

**EB5202 Web Analytics – Web Usage Mining**

**Microsoft Report**

Team Members

Ang Wei Hao Max - A0178278L

Bharat Nagaraju - A0178258N

Khine Zin Win - A0163222U

Koo Chee Kit - A0178200N

Li Xiaoguang - A0178450B

Lin JunLiang - A0178295M

Contents

[1. Objectives 2](#_Toc528747163)

[2. Executive Summary 2](#_Toc528747164)

[3. Dataset 3](#_Toc528747165)

[4. Sequence Mining - Methodological Procedures 3](#_Toc528747166)

[4.1. Data Pre-processing 3](#_Toc528747167)

[4.2. Data Analysis 4](#_Toc528747168)

[4.3. Recommendations 5](#_Toc528747169)

[5. Association Mining - Methodological Procedures 5](#_Toc528747170)

[5.1. Data Pre-processing 5](#_Toc528747171)

[5.2. Data Analysis 6](#_Toc528747172)

[5.3. Recommendations 12](#_Toc528747173)

[**6.** Conclusion 12](#_Toc528747174)

# Objectives

The objective of this exercise is to analyse the logs from websites to find user behaviour on the website and derive useful insights based on various log analysis techniques. In this assignment we will employ association & sequence mining to understand if there are any pages on Microsoft website that are closely associated and being frequently visited in a sequence, based on which pages can be recommended to users. And before applying the recommendations on the website, we will also test page recommendations to users by using back testing techniques.

# Executive Summary

Our team consists of students from EBAC, KE and SE. After our basic analysis on the available datasets and considering their size and basic characteristics, we chose to use MSFT-VROOTS dataset, a website log holding about 132,000 records pertaining to 38000 users randomly sampled from Microsoft’s website. The dataset was big enough to be run on a personal laptop and was smaller than other available datasets. That ensured our PC’s wouldn’t collapse while pre-processing the datasets. As such the analysis techniques used across different datasets would be same, choosing a smaller dataset would mean we had to struggle a bit to discover insights presuming that the number of patterns that would be mined would be less comparatively and we would have to play around with different parameters to discover insights. The team was ok with it.

Next step was to transform this raw data of weblogs into a usable (transactional) format that can be used for further analysis. The format of the raw data was such that there were no key columns to associate user and the pages visited. They were just independent records placed in an order. The challenge was to preserve the sequence of the original data records since we were mining the sequence. This was a little cumbersome as there were no direct methods to convert the file format using data frame interpolation / transformation & We had to loop through all the records manually to create the file in the required format. Looping through 132K records took almost 30-45 mins and we had to wait patiently to convert the raw data to transactional format. Once the pre-processed file was available, Coding the association mining, back testing and Sequence mining was not that difficult.

We Started off with the Sequence analysis using R. For more details, please refer to section 4.1. Using a R library called ***arulesSequence*** , we applied cSpade algorithm to mine the sequence. 8 set of sequences were discovered. Further analysis using rule Induction function, Free Download and Internet Explorer (IE) are two occurrences are likely dependent on one another. The result is correct because IE was free for download. The recommendation is make use of the 2 web pages below to promote new products. The link between the two pages should not remove or if there is no link between the 2 pages, consider adding a link to improve user experience. We should spend effort to discover webpages that do not have sequence. If sequence is expected from business goals, do make changes to those websites and redo the sequence analysis when there is good enough raw data been logged after the changes.

|  |  |  |  |
| --- | --- | --- | --- |
| Rule | Support | Confidence | Lift |
| {Free Downloads|/msdownload}> => <{Internet Explorer|/ie} | 0.1591208 | 0.4803433 | 1.674572 |

For association mining, we split the data into Training and test data with a ratio of 80:20 We used a library called ***aRules*** and a function called ***Apriori*** to build association rules between pages using the 80% of the data available – labelled as “Training data”. These rules do not consider any sequence. For detail, please refer to section 5.1. We were able to mine only two rules that are same as mined in Sequence mining with min. support as 10%, min. confidence as 10% & considering association between two webpages at a time. The yielded rules had a max confidence of 43%. Mined rules were run against the remaining 20% of the data – labelled as “Test Data”. The results were quite underwhelming since the precision was around 12% from approximately 3800 Total predictions made. So, we reduced the support further to 4% and reran the mining to discover 12 rulesets this time, but with lower confidence levels & Back testing these rulesets yielded even lesser precision. On the downside, the confidence factor of the rules mined were lower. But on the upside, we can recommend some changes to the webpages on the Microsoft website as described in section 5.2 and 5.3. Nonetheless, In the real world – We do not have enough data to make page recommendations and may necessarily involve some changes in the webpages in order to collect more useful data and is a continuous improvement process. Same can be said about the results of this exercise.

# Dataset

The data was created by sampling and processing the www.microsoft.com logs. The data

records the use of www.microsoft.com by 38,000 anonymous, randomly‐selected users. For

each user, the data lists all the areas of the web site (Vroots) that the user visited in a one

week timeframe.

1st dataset file: **anonymous-vrootnames-msweb.csv**

The following table is a sample view of the dataset format and header

|  |  |  |
| --- | --- | --- |
| vroot | title | url |
| 1287 | International AutoRoute | /autoroute |
| 1288 | library | /library |
| 1289 | Master Chef Product Information | /masterchef |

“vroot” is the identifier of the page title and the URL of where the page is located, relative to www.microsoft.com

2nd dataset file: **anonymous-msweb.data**

The following table is a sample view of the dataset format without header

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A | 1277 | 1 | "NetShow for PowerPoint" | "/stream" |
| C | 10164 | 10164 |  |  |
| V | 1009 | 1 |  |  |
| V | 1052 | 1 |  |  |

1st column is the attribute type column. Whereby “A” refers to an attribute record, “C” refers to a Case line, in our case here, we use this as a proxy for user id, followed by “V”, which refers to the vote line, which indicate the page where the user visited.

In 2nd column, if 1st column is “A” or “V”, then the number in the 2nd column will be referring to the vroot, which is the ID of a page. If its “C”, then it is a user ID.

3rd column will be ignored since it’s not required.

4th and 5th column refers to the page title and URL respectively.

# Sequence Mining - Methodological Procedures

## Data Pre-processing

The following steps are taken to pre-process the data before we can use the data for any mining purposes.

First, we remove all “A” lines, from the 2nd dataset because it is already in the 1st dataset.

Next, we need to provide an event ID for each vote (“V”) line and also and additional columns to keep track of which case (“C”) line does that vote belongs to. The assumption applied here is that the sequence of the vote is as per the rows entries in the file.

To do the above, we used a for-loop and a temporary value to store the sequence. For each record with “C” in the 1st column, we will reset the event ID number to 1. For each recode without “C” in the 1st column, we will increment the value by 1 to indicate the event sequence.

After the data frame is created, we then match it with the 1st dataset file (anonymous-vrootnames-msweb.csv), which contains all the vroot page name and URL. As we want to keep this information to aid our analysis further on, we will combine these 2 columns into 1 single column, delimited by a pipe “|” symbol.

An extra column “size” is also added, because the sequencing function “read\_basket()” takes in a transaction data type and “size” is required. We have used a fixed number “1” as the size because each record represents 1 Page & URL

Finally, we save the file into a file name as "sequenceDataWithVroot.txt".

The resulting file, "sequenceDataWithVroot.txt”, sample format looks like this:

|  |  |  |  |
| --- | --- | --- | --- |
| Case ID | Event ID | Size | vroot |
| 10001 | 1 | 1 | regwiz|/regwiz |
| 10001 | 2 | 1 | Support Desktop|/support |
| 10001 | 3 | 1 | End User Produced View|/athome |

## Data Analysis

In summary, there are 98654 rows (elements/itemsets/transactions) and 285 columns (items) and a density of 0.003508772

Most frequent items:

|  |  |
| --- | --- |
| Item | Occurrence |
| Free Downloads|/msdownload | 10836 |
| Internet Explorer|/ie | 9383 |
| Microsoft.com Search|/search | 8463 |
| isapi|/isapi | 5330 |
| Products |/products | 5108 |
| (Other) | 59534 |

We are going to mine frequent sequential patterns with the cSPADE algorithm provided by arules sequence library using R. This algorithm utilizes temporal joins along with efficient lattice search techniques and provides for timing constraints.

R code snippets:

*seqs = cspade(transactionalData, parameter = list(support = 0.1), control = list(verbose = TRUE, summary = TRUE))*

Support is an indication of how frequently the itemset appears in the dataset. We set the support to 0.1 as we are interested to find sequence of items that appear in at least 10% of the data set

The algorithm found set of 8 sequence that fulfil the requirements we indicated.

Summary:

|  |  |  |
| --- | --- | --- |
| No of Transactions | No of Sequences | Support |
| 98654 | 32711 | 0.1 |

8 set of sequences:

|  |  |
| --- | --- |
| Sequence | Support |
| {Free Downloads|/msdownload} | 0.3312647 |
| {Internet Explorer|/ie} | 0.2868454 |
| {isapi|/isapi} | 0.1629421 |
| {Microsoft.com Search|/search} | 0.2587203 |
| {Products |/products} | 0.1561554 |
| {Support Desktop|/support} | 0.1360704 |
| {Windows Family of OSs|/windows} | 0.1414815 |
| {Free Downloads|/msdownload},{Internet Explorer|/ie} | 0.1591208 |

Using rule Induction function, we want to find the sequence with the minimum confidence level of 10%

R code snippets:

*rules = ruleInduction(seqs, confidence = 0.1,control = list(verbose = TRUE))*

|  |  |  |  |
| --- | --- | --- | --- |
| Rule | Support | Confidence | Lift |
| {Free Downloads|/msdownload}> => <{Internet Explorer|/ie} | 0.1591208 | 0.4803433 | 1.674572 |

From the results, we can interpret that this sequence has a confidence level of 48% to appear again.

The support is 15%, means that this sequence appears in 15% of the dataset. The lift, which considers both the confidence of the rule and the overall data set, is at 1.6, which means that these two occurrences are likely dependent on one another.

## Recommendations

For the sequence mining results, it seems like use who visited “Free Downloads” page has higher probability of visiting the “Internet Explorer” page. Depending on the objective of the site owner, if it’s to promote a new product, the new product info could be highlighted in one of these 2 pages, since the confident of these 2 pages appearing in sequence in near to 50%. There is a high chance that for every 2 users, 1 of them will see the new product information.

The link between {Free Downloads|/msdownload}> => <{Internet Explorer|/ie} should not remove or if there is not link between the 2 pages, consider adding a link to improve user experience. We should spend effort to discover webpages that do not have sequence. If sequence is expected from business goals, do make changes to those websites and redo the sequence analysis when there is good enough raw data been logged after the changes.

# Association Mining - Methodological Procedures

## Data Pre-processing

For association mining, we are using back the same processed dataset from sequence mining. However, there some further pre-processing required, which is to split the dataset into training dataset (80% of the full dataset) and testing data (20% of the full dataset).

To split the dataset, we use “caret”, a R library, and make use of the createPartion function.

Seed is also set to ensure consistent split of dataset so that we can reproduce the results if necessary.

After splitting the training dataset, we saved into a file named as “basketTrain.txt".

Sample record of the “basketTrain.txt" file:

|  |  |
| --- | --- |
| userid | Vroot |
| 10001 | regwiz|/regwiz |
| 10001 | Support Desktop|/support |

Sample record of the “basketTest.txt" file, which is the same as training dataset:

|  |  |
| --- | --- |
| userid | vroot |
| 10002 | Knowledge Base|/kb |
| 10006 | Knowledge Base|/kb |

## Data Analysis

We used apriori function from “arules” R library to find rules set that consist of association that has minimum support of 10% and confidence level of 10% with minimum association length of 2.

R code snippets:

*rules <- apriori(trainegs, parameter = list(supp=0.1, conf=0.1, minlen=2))*

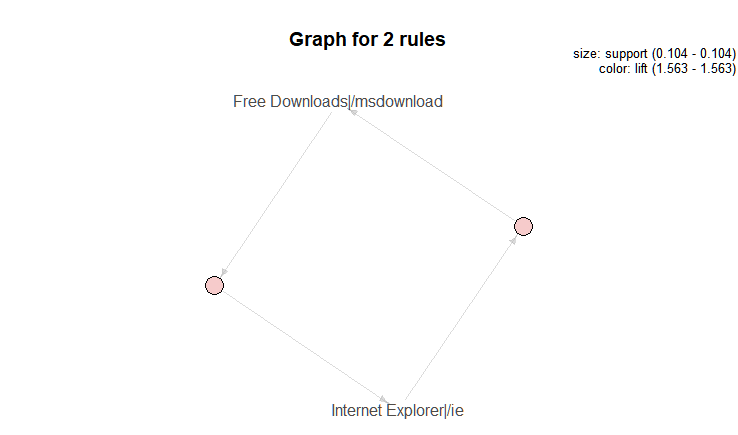
Summary:

|  |  |  |
| --- | --- | --- |
| No of Association | Support | Confidence |
| 16403 | 0.1 | 0.1 |

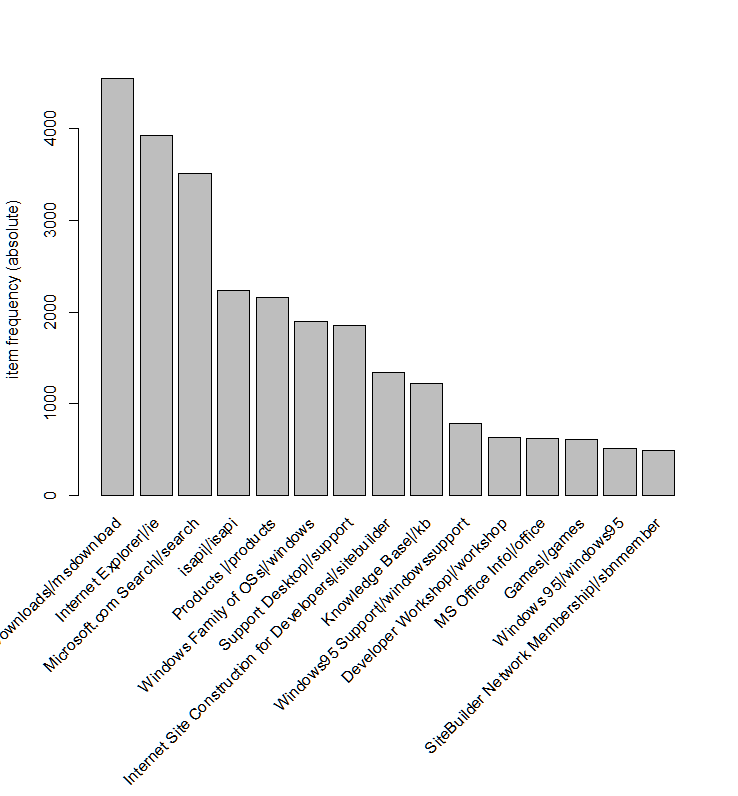
2 rulesets are found as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Internet Explorer|/ie} =>  {Free Downloads|/msdownload} | 0.1038225 | 0.433554 | 1.563329 |
| {Free Downloads|/msdownload} =>  {Internet Explorer|/ie} | 0.1038225 | 0.374368 | 1.563329 |

Given the above results, it appears that both rules have the same lift. However, the confidence for the first rule is higher, at 43%



We are also interested to find out the item frequency of the item appearing in training dataset.



We then proceed to make prediction with the test data by executing the rules against test data

Prediction Results:

**Precision= 11.31973 Correct Predictions = 416 Total Predictions= 3675**

Precision is the probability that a (randomly selected) retrieved document is relevant. Our results show that out of the total prediction made, 416 was predicted correctly.

We then decided to set a lower support to see if results will differ with a better prediction.

R code snippets:

*rules <- apriori(trainegs, parameter = list(supp=0.04, conf=0.1, minlen=2))*

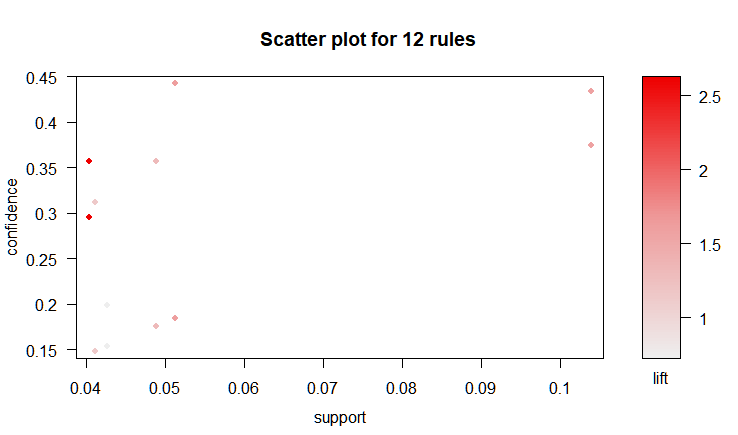
12 Ruleset found:

|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Support Desktop|/support} => {isapi|/isapi} | 0.04029751 | 0.3572973 | 2.6187433 |
| {isapi|/isapi} => Support Desktop|/support} | 0.04029751 | 0.295353 | 2.6187433 |
| {Windows Family of OSs|/windows} => {Free Downloads|/msdownload} | 0.05121014 | 0.4428044 | 1.5966852 |
| {Free Downloads|/msdownload} => {Windows Family of OSs|/windows} | 0.05121014 | 0.184656 | 1.5966852 |
| {Internet Explorer|/ie} => {Free Downloads|/msdownload} | 0.10382247 | 0.433554 | 1.5633295 |
| {Free Downloads|/msdownload} => {Internet Explorer|/ie} | 0.10382247 | 0.374368 | 1.5633295 |
| {isapi|/isapi} => {Free Downloads|/msdownload} | 0.04877157 | 0.357462 | 1.2889535 |
| {Free Downloads|/msdownload} => {isapi|/isapi} | 0.04877157 | 0.1758628 | 1.2889535 |
| {Products |/products} => {Free Downloads|/msdownload} | 0.04102908 | 0.3117184 | 1.124009 |
| {Free Downloads|/msdownload} => {Products |/products} | 0.04102908 | 0.1479446 | 1.124009 |
| {Microsoft.com Search|/search} => {Free Downloads|/msdownload} | 0.04261416 | 0.1990319 | 0.7176786 |
| {Free Downloads|/msdownload} => {Microsoft.com Search|/search} | 0.04261416 | 0.1536601 | 0.7176786 |

Prediction Results:

**precision= 6.968808 Correct Predictions = 1164 Total Predictions= 16703**

The results aren’t as good, but we can see more rules with a higher lift. Which may be useful for recommending possible changes to the pages in Microsoft website.



Using the scatter plot can quickly help us identify interesting items/association

The 2 red dots on the most left:

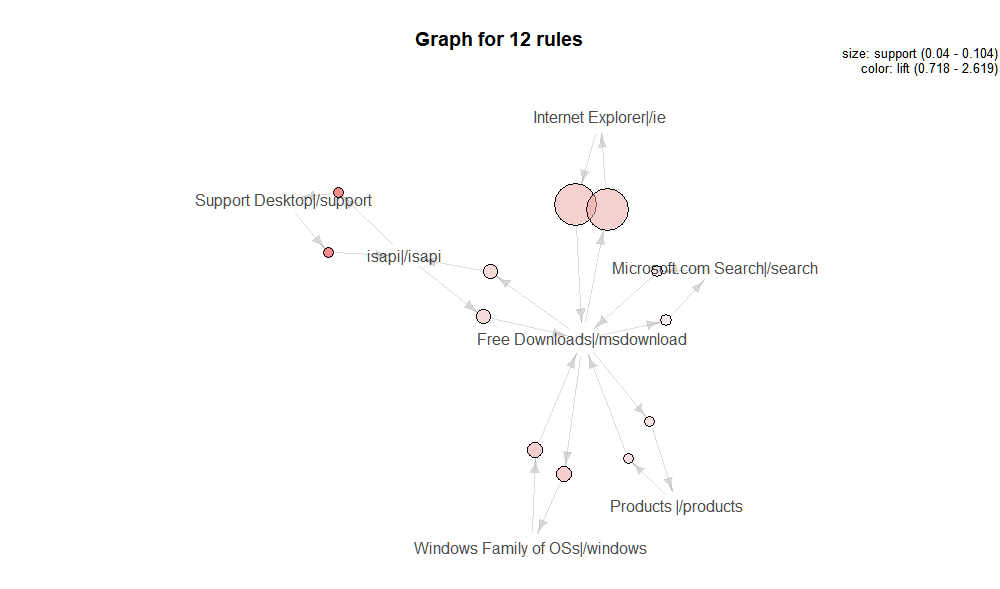
|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Support Desktop|/support} => {isapi|/isapi} | 0.04029751 | 0.3572973 | 2.618743 |
| {isapi|/isapi} => {Support Desktop|/support} | 0.04029751 | 0.2953530 | 2.618743 |

The top right most dim red dots:

|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Internet Explorer|/ie} =>  {Free Downloads|/msdownload} | 0.1038225 | 0.433554 | 1.563329 |
| {Free Downloads|/msdownload} =>  {Internet Explorer|/ie} | 0.1038225 | 0.374368 | 1.563329 |

The top most left dot:

|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Windows Family of OSs|/windows} =>  {Free Downloads|/msdownload} | 0.05121014 | 0.4428044 | 1.596685 |

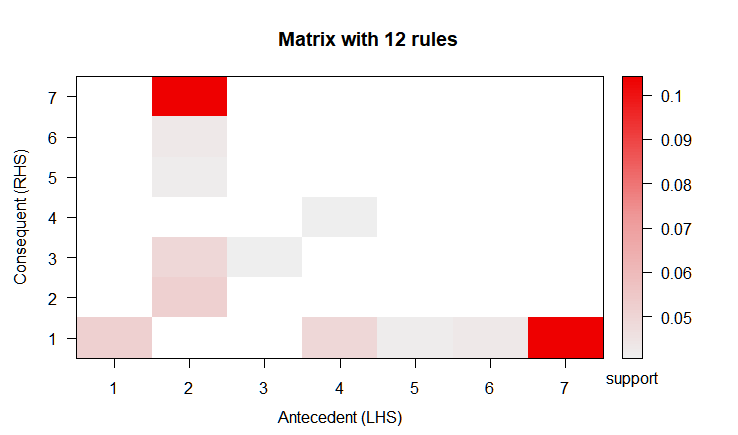


From the graph above, we can see the biggest size, which represent higher support.

But what we are interested at is the redder colours, which, means higher lift.

The 2 red dots are:

|  |  |  |  |
| --- | --- | --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support | Confidence | Lift |
| {Support Desktop|/support} => {isapi|/isapi} | 0.04029751 | 0.3572973 | 2.618743 |
| {isapi|/isapi} => {Support Desktop|/support} | 0.04029751 | 0.2953530 | 2.618743 |



|  |  |
| --- | --- |
| No. | Itemsets in Antecedent (LHS) |
| 1 | {Windows Family of OSs|/windows} |
| 2 | {Free Downloads|/msdownload} |
| 3 | {Support Desktop|/support} |
| 4 | {isapi|/isapi} |
| 5 | {Products |/products} |
| 6 | {Microsoft.com Search|/search} |
| 7 | {Internet Explorer|/ie} |

|  |  |
| --- | --- |
| No. | Itemsets in Consequent (RHS) |
| 1 | {Free Downloads|/msdownload} |
| 2 | {Windows Family of OSs|/windows} |
| 3 | {isapi|/isapi} |
| 4 | {Support Desktop|/support} |
| 5 | {Products |/products} |
| 6 | {Microsoft.com Search|/search} |
| 7 | {Internet Explorer|/ie} |

Top red:

|  |  |
| --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support |
| {Free Downloads|/msdownload} => {Internet Explorer|/ie} | 0.1038225 |

Bottom red:

|  |  |
| --- | --- |
| Left hand side (lhs) => Right hand side (rhs) | Support |
| {Internet Explorer|/ie} => {Free Downloads|/msdownload} | 0.1038225 |

## Recommendations

Although the first results look better, based on precision, however, the ruleset is only 2, which can be limiting in terms of the associated pages and recommendation, which is the same as the results from sequence mining.

For our next test when we lowered the support parameters, we got 12 rulesets. we managed to mine more data. Which is a good thing because there are now more association between pages.

There is high lift, at 2.6, between “Desktop Support” page and “ïsapi” page, which indicate the dependency of these 2 pages are higher than others. “Free Downloads” page also appears among most of the ruleset with more than 30% confidence level. One of the recommendations would be showing more download page link to users at pages with high lift. This will try to drive users to stay on site more as 30% - 40% of users tend to visit “Free Download” pages, and by doing so, since there are many page association with “Free Download” pages, this will intern increase user’s time on site, provided the objective is to keep user engaged on site by showing more relevant content/pages.

# Conclusion

We believe more extensive sequence mining and association mining can be achieved if there was more data. Currently data is only 1 weeks’ worth of data. However, given more data would also means more computing power is required. For the current dataset, it took about 30-45 mins for us to pre-process the data into the format we need in the for-loop function. This might be due to the lower specification of the laptop used. If computational power permits, we could event tweak more changes to the support and confidence level to mine more page association to help provide more recommendation.

The output of the various graphs and tables will enable stakeholders to better visualise the web traffic in log format. Hence allows them to ask questions among themselves to formulate changes and make the website successful. Likewise, the key metrics recorded from the first analysis can be compare after the website is improved against the sequence analysis of the website.