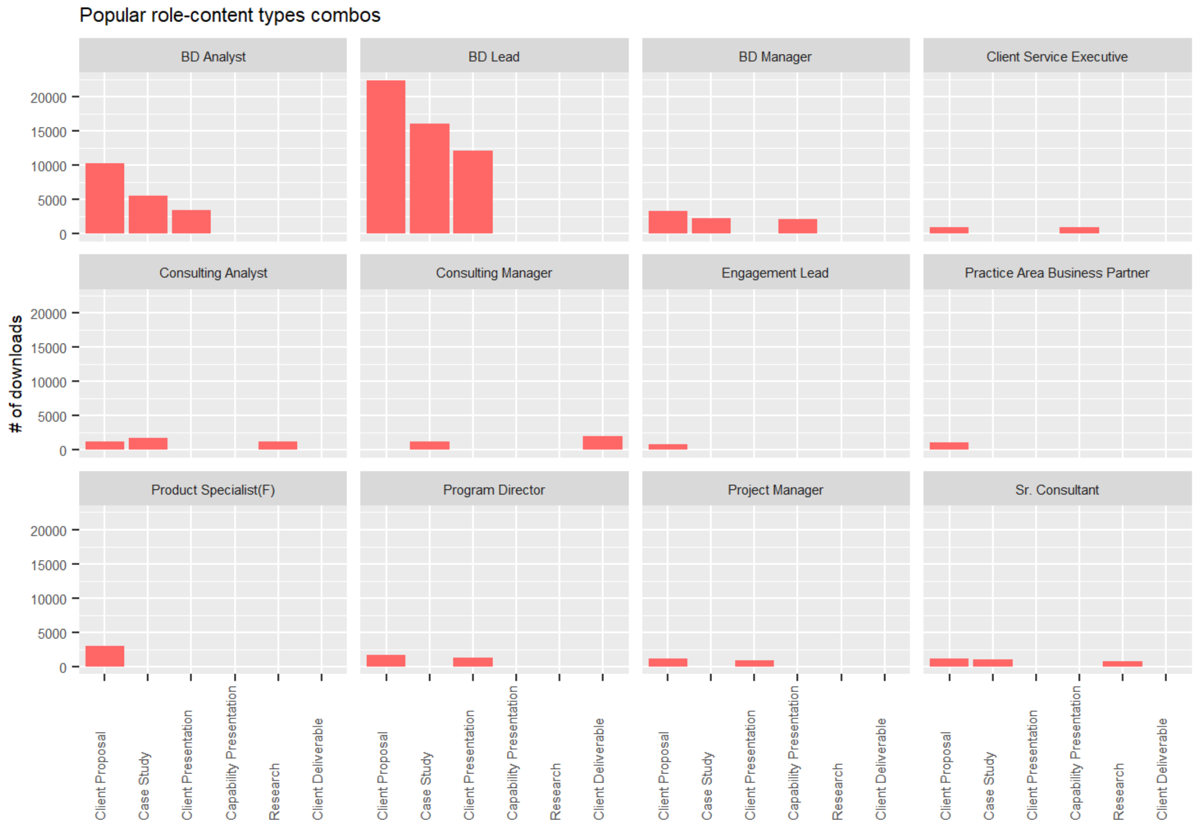
Asset Recommendation

Predictive analytics, Delivery Excellence

May 22, 2020



Context:

With ~200,000 artifacts in the SharePoint repository, it is a challenge for users to find useful content. A series of solutions were built to address this, including,

1. Semantic search: This suggests content similar to the user's search condition. This is described in a separate white paper (https://github.com/balawillgetyou/dy/blob/master/SemanticSearch20191126.pdf).
2. Subscription channels: Users can register their interests and have relevant content on their personalized home page.
3. Recommended content: Based on the user's past behavior, recommendations are made.

This documents covers the third item on the list above, recommended content.

Approach:

Users show clear patterns in their choice of content types. This has been studied and analyzed using the following approaches,
1. Visual analyis of content type wise download frequency. One of the resulting plots is shown above.
2. Collaborative filtering to help users discover new content types they may like, based on the preferences expressed by other users that share the same tastes.
3. Association rules mining that identifies popular content types for a given classification of user.

Collaborative filtering:

We begin by creating a Role v/s Content Type rating matrix, where we treat the number of downloads as the rating. The rating matrix provides insights into similar/ dis-similar tastes in content types, by role. But the results can vary depending on the distance measure used. For example, see that *Consulting Analyst* v/s *Consulting Director* similarity differs by distance measure.

Dissimilarity using cosine distance

```
##                          Biz Dev Manager Chief Architect
## Chief Architect                       NA                
## Client Service Executive      0.00000000      0.00000000
## Consulting Analyst            0.00000000      0.00000000
## Consulting Director           0.00000000      0.00000000
## Consulting Manager            0.00000000      0.00000000
##                          Client Service Executive Consulting Analyst
## Chief Architect                                                     
## Client Service Executive                                            
## Consulting Analyst                     0.34458924                   
## Consulting Director                    0.11706755         0.09147596
## Consulting Manager                     0.29780371         0.53068370
##                          Consulting Director
## Chief Architect                             
## Client Service Executive                    
## Consulting Analyst                          
## Consulting Director                         
## Consulting Manager                0.13034075
```

Similarity using cosine distance

```
##                          Biz Dev Manager Chief Architect
## Chief Architect                       NA
## Client Service Executive       1.0000000       1.0000000
## Consulting Analyst             1.0000000       1.0000000
## Consulting Director            1.0000000       1.0000000
## Consulting Manager             1.0000000       1.0000000
##                          Client Service Executive Consulting Analyst
## Chief Architect
## Client Service Executive
## Consulting Analyst                      0.6554108
## Consulting Director                     0.8829325          0.9085240
## Consulting Manager                      0.7021963          0.4693163
##                          Consulting Director
## Chief Architect
## Client Service Executive
## Consulting Analyst
## Consulting Director
## Consulting Manager                 0.8696593
```

Similarity using Pearson distance

```
##                          Biz Dev Manager Chief Architect
## Chief Architect                       NA
## Client Service Executive              NA              NA
## Consulting Analyst                    NA              NA
## Consulting Director                   NA              NA
## Consulting Manager                    NA              NA
##                          Client Service Executive Consulting Analyst
## Chief Architect
## Client Service Executive
## Consulting Analyst                      0.5677631
## Consulting Director                     0.6603352          0.5879119
## Consulting Manager                      0.5307377          0.5378860
##                          Consulting Director
## Chief Architect
## Client Service Executive
## Consulting Analyst
## Consulting Director
## Consulting Manager                 0.5758613
```

Similarity using Jaccard distance

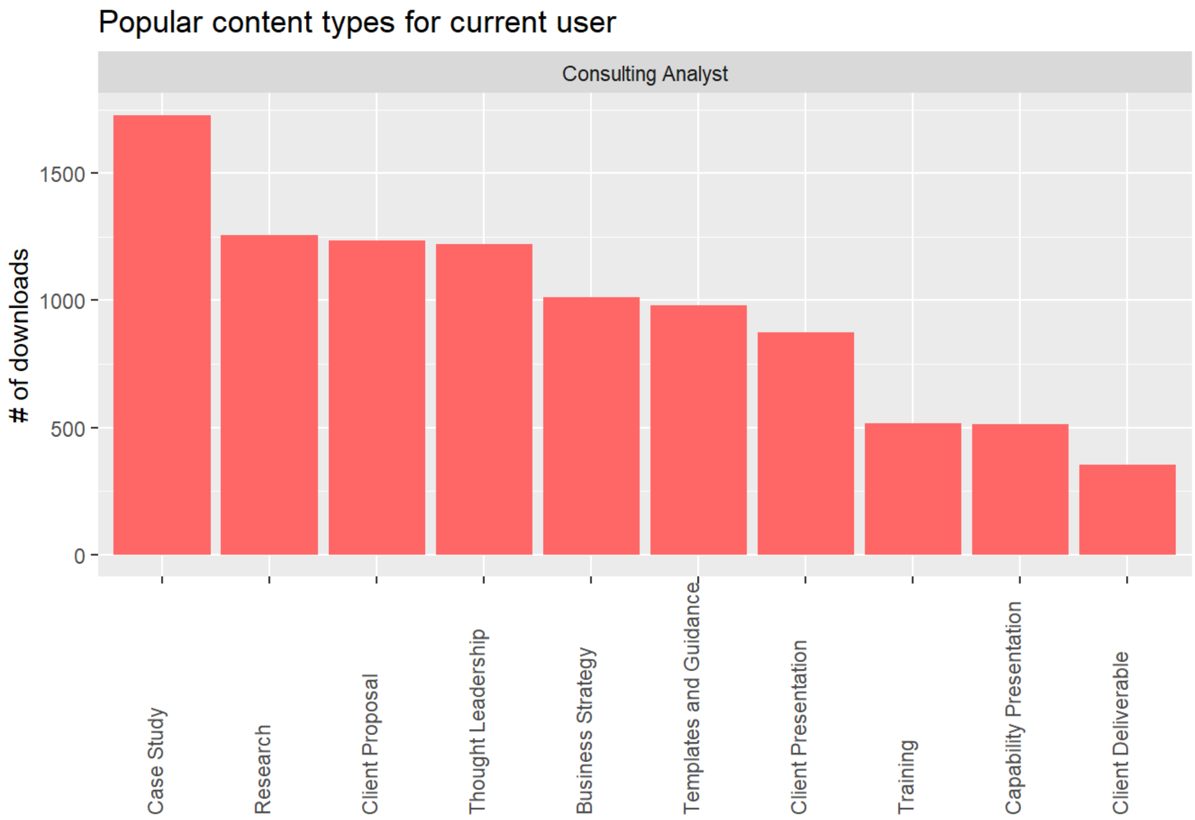
```
##                          Biz Dev Manager Chief Architect
## Chief Architect                       NA                
## Client Service Executive               1               1
## Consulting Analyst                     1               1
## Consulting Director                    1               1
## Consulting Manager                     1               1
##                          Client Service Executive Consulting Analyst
## Chief Architect                                                     
## Client Service Executive                                            
## Consulting Analyst                               1                  
## Consulting Director                              1                  1
## Consulting Manager                               1                  1
##                          Consulting Director
## Chief Architect                             
## Client Service Executive                    
## Consulting Analyst                          
## Consulting Director                         
## Consulting Manager                         1
```

Next we fit the collaborative filtering model itself, using the rating matrix. Note that there are many parameters to tune in this step. While the usual confusion matrix driven comparison is possible, we've produced output that a business user can use to compare differences between the new content types the recommender suggested v/s those the user currently uses.

The sample user role (user) and the top 10 suggested content type (item), with rating are below. A plot showing the top 10 content types the same user currently downloads is shown next to contrast and show that new content types are suggested.

```
## [1] "Consulting Analyst"
```

```
##                  user                       item    rating
## 1  Consulting Analyst            Golden Samples 0.2037462
## 2  Consulting Analyst                   Reports 0.2024447
## 3  Consulting Analyst         Training Material 0.2019691
## 4  Consulting Analyst Marketing and Bid Material 0.1998004
## 5  Consulting Analyst        Poster or Brochure 0.1986742
## 6  Consulting Analyst                 Checklist 0.1925662
## 7  Consulting Analyst         Proposal Reusable 0.1904017
## 8  Consulting Analyst                  Abstract 0.1832242
## 9  Consulting Analyst        Marketing Brochure 0.1810608
## 10 Consulting Analyst                    Mailer 0.1778555
```



Association rules mining:

First we convert our dataset into the transactions class and look at the head of the dataset.

```
##     items                       transactionID
## [1] {CCA_Role=Delivery Partner,
##      UserBU=Insurance,
##      ContentBU=Chubb,
##      ContentType=Best Practice}             1
## [2] {CCA_Role=Delivery Partner,
##      UserBU=Insurance,
##      ContentBU=Chubb,
##      ContentType=Case Study}                2
## [3] {CCA_Role=Delivery Partner,
##      UserBU=Insurance,
##      ContentBU=Chubb,
##      ContentType=Case Study}                3
```

Then, we create rules that associate user classification metadata with the content type and look at the head of the data frame.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.2    0.1    1 none FALSE            TRUE       5   0.001      4
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 290
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[576 item(s), 290522 transaction(s)] done [0.11s].
## sorting and recoding items ... [215 item(s)] done [0.01s].
## creating transaction tree ... done [0.16s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [420 rule(s)] done [0.00s].
## creating S4 object  ... done [0.02s].
```

```
##                                                                                                      rules
## 113 {CCA_Role=Product Specialist(F),UserBU=DE,ContentBU=PNR-MLEU-BDSpace} => {ContentType=Client Proposal}
## 112        {CCA_Role=Consulting Senior Manager,UserBU=EIM,ContentBU=EIM} => {ContentType=Capability Deck}
## 104       {CCA_Role=Consulting Manager,UserBU=NULL,ContentBU=CBC-PQC} => {ContentType=Client Deliverable}
## 4                 {CCA_Role=Solution Architect,UserBU=MSI,ContentBU=MSI} => {ContentType=Client Proposal}
## 192                  {CCA_Role=BD Lead,UserBU=EAS-SAP,ContentBU=EAS-CRM} => {ContentType=Client Proposal}
## 100        {CCA_Role=Consulting Manager,UserBU=EIM,ContentBU=CBC-AIM} => {ContentType=Client Deliverable}
##         support confidence      lift count
## 113 0.008474401  0.7751889  3.767115  2462
## 112 0.001273570  0.6313993 10.899958   370
## 104 0.004230316  0.5491510 55.920244  1229
## 4   0.001762345  0.5372508  2.610829   512
## 192 0.001297664  0.4504182  2.188856   377
## 100 0.002753664  0.4203889 42.808346   800
```

Code:

```
library(tidyverse)
library(ggplot2)
library(recommenderlab)
library(arules)

#data load & wrangling
setwd('C:\\Users\\654829\\Documents\\Persona')
data1 <- readxl::read_excel('Knowhub_UserData_Role_Updated_ARM_Jul19.xlsx', sheet='User Details')

#column renaming
names(data1) <- c("UserID","Designation","UserBU","Location","Date","ContentBU","ContentType","RHMS_Role","CCA_Role")

data1x <- data1[!is.na(data1$CCA_Role),]
data1x <- data1x[!is.na(data1x$ContentType),]
data1x <- cbind(newColName = rownames(data1x), data1x)
data1x$CCA_Role <- str_replace_all(data1x$CCA_Role , "Biz Dev Analyst", "BD Analyst")

df <- cbind(newColName = rownames(df), df)

#exploring data rationalization
countRole1 <- data1x %>% group_by(CCA_Role) %>% summarise(countRole = n_distinct(UserID))

#visual exploration and filtering for CCA_Role and ContentType that have more than a minimum number of downloads
data2 <- data1x %>%  group_by(CCA_Role, ContentType) %>%
  summarise(countCombo = n()) %>% filter(countCombo>100)

data2x <- inner_join(data1, data2, by = c('CCA_Role', 'ContentType'))

data2x %>%  group_by(CCA_Role, ContentType) %>%
  summarise(countCombo = n()) %>% filter(countCombo>800) %>% top_n(3) %>%
  inner_join(data1x, by = c('CCA_Role', 'ContentType')) %>%
  ggplot(aes(fct_infreq(factor(ContentType)))) +
  geom_bar(fill = "#FF6666") +
  labs(x='', y = "# of downloads") +
  theme(text = element_text(size=7)) +
  facet_wrap(CCA_Role~.) +
  theme(axis.text.x = element_text(angle = 90, hjust = 0, vjust = 0)) +
  ggtitle('Popular role-content types combos')

data3 <- data2x %>% select(-countCombo) %>% group_by(CCA_Role, ContentType)%>%
  summarise(countCombo = n()) %>% select(CCA_Role, ContentType, countCombo)

#CCA Role v/s Content Type rating matrix
ratingMat <- data3 %>% spread(ContentType, countCombo) %>% select(-CCA_Role)
ratingMat <- ratingMat[,-1]
```

```
ratingMat[is.na(ratingMat)] <- 0
ratingMat <- as.matrix(ratingMat)
ratingMat <- ratingMat %*% diag(1/colSums(ratingMat))#this normalization step is important


dimension_names <- list(CCA_Role_id = sort(unique(data2x$CCA_Role)), ContentType_id = sort(unique(data2x$ContentType)))
dimnames(ratingMat) <- dimension_names

ratingMat0 <- ratingMat
ratingMat0[is.na(ratingMat0)] <- 0
sparse_ratings <- as(ratingMat0, "sparseMatrix")

real_ratings <- new("realRatingMatrix", data = sparse_ratings)

dissimilarity(real_ratings[8:13], method = "cosine")
similarity(real_ratings[8:13], method = "cosine")
similarity(real_ratings[8:13], method = "pearson")
similarity(real_ratings[8:13], method = "jaccard")


#collaborative filter model:
modelUBCF <- Recommender(real_ratings, method = "UBCF", param = list(method = "cosine", nn = 40))

#current_user <- "BD Manager"
#current_user <- "BD Analyst"
current_user <- 'Consulting Analyst'
#current_user <- 'BD Lead'
#current_user <- 'Program Director'
#current_user <- 'Client Service Executive'

prediction <- predict(modelUBCF, real_ratings[current_user, ], type = "ratings")
prediction <- as(prediction, 'data.frame')
predictionOut <- prediction %>% arrange(desc(rating))
predictionOut[1:10,]

data2User <- data1x %>% filter(stringr::str_detect(CCA_Role, current_user)) %>%
  group_by(CCA_Role, ContentType) %>% summarise(countCombo = n()) %>% top_n(10)

data2xUser <- inner_join(data1x, data2User, by = c('CCA_Role', 'ContentType'))

data2xUser %>%  group_by(CCA_Role, ContentType) %>%
  summarise(countCombo = n()) %>% filter(countCombo>100) %>% top_n(10) %>%
  inner_join(data1x, by = c('CCA_Role', 'ContentType')) %>%
  ggplot(aes(fct_infreq(factor(ContentType)))) +
  geom_bar(fill = "#FF6666") +
```

```
  labs(x='', y = "# of downloads") +
  facet_wrap(CCA_Role~.) +
  theme(axis.text.x = element_text(angle = 90, hjust = 0, vjust = 0)) +
  ggtitle('Popular content types for current user')

#Association Rules Mining:
tData <- data1x %>% select(CCA_Role, UserBU, ContentBU, ContentType) %>% as("transactions") # convert to 'transactions' clas
s
inspect(head(tData,3))
frequentItems <- eclat(tData, parameter = list(supp = 0.07))
inspect(frequentItems)
itemFrequencyPlot(tData, topN=10, type="absolute", main="Item Frequency")
#rules, various slices/ insights
rules <- apriori (tData, parameter = list(supp = 0.001, conf = 0.2, minlen=4))
#rules, various slices/ insights
rules_conf <- sort (rules, by="confidence", decreasing=TRUE)
inspect(head(rules_conf))
#
rules_contentType <- subset(rules_conf, rhs %pin% c('ContentType'))
#arules::inspect(head(rules_contentType))
rules_contentTypeDF <- as(rules_contentType, 'data.frame')
head(rules_contentTypeDF)
```