



ASSOCIATION MINING

Website Operation Pattern Analysis

Mitkumar Patel	A00444857
Chirag Panasuriya	A00442907
Wensho Li	A00445457

Contents

Executive Summary	1
Objective.....	2
About Data	2
Association Mining	3
User Level	3
Session Level.....	3
Design/Methodology/Approach	4
Apriori Algorithm.....	4
Steps for Apriori Algorithm	4
Parameters	4
Data Preparation for User Analysis	5
Raw data	5
User Level	6
Session Level.....	6
Rule Discovery	6
Rules Analysis	7
User Level	8
Session Level.....	10
Conclusion	11
Appendix.....	12

Executive Summary

An analysis of user sessions data was performed to find association between user behaviors so that we can better understand the pattern of users' operating behaviors online. The analysis focused on users' operating behaviors on a specific website, which behaviors are from following table:

ResetAccountSett	SaveAccountSetti	SendLater	SpellSearch
ManageTab	SetDefaultAccoun	SendRecurring	TxtIndentJustify
ReportsTab	ViewAccountSetti	SocialAutoPost	OpenReportList
AddImage	AccountSettingsA	SocialShareEleme	InsertList
ReEditProj	AssignAccountSet	testTest	TxtMergeTags
SEDragIn	ChangeAccountSet	TextOnlyProjCrea	SEBackColors
SendNow	CreateAccountSet	UseBlankTemplate	ManageFilters
CaptionedImages	RenameAccountSet	ActivateAutoresp	MobileFriendly
ImageProperties	RevokeAccountSet	AutoresponderRep	PrevMobileVersio
OpenReport	UploadImage	AutoresponderSum	PrevTextVersion
SEBorders	InsertLinkAnchor	CopyAutoresponde	TableSort
SEDragResize	InsertTableImage	CopyCrossMarkete	ABSplitTools
SELinkElements	SEDeleteElement	CreateAutorespon	ConditionalEleme
SEPadding	TxtViewSource	CreateCrossMarke	ToggleSocialSite
SimpleProjCreate	TestSend	CrossMarketerCon	ActivateCrossMar
ABSplitProj	TxtFontSizeColor	CrossMarketerRep	PublishTemplate
AddFB	LinkActivity	CrossMarketerSum	TableSearch
AddTwitter	Location	DeactivateCrossM	MultImageElemen
AdvProjCreate	OpenData	DeleteAutorespon	SubjectMergeTags
CopyProj	OpensAndBounces	DeleteCrossMarke	SocialFollowElem
DeleteProj	OpensByPlatform	EditAutoresponde	SetCompanyDefaul
ImportProjEmail	UnopenedData	EditCrossMarkete	CompareReports
ImportProjFileUR	UserAgent	RenameAutorespon	ReportsFilters
ProjPreview	UseReportCharts	RenameCrossMarke	EditDefaultStyle
RenameProj	SECreateColumn	SaveAutoresponde	MultImageProper
ReSchedProj	TxtBIUS	SaveCrossMarkete	HideOnMobile
SEDragMove	TxtStylesFormats	AddAccountSettin	RSSElement
AutoresponderCon	SEPublishElement	DeleteAccountSet	RSSProperties

Table-1

And the analysis is processing on two different level: user level and session level.

Aim is to find the association between button and tabs that are present in the Simplicast.com.

The user activity like clicks on the tabs and buttons are recorded and the data has been stored in the database. With the given data, association mining using apriori algorithm is performed at user level and session level. Consequently, the different sets of tabs and buttons that were frequently used at their respective levels.

Objective

Website/app is becoming an important part of our daily life, which brings a huge amount of business opportunities that could make billions of profits for us. So, there is a necessary that gaining insights into our customers by tracking all their activity on website and in app, and then continue to follow through by tracking all the interactions we have with them. Once we got the correct pattern of customer behaviors, we could easily optimize the way our website/app interacting with customers and organize customers into pipelines for sales and support, which could both improve user experience and reduce maintenance pressure and cost of website/app. Finally, we can take more market sharing and make more profits. It is the report author's intent to perform various data preparation, association mining, and association analysis to gain insight of user behavior pattern that will provide beneficial information to the business owner. In the absence of specific requirements from the business, the report's authors will consider the goal of defining user operating pattern to be knowledge generation for association between different operating behaviors both on user level and session level. A successful analysis will produce delineated and communicable profiles for user behavior pattern that will aid in decision making and support future development plan for website/app.

About Data

To support the analysis, website operation behavior dataset captured between July 2015 to December 2015. These operation data provide operation behavior pattern of users as it relates to both user level and session level, which show difference of different users and common pattern of all users. A brief description of user and session are provided herein:

User Table:

Provides additional information for each user in defining the operation behavior he/she have on the website (distinct behavior). It does not include the time and date of operation. So, this table is used to mine the association between among the user level.

Number of records:	39096
Number of unique records:	39096
Reference Field:	N/A(only one table)

Session Table:

Provide additional information of operation behavior processed by each user in any specific day. This table is used to identify association between session by specify date. We focus on what behavior user have to this website on one day so that we can know which operations have association in session level.

Number of records:	173082
Number of unique records:	173082
Reference Field:	N/A(only one table)

Association Mining


Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories. We had two CSV files; first file is related to user's distinct milestone and another file has same information with the date.

User Level

We transposed CSV data and converted into row format for the association mining. One user performs various task, and those steps are stored in various rows. For association mining this format is not suitable so various tasks of single user is stored in one row. This transformed data is useful to predict next transaction of user based on current transaction, but we can't say when next transaction occurs. Next transaction might be performed after one day, one month or one year.

Example shown below:

ID	Transaction
A	1
A	2
B	3
C	1
A	5



ID	Transaction		
A	1	2	5
B	3		
C	1		


Figure-1

Session Level

This CSV file has more data and information compare to previous file. In the User level we considered only id for combining milestone but in this case, we used id and transaction date for combining milestone. This transformed data is quite useful and helps to discover relationship between transactions. Each row stores information about user's transaction for any given day so we can predict next move of user for the same day.

Simple version is shown below:

ID	DATE	Transaction
A	a	1
A	a	2
B	a	3
C	b	1
A	b	5



ID	DATE	Transaction	
A	a	1	2
A	b	5	
B	a	3	
C	b	1	

Figure-2

Design/Methodology/Approach

There are many algorithms for association mining but we choose Apriori algorithm. This is the most simple and easy-to-understand algorithm among association rule learning algorithms. The resulting rules are intuitive and easy to communicate to an end user.

Apriori Algorithm

The Apriori algorithm uses frequent item sets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected.

Frequent item sets are those items whose support is greater than the threshold value or user-specified minimum support. It means if A & B are the frequent item sets together, then individually A and B should also be the frequent itemset.

Suppose there are the two transactions: A= {1,2,3,4,5}, and B= {2,3,7}, in these two transactions, 2 and 3 are the frequent item sets.

Steps for Apriori Algorithm

Below are the steps for the apriori algorithm:

- Step-1: Determine the support of item sets in the transactional database and select the minimum support and confidence.
- Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.
- Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
- Step-4: Sort the rules as the decreasing order of lift.

Parameters

Support:

Support us an indication of how frequently the item appears in the data.

$$Support(\{X\}) = \frac{\text{Transactions containing } X}{\text{Total number of transactons}} = P(X)$$

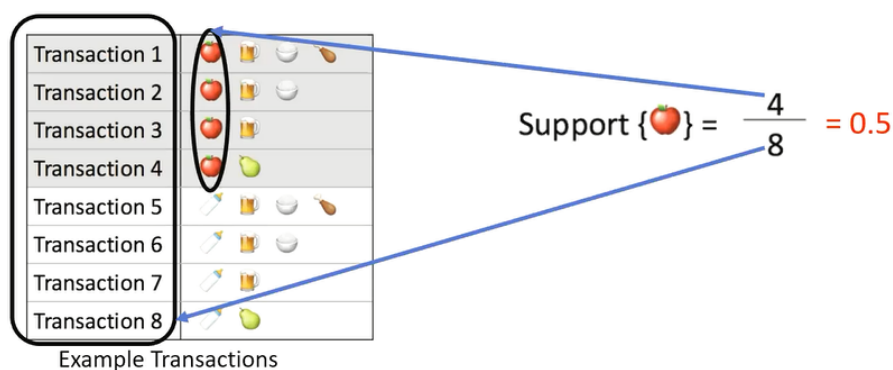


Image Source: <https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>

Figure-2

Confidence

Confidence is how likely Y is purchased when item X is purchased.

$$\text{Confidence}(\{X\} \rightarrow \{Y\}) = \frac{\text{Transactions containing both } X \text{ and } Y}{\text{Transaction containing } X}$$

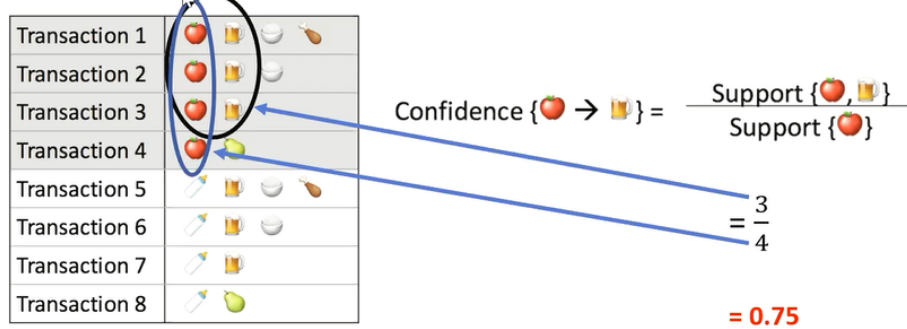


Image Source: <https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>

Figure-3

Lift

Lift is how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

$$\text{Lift}(\{X\} \rightarrow \{Y\}) = \frac{(\text{Transactions containing both } X \text{ and } Y)}{(\text{Transaction containing } X)(\text{Transaction containing } Y)} = \frac{P(X, Y)}{P(X) * P(Y)}$$

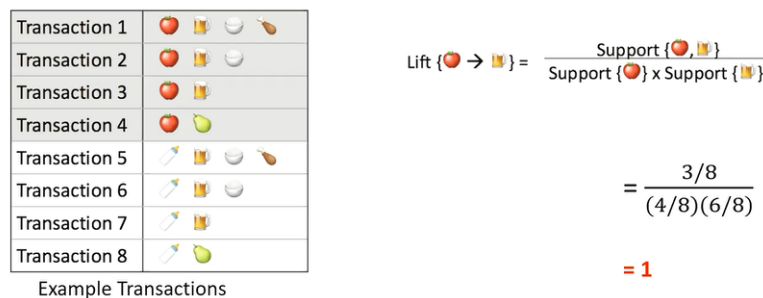


Image Source: <https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>

Figure-4

Data Preparation for User Analysis

As the data for this assignment is already provided by instructor of MCDA5580, Trishla. This part is just the simulation of prepare dataset procession.

Raw data

The original data has 5 column, which are "id" (primary key of raw data table), "user_id" (user id), "milestone_name" (operation of user), "date" (operation date) and "time" (operation time).

Showing as below:

id	user_id	milestone_name	date	time
76085	5813621	ManageTab	2015-07-20	18:34:49
76086	5813621	SimpleProjCreate	2015-07-20	18:35:31
76087	5813621	UseBlankTemplate	2015-07-20	18:36:14
76088	5813621	SEDragIn	2015-07-20	18:36:31
76089	5813621	SEDragIn	2015-07-20	18:36:32
76090	5813621	SEDragIn	2015-07-20	18:36:35
76091	5813621	AddImage	2015-07-20	18:36:40
76092	5813621	SEDragIn	2015-07-20	18:36:43
76093	5813621	ImageProperties	2015-07-20	18:36:50
76094	5813621	ImageProperties	2015-07-20	18:36:50
76095	5813621	SELinkElements	2015-07-20	18:37:13
76096	5813621	SendNow	2015-07-20	18:38:33
76097	5813621	SendNow	2015-07-20	18:40:23
76098	5813621	SendNow	2015-07-20	18:40:48
76099	5813621	ManageTab	2015-07-20	18:41:05
76100	5813621	ReportsTab	2015-07-20	18:41:34
76101	5813621	OpenReport	2015-07-20	18:41:43

Figure-5

User Level

Retrieve user level data from raw data.

SQL Statement:

```
CREATE TABLE userMilestone as
SELECT user_id, milestone_name FROM dataset03.rawdataDec15 ORDER BY user_id,
milestone_name
```

```
CREATE TABLE userDistinctMilestoneDec15 as SELECT DISTICT * from userMilestone
```

The reason why we choose distinct record of user operation is that we intend to analysis operation pattern in user level without influence of counts of operation.

Session Level

Retrieve user level data from raw data.

SQL Statement:

```
CREATE TABLE sessionMilestone as
SELECT user_id, milestone_name, date FROM dataset03.rawdataDec15 ORDER BY user_id,
milestone_name
```

The reason why we retrieve date as part of session dataset is that the authors of this report consider one day will be a more appropriate time frame to analysis the association between operation behavior on website. Because if we choose both date and time as part of session dataset, there will be too much noise in data, as each operation of user could be influenced by multiple influence source.

Rule Discovery

Apriori generate lots of rules based on given data. All rules are not useful and effective, so we need to ignore rules. We can use confidence and support to decide whether rule is useful or not and sometimes it is desirable to remove the rules that are subset of larger rules. To

perform Rule discovery in R, matrix operations are required because “subset” function return matrix in execution.

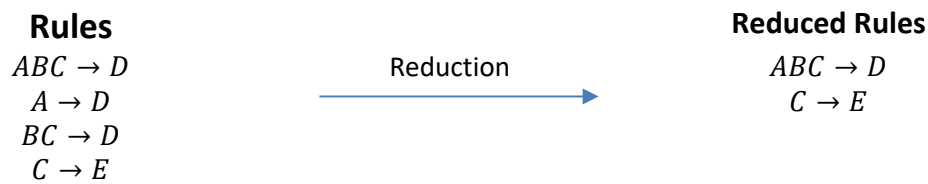


Figure-6

Rules Analysis

Association rule mining, it involves the use of machine learning models to analyze data for patterns, or co-occurrences, in a database. It defines frequent if-then associations.

An association rule consists of two parts:

1. antecedent (if)
2. consequent (then)

An antecedent is an item with If condition. A consequent is an item found in combination with the antecedent.

Association rules are generated by finding data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships.

Support is an indication of how frequently the items appear in the data.

Confidence indicates the number of times the if-then statements are found true.

lift is used to compare confidence with expected confidence, or how many times an if-then statement is expected to be found true.

If the lift value is a negative value, then there is a negative correlation between datapoints. If the value is positive, there is a positive correlation, and if the ratio equals 1, then there is no correlation.

User Level

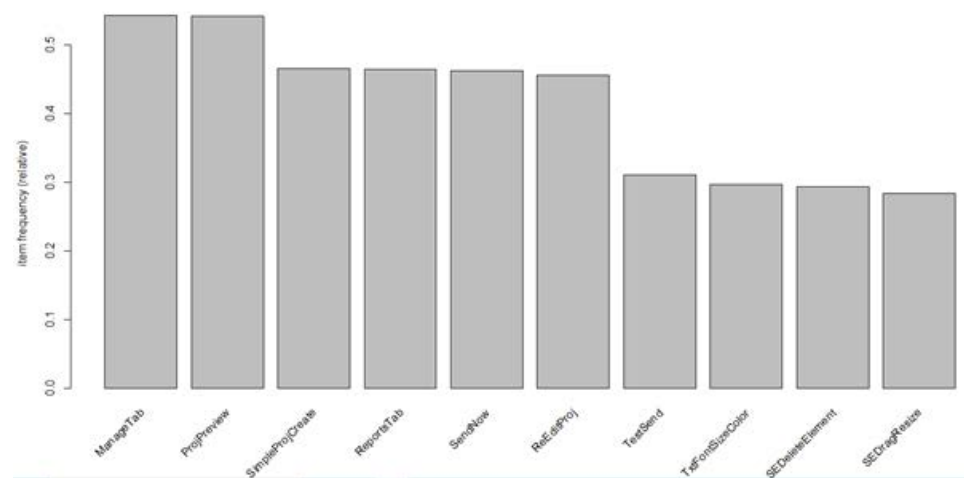


Figure-7: Top 10 frequently occurred milestones for user level association

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{ManageTab, ReEditProj}	=> {SendNow}	0.3469452	0.9087894	0.3817664	1.962314	1096
[2]	{ReEditProj, ReportsTab}	=> {SendNow}	0.3108579	0.9042357	0.3437797	1.952482	982
[3]	{ReEditProj, ReportsTab, SendNow}	=> {ManageTab}	0.3038936	0.9775967	0.3108579	1.799667	960
[4]	{ProjPreview, ReEditProj, SendNow}	=> {ManageTab}	0.3206711	0.9502814	0.3374486	1.749382	1013
[5]	{ReEditProj, SendNow}	=> {ManageTab}	0.3469452	0.9383562	0.3697373	1.727428	1096
[6]	{ManageTab, ReEditProj, ReportsTab}	=> {ProjPreview}	0.3089585	0.9357622	0.3301678	1.723658	976
[7]	{ReEditProj, ReportsTab}	=> {ProjPreview}	0.3194049	0.9290976	0.3437797	1.711382	1009
[8]	{ManageTab, ReEditProj, SendNow}	=> {ProjPreview}	0.3206711	0.9242701	0.3469452	1.702489	1013
[9]	{ReportsTab}	=> {ManageTab}	0.4283001	0.9222904	0.4643875	1.697853	1353
[10]	{ManageTab, ReportsTab, SendNow}	=> {ProjPreview}	0.3187718	0.9179581	0.3472618	1.690863	1007
[11]	{ReportsTab, SendNow}	=> {ProjPreview}	0.3266857	0.9132743	0.3577081	1.682235	1032
[12]	{ManageTab, ReEditProj}	=> {ProjPreview}	0.3485280	0.9129353	0.3817664	1.681611	1101
[13]	{ReEditProj, SendNow}	=> {ProjPreview}	0.3374486	0.9126712	0.3697373	1.681124	1066
[14]	{SendNow}	=> {ManageTab}	0.4169041	0.9002051	0.4631212	1.657196	1317

Figure-8

Let's understand couple of rules in with explanation.

Rule 1.

{ManageTab, ReEditProj} => {SendNow}

User who clicks Milestone ManageTab and ReEditProj are likely to click SendNow also.

These both are the antecedents {ManageTab, ReEditProj}

{SendNow} is the consequence. We can consider them as LHS – RHS

- Support 0.3469452

Approx. 35 % of all the buttons and tabs represent this LHS and RHS combination. This is indication of how frequently the items appear in the data.

- Confidence 0.9087894

It means 91 % of all events likely to have SendNow button click.. Confidence indicates the number of times the if-then statements are found true.

- Lift 1.96

Lift value basically tell us how significant as the consequent with respect to the antecedent so we see that this is like all these values are approx. two times more significant. It is the ratio of confidence to support.

Rule 2.

$\{\text{ReEditProj}, \text{ReportsTab}, \text{SendNow}\} \Rightarrow \{\text{ManageTab}\}$

User who clicks Milestone ReEditProj , ReportsTab and SendNow are likely to click ManageTab also.

These are the antecedents { ReEditProj,ReportsTab,SendNow }. { ManageTab } is the consequence.

- Support 0.31085

Approx. 31 % of all the buttons and tabs represent this LHS and RHS combination. This is indication of how frequently the items appear in the data.

- Confidence 0.904

It means 90 % of all events likely to have ManageTab button click. Confidence indicates the number of times the if-then statements are found true.

- Lift 1.95

Lift value basically tell us how significant as the consequent with respect to the antecedent so we see that this is like all these values are approx. 1.95 times more significant. It is the ratio of confidence to support.

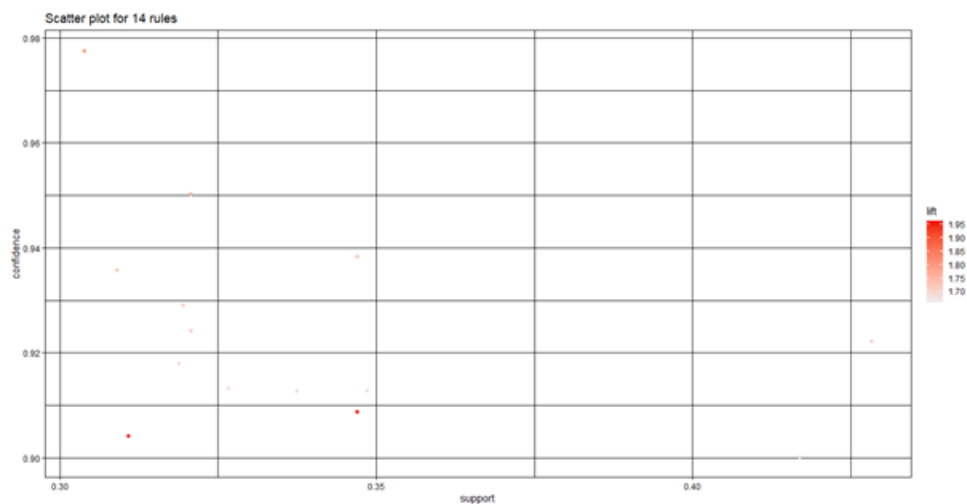


Figure-9

- We plot the rules here on the graph and the y-axis we have confidence, and, on the x-axis, we have support.
- From the heat map we can say that the left side darker red dots are mostly occurring at low to medium support

Session Level

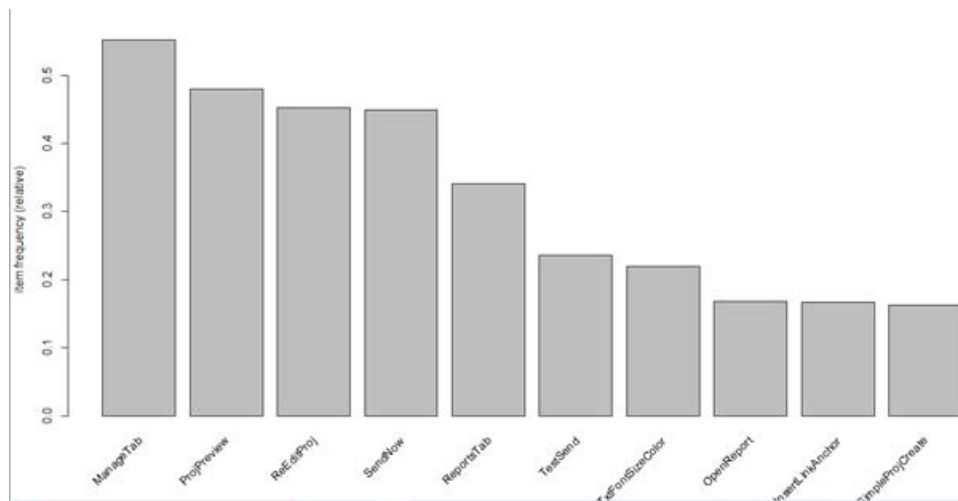


Figure-10: Top 10 frequently occurred milestones at session level association

```
> inspect(rules)
```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{ProjPreview}	=> {ManageTab}	0.3190224	0.6655973	0.4793024	1.206285	7884
[2]	{SendNow}	=> {ManageTab}	0.4047263	0.9010811	0.4491563	1.633061	10002
[3]	{ManageTab}	=> {SendNow}	0.4047263	0.7334996	0.5517744	1.633061	10002

Figure-11

Rule 1.

{ProjPreview} => {ManageTab}

User who clicks Milestone ProjPreview is likely to click ManageTab also.

{ ProjPreview } is the antecedents, {SendNow} is the consequence.

- Support 0.3190224

Approx. 31 % of all the buttons and tabs represent this LHS and RHS combination. This is indication of how frequently the items appear in the data.

- Confidence 0.6655973

It means 66 % of all events likely to have ManageTab button click.. Confidence indicates the number of times the if-then statements are found true.

- Lift 1.20

Lift value basically tell us how significant as the consequent with respect to the antecedent so we see that this is like all these values are approx. 1.2 times more significant. It is the ratio of confidence to support.

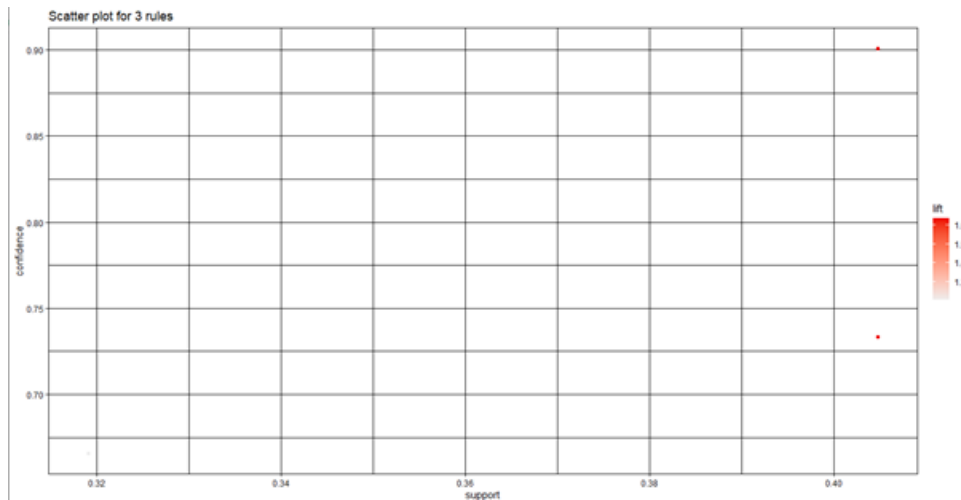


Figure-12

- We plot the rules here on the graph and the y-axis we have confidence, and, on the x-axis, we have support.
- From the heat map we can say that the right side darker red dots are mostly occurring at high support
-

Conclusion

From our analysis, we got some rules that the website may take use for future maintenance. And suggestions are as following:

- ManageTab button should be placed in the most conspicuous place of n navigation menu, which is not only because it is the most common use button and has a high support but also associates with a lot of other functions.
- ManageTab should have a drop list when cursor moving on it. And the list could contain ReEditProj, ReportsTab and ProjPreview. For one reason, the logic make sense, which could help users find function more easily. Besides, it fits the rule $\{ManageTab, ReEditProj\} \Rightarrow \{SendNow\}$. Move cursor on ManageTab and then choose ReEditProj.
- SendNow should on the highlight place on edit project page. First, it is in logic order: edit project \Rightarrow send; Second, it fits the rule: $\{ManageTab, ReEditProj\} \Rightarrow \{SendNow\}$.
- ProjPreview should be in the manage page of project and every project should have their own ProjPreview. After the preview, users could back to manage page by ManageTab button which is always on a navigation menu.

Appendix

R Script

```
library(arules)
library(plyr)
df_user= read.csv("./user.csv")
df_user = ddply(df_user,c("id"),function(dfl)paste(dfl$milestone, collapse = ","))
df_user$id = NULL
write.table(df_user,"./Milestones2.csv", quote=FALSE, row.names = FALSE, col.names = FALSE)
tr = read.transactions("./Milestones2.csv",format="basket",sep=",")
summary(tr)
itemFrequencyPlot(tr, topN=10)
rules = apriori(tr,parameter = list(supp=0.3,conf=0.9))
inspect(rules)
inspect(sort(rules,by='lift'))
itemsets=unique(generatingItemsets(rules))
itemsets
inspect(itemsets)

rules = apriori(tr,parameter = list(supp=0.30,conf=0.9))
rules=sort(rules,by='lift')
subset.matrix = is.subset(rules, rules)
precolumn=colSums(subset.matrix, na.rm=T)

#REMOVE SUB RULES RULE, Consider only Main rule
x=subset.matrix@Dimnames[[1]]
subset.matrix[lower.tri(subset.matrix, diag=T)] = 0
column=colSums(subset.matrix, na.rm=T)
notredundant=c()
for(i in 1:length(column)){
  if(column[i]<1){
    if(precolumn[i]-column[i]==1){
      notredundant[i]=TRUE
    }
    else{
      notredundant[i]=FALSE
    }
  }
  else{
    if(column[i]==1){
      temp=which(x %in% x[i])
      notredundant[temp]=TRUE
    }
    else{
      notredundant[i]=TRUE
    }
  }
}
rules=rules[notredundant]
rules=sort(rules,by='lift')
inspect(rules)
```

R Script1: User level association mining

```
library(arules)
library(plyr)
```

```

df_user= read.csv("./session.csv")
df_user = ddpily(df_user,c("user_id","date"),function(dfl)paste(dfl$milestone, collapse = ","))
df_user$user_id = NULL
df_user$date = NULL
write.table(df_user,"./Milestones2.csv", quote=FALSE, row.names = FALSE, col.names = FALSE)
tr = read.transactions("./Milestones2.csv",format="basket",sep=",")
summary(tr)
itemFrequencyPlot(tr, topN=10)
rules = apriori(tr,parameter = list(supp=0.3,conf=0.6))
inspect(rules)
inspect(sort(rules,by='lift'))
itemsets=unique(generatingItemsets(rules))
itemsets
inspect(itemsets)

rules = apriori(tr,parameter = list(supp=0.3,conf=0.5))
inspect(rules)
inspect(sort(rules,by='lift'))
itemsets=unique(generatingItemsets(rules))
itemsets
inspect(itemsets)

rules = apriori(tr,parameter = list(supp=0.3,conf=0.6))
subset.matrix = is.subset(rules, rules)
precolumn=colSums(subset.matrix, na.rm=T)
x=subset.matrix@Dimnames[[1]]

subset.matrix[lower.tri(subset.matrix, diag=T)] = 0
column=colSums(subset.matrix, na.rm=T)
notredundant=c()
for(i in 1:length(column)){
  if(column[i]<1 & precolumn[i]<1){
    if(precolumn[i]-column[i]==1){
      notredundant[i]=TRUE
    }
    else{
      notredundant[i]=FALSE
    }
  }
  else{
    if(column[i]==1){
      temp=which(x %in% x[i])
      notredundant[temp]=TRUE
    }
    else{
      notredundant[i]=TRUE
    }
  }
}
rules=rules[notredundant]
inspect(rules)

```

R Script2: Session level association mining

Reference:

<http://r-statistics.co/Association-Mining-With-R.html>

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781784390815/9/ch09lvl1sec100/pruning-redundant-rules