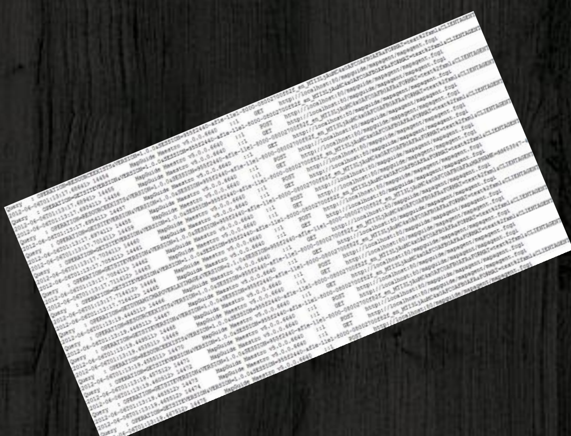
Analysis of DULS using Web Usage Mining

Contents

Introduction

- What are Web Logs?
- What is Web Usage Mining?
- Usefulness in the case of library logs
- Some important terminologies

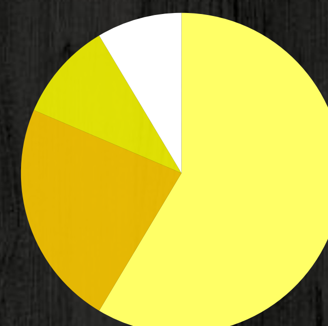


Work Done

- Aim of the experiments
- Data Preparation
- Experiment no. 1
- Experiment no. 2

Results

- Results of Exp 1
- Results of Exp 2
- Future scope
- References



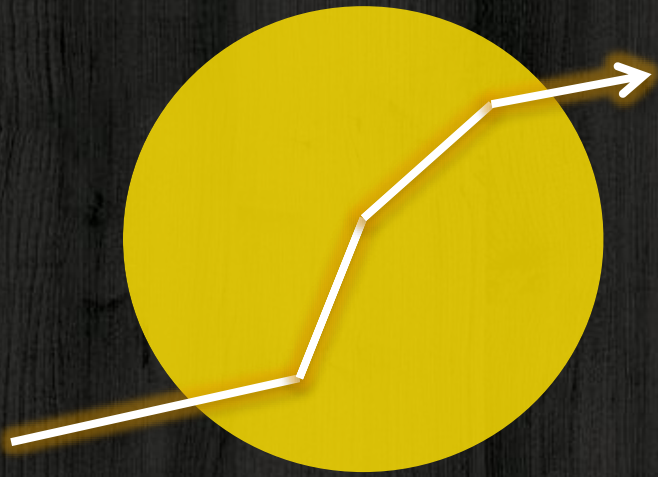
Introduction

What, Why and How of the project

Web Access Logs

What are Web Access Logs?

An **access log** is a list of all the requests for individual files that people have requested from a **Web** site.



Why do we need them?

- Pattern Analysis
- Trend Predictions

What is Web Usage Mining

Data Collection

Pre-processing

Pattern Discovery

Analysis and Results

Digital Library's Web Log Analysis

Insights we can gather by this analysis

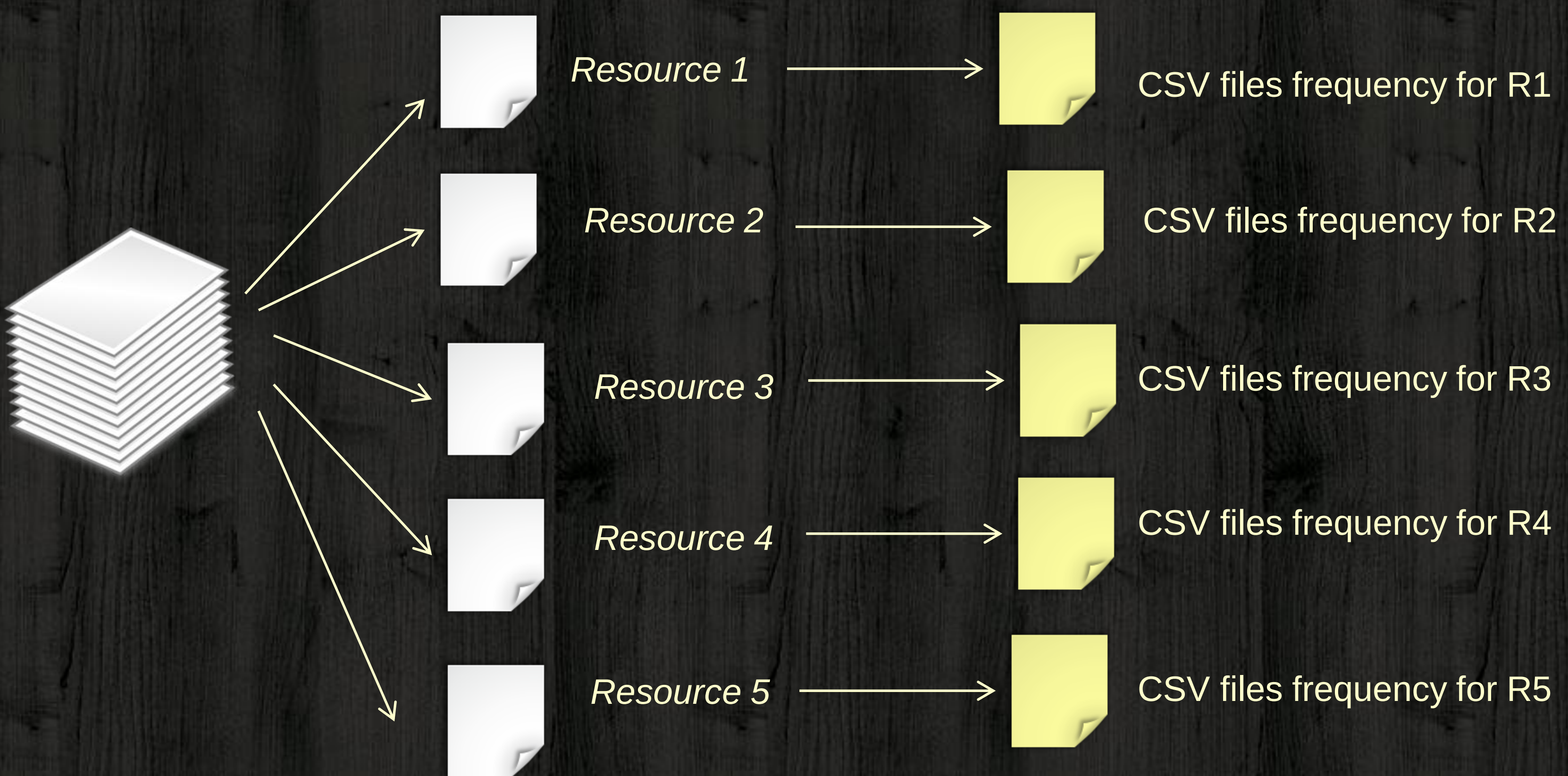
- **Most used e-resources:** Can tell us which e-resources we need to continue subscribing.
- **Least used e-resources:** We can work on how to make these resources popular.
- **Consistently used e-resources**: Whether a resources is used only for some time or is always/constantly used.
- **Is there any link or a pattern in the way the resources are been accessed:** Which resources are accessed together.

Work Done

Experiments on Web Logs of Delhi Unversity Library System (DULS)

Data Pre-Processing

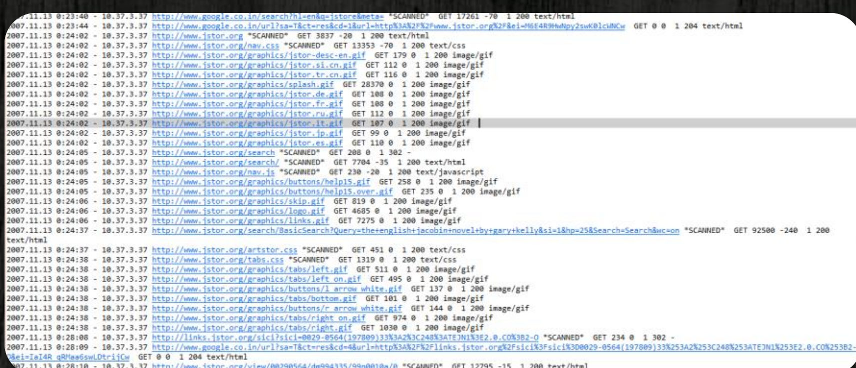
Data log format: The format of the log files was DansGuardian. The fields that are present are as follows: Date, Time, Requesting IP address, Complete requested URL, Actions, Methods, Size and HTTP status return code.



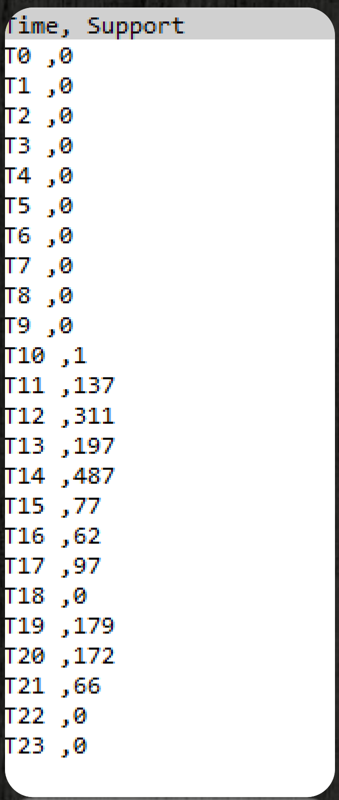
Using UNIX commands **Using UNIX commands**

Data Pre-Processing

Raw logs



Resource specific log file



```
Time, Support
T0 ,0
T1 ,0
T2 ,0
T3 ,0
T4 ,0
T5 ,0
T6 ,0
T7 ,0
T8 ,0
T9 ,0
T10 ,1
T11 ,137
T12 ,311
T13 ,197
T14 ,487
T15 ,77
T16 ,62
T17 ,97
T18 ,0
T19 ,179
T20 ,172
T21 ,66
T22 ,0
T23 ,0
```

CSV file, with time and frequency

Ranking of e-resources

Resources	Average Frequency
Jstor	112
Springer	74
Scopus	50
Nature	41
Inter science	30
Ebsocohost	7
Oxford	6
Emerald Insight	2.4
Portal.acm	1.29

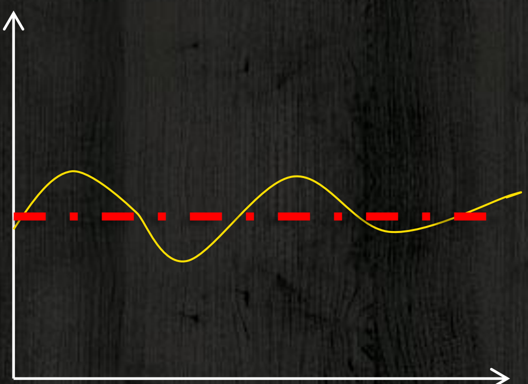
- First, all the files were imported into excel.
- Then, their average frequency was calculated.
- Finally, the resources were ranked accordingly.

Why is studying only frequency not enough???

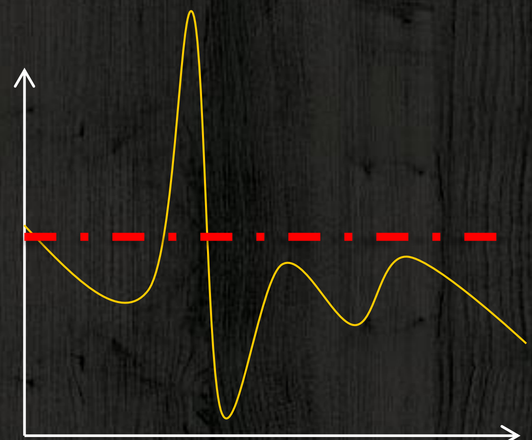
Why is Frequency not enough?

Taking average of frequency has two drawbacks

- It gets biased by peaks and pits of the graphs
- It is unable to account for fluctuations.



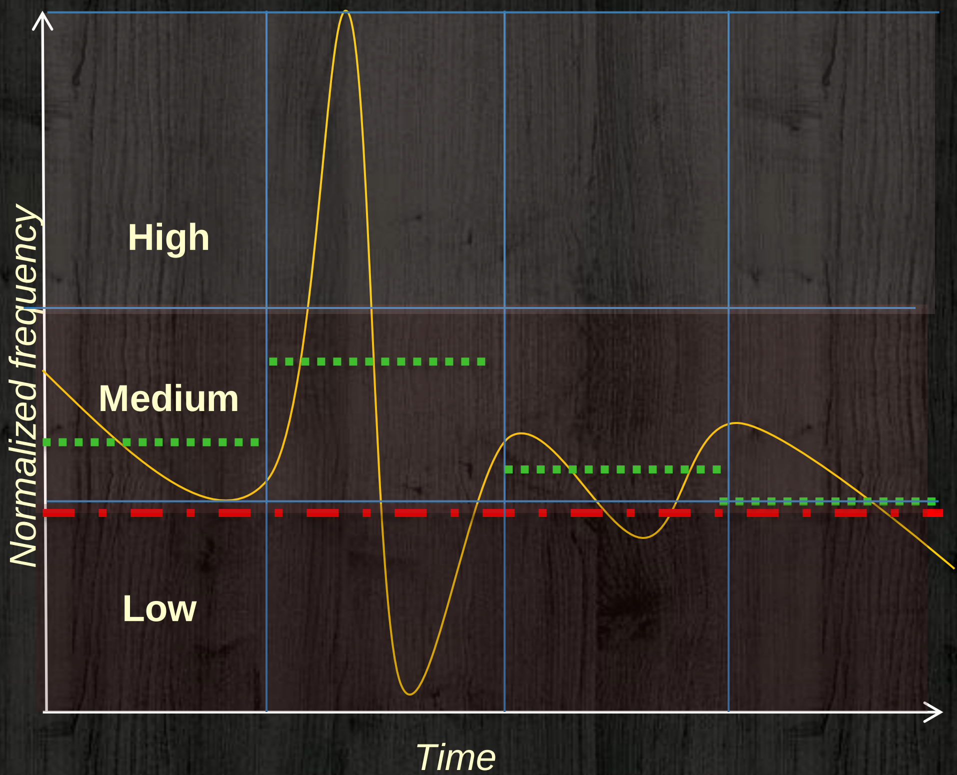
Resource 1



Resource 2

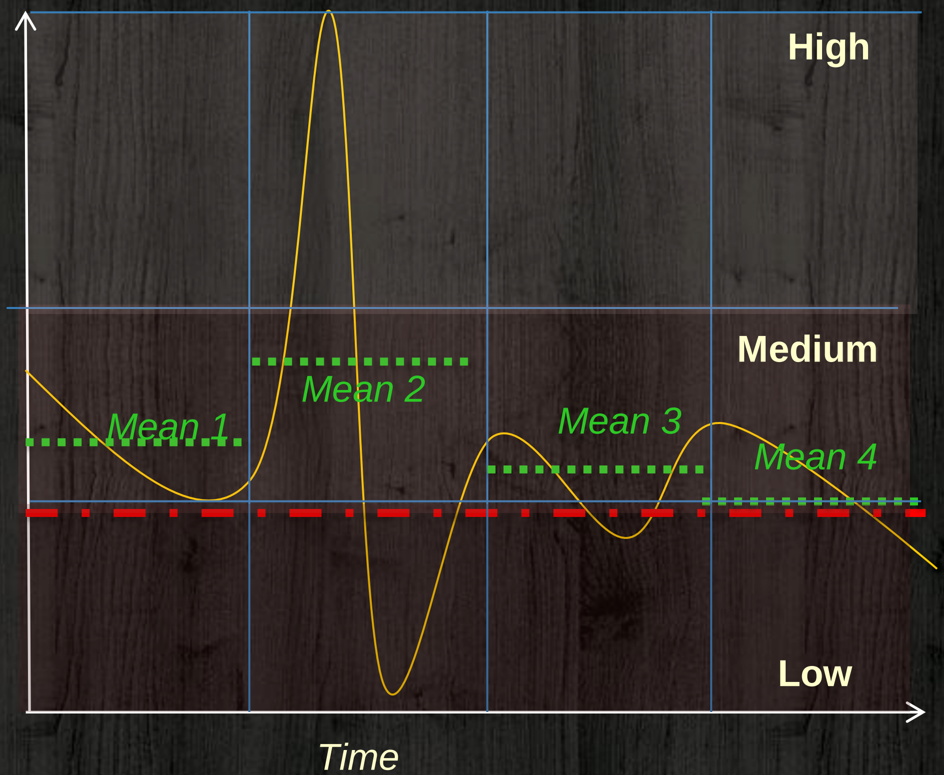
We can see that resource 2 has a higher overall average as compared to resource 1 but, when we are deciding about e-resources, we should not say that resource 2 is better than resource 1. This is because our resource should be as consistent as possible, and should not be biased by a few very high instances and since, frequency analysis fails to take this into account, we believe it is not good enough.

Frequency and Fluctuation Analysis



- Y axis: Normalized usage frequency
- X axis: Time (hourly)
- The time has been divided into some time windows (*So that we can isolate parts of the graph where fluctuations are high).*
- *The frequency has been divided into thresholds (for low, medium and high usage).*
- *Calculated mean per time window.*
- *Also, calculated average standard deviation per time window.*
- *Spread is the number of partitions our graph has been to in a specific time window.*

Frequency and Fluctuation Analysis



Spread=1	Spread=3	Spread =2	Spread =2
SD=low	SD=high	SD= low	SD=low
(OK)	(PENALTY)	(OK)	(OK)

Fluctuation: A true fluctuation occurs when the graph crosses the thresholds and roams in more than one partition and at the same time, has a high standard deviation.

$$\text{RESULT} = \frac{\text{Mean1}+(\text{Mean2}/3)+\text{Mean3}+\text{Mean4}}{4}$$

Ranking of e-resources

Resource	(Frequency + Fluctuation)
Jstor	0.23118
Springer	0.1043
Scopus	0.10015
Nature	0.0616
Inter science	0.0604
Ebsocohost	0.0143
Oxford	0.01292
Emerald Insight	0.00505
Portal.acm	0.0026

- Now, we have the ranks that are influenced by frequency as well as fluctuations in the graph.

Important Terminologies

- **Support:** Number of times a rule “did” occur divided by the number of observations in the dataset.
- **Support Count:** Number of occurences of an item.
- **Association rule**: They are used for finding patterns. These rules are basically “If/then” statements that help us uncover relationships amongst dataset elements.
- **Confidence:** It measures how positive we are that if one attribute is flagged true, then the other to-be-associated attribute will also be flagged true.
- **FP Growth:** FP=Frequent Pattern; It aims to find out the most frequent “itemset” from a dataset. Necessarily used for finding association rules.

Association Rules

MOTIVE

- To look for patterns in usage.
- This experiment tells us which resources were accessed along with which resource.

HOW

- Software used : Rapid Miner
- Algorithm : Association Rule operator provided by s/w along with FP growth rule.

Minimum Confidence

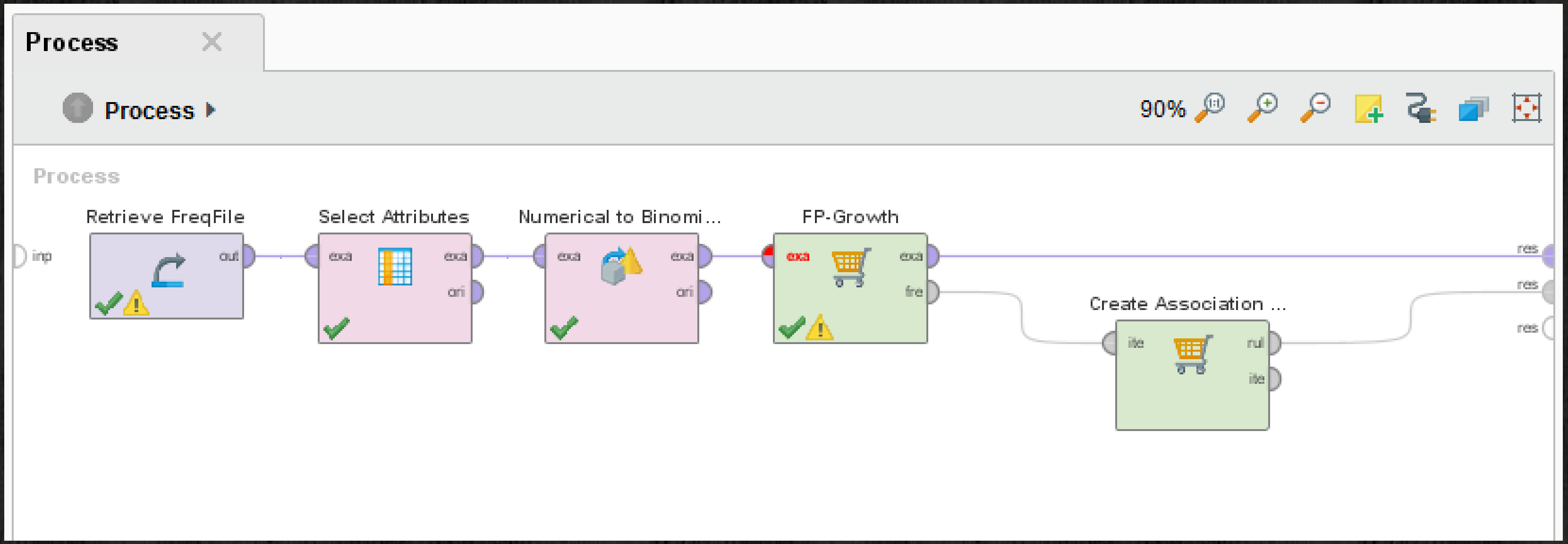
Minimum Support

Attributes

FP Growth

Association Rules

Creating Association Rules

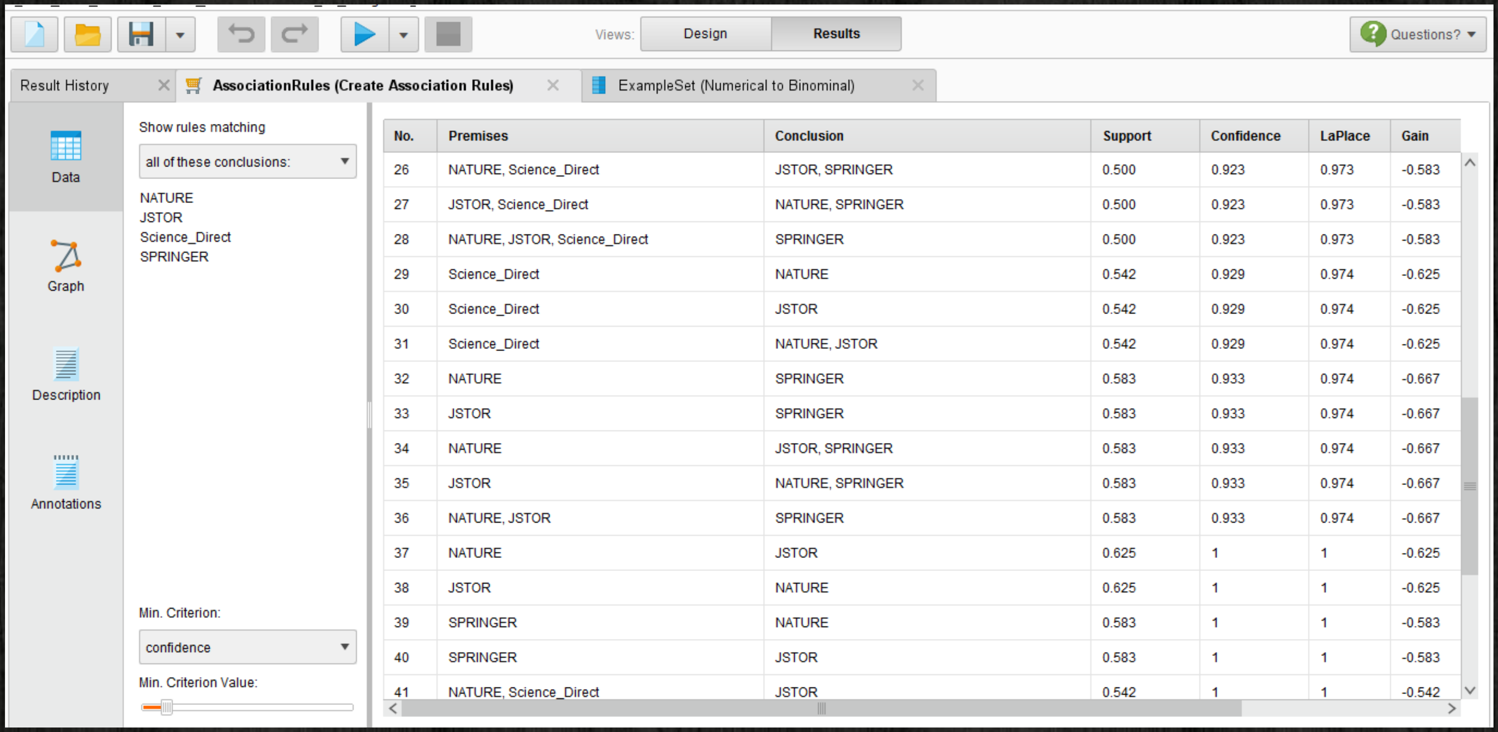


Model of miner to create association rules

- **This model helps us select some/all attributes that we want to analyze.**
- **Uses FP growth algorithm and the operators provided by the Rapid Miner.**
- **Five combinations of attributes, minimum support and minimum confidence were experimented with in the project. Two of them are discussed in the next slides.**

Condition Set 2

Conditions set: All attributes selected; Minimum support=0.5; Minimum confidence=0.6; Minimum number of item sets- Disabled.



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain
26	NATURE, Science_Direct	JSTOR, SPRINGER	0.500	0.923	0.973	-0.583
27	JSTOR, Science_Direct	NATURE, SPRINGER	0.500	0.923	0.973	-0.583
28	NATURE, JSTOR, Science_Direct	SPRINGER	0.500	0.923	0.973	-0.583
29	Science_Direct	NATURE	0.542	0.929	0.974	-0.625
30	Science_Direct	JSTOR	0.542	0.929	0.974	-0.625
31	Science_Direct	NATURE, JSTOR	0.542	0.929	0.974	-0.625
32	NATURE	SPRINGER	0.583	0.933	0.974	-0.667
33	JSTOR	SPRINGER	0.583	0.933	0.974	-0.667
34	NATURE	JSTOR, SPRINGER	0.583	0.933	0.974	-0.667
35	JSTOR	NATURE, SPRINGER	0.583	0.933	0.974	-0.667
36	NATURE, JSTOR	SPRINGER	0.583	0.933	0.974	-0.667
37	NATURE	JSTOR	0.625	1	1	-0.625
38	JSTOR	NATURE	0.625	1	1	-0.625
39	SPRINGER	NATURE	0.583	1	1	-0.583
40	SPRINGER	JSTOR	0.583	1	1	-0.583
41	NATURE, Science_Direct	JSTOR	0.542	1	1	-0.542

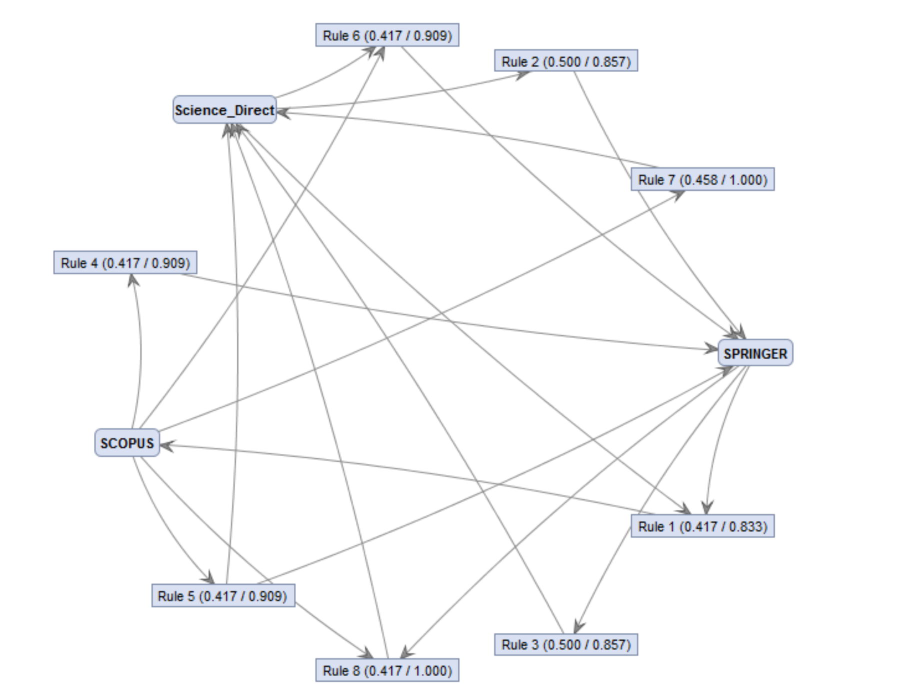
Condition Set 2: Association Rules obtained

Conditions set: All attributes selected; Minimum support=0.5; Minimum confidence=0.6; Minimum number of item sets- Disabled.

- [NATURE] --> [Science_Direct, SPRINGER] (confidence: 0.800)
- [JSTOR, Science_Direct] --> [SPRINGER] (confidence: 0.923)
- [NATURE, JSTOR, Science_Direct] --> [SPRINGER] (confidence: 0.923)
- [NATURE] --> [JSTOR, SPRINGER] (confidence: 0.933)
- [NATURE] --> [JSTOR] (confidence: 1.000)[JSTOR] --> [NATURE] (confidence: 1.000)[SPRINGER] --> [NATURE] (confidence: 1.000)[SPRINGER] --> [JSTOR] (confidence: 1.000)[NATURE, Science_Direct] --> [JSTOR] (confidence: 1.000)
- [NATURE, SPRINGER] --> [JSTOR] (confidence: 1.000)[JSTOR, SPRINGER] --> [NATURE] (confidence: 1.000)[Science_Direct, SPRINGER] --> [NATURE] (confidence: 1.000)
- [Science_Direct, SPRINGER] --> [JSTOR] (confidence: 1.000)[Science_Direct, SPRINGER] --> [NATURE, JSTOR] (confidence: 1.000)
- [NATURE, Science_Direct, SPRINGER] --> [JSTOR] (confidence: 1.000)
- [JSTOR, Science_Direct, SPRINGER] --> [NATURE] (confidence: 1.000)

Condition Set 4

Conditions set:
Four attributes selected- ACM, Springer, Scopus, Science Direct; Minimum support=0.95; Minimum confidence=0.8; Minimum number of item sets= Enabled and set to 100.



Condition Set 4: Results

Conditions set: Four attributes selected- ACM, Springer, Scopus, Science Direct; Minimum support=0.95; Minimum confidence=0.8; Minimum number of item sets- Enabled and set to 100.

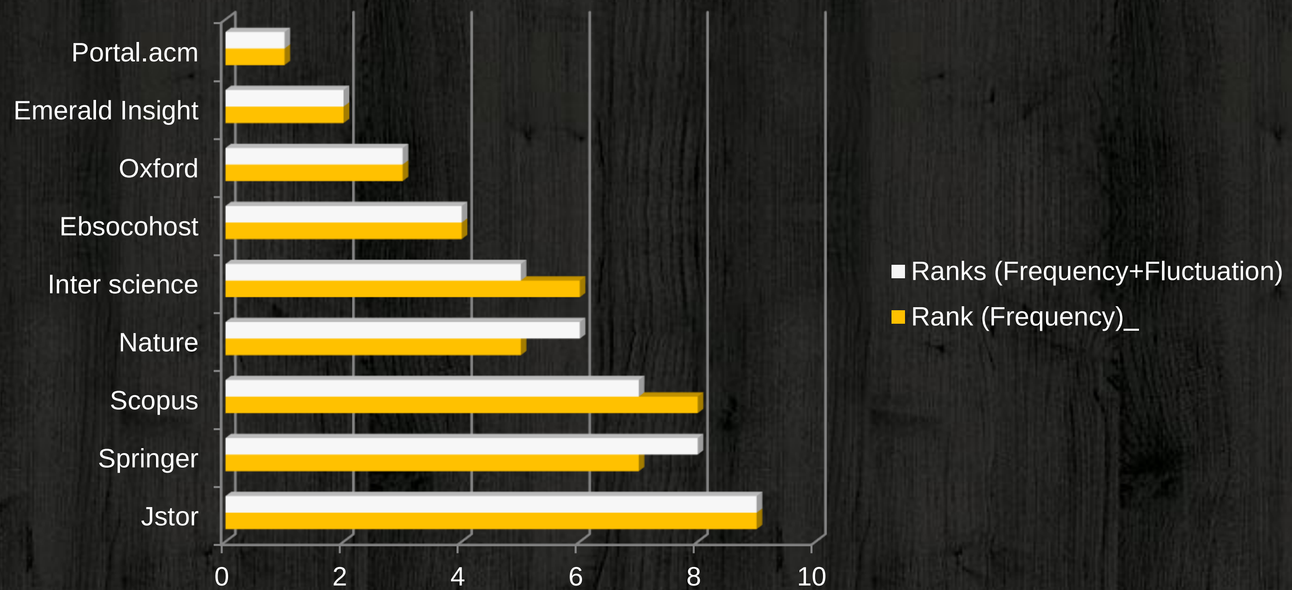
- [Science_Direct, SPRINGER] --> [SCOPUS] (confidence: 0.833)
- [Science_Direct] --> [SPRINGER] (confidence: 0.857)
- [SPRINGER] --> [Science_Direct] (confidence: 0.857)
- [SCOPUS] --> [SPRINGER] (confidence: 0.909)
- [SCOPUS] --> [Science_Direct, SPRINGER] (confidence: 0.909)
- [Science_Direct, SCOPUS] --> [SPRINGER] (confidence: 0.909)
- [SCOPUS] --> [Science_Direct] (confidence: 1.000)
- [SPRINGER, SCOPUS] --> [Science_Direct] (confidence: 1.000)

Result

Final Results and Conclusions

Results

- As mentioned before, study of fluctuation gives us a better idea about consistency of the resources. They do not allow our results to get biased by outliers. The first experiment gave us the following ranking for frequency and frequency with fluctuations.



- Identifying patterns like these can guide us in subscriptions and even their placements. If used wisely we may even be able to affect the crawling behaviour of users. This can also help us in bringing attention to other e-resources that are not accessed so frequently.

Future Scope

Web logs definitely prove to be helpful in this analysis. But, we believe that a better system could still be proposed where the evaluation is continuous in nature and the parameters for analysis are not merely based on the number of times resources have been used. As our results depicted, the fluctuations and frequency together if studied with real focus can be much more insightful. Also, the association rules found might offer a great deal of help when put to use for predicting user needs and in building recommender systems. Marveling with such minute details is what modern day data studying requires.

A good qualitative analysis is the future of mining. Bringing together all dependent features, such as pattern fluctuation of frequency in our case, might seem tedious but it would be much perceptive than state-of-the-art systems and techniques.

Thank You

Presentation by :
URVASHI CHOUDHARY
VASUNDHRA DAHIYA