

# Data Preparation for User Profiling

Marek Kumpošt

DATAKON, 20-23.10.2007  
Hotel Santon, Brno

Faculty of informatics  
Masaryk university  
Brno



# Contents

- 1 Introduction
- 2 Ways to filter data
- 3 Ideas for future work

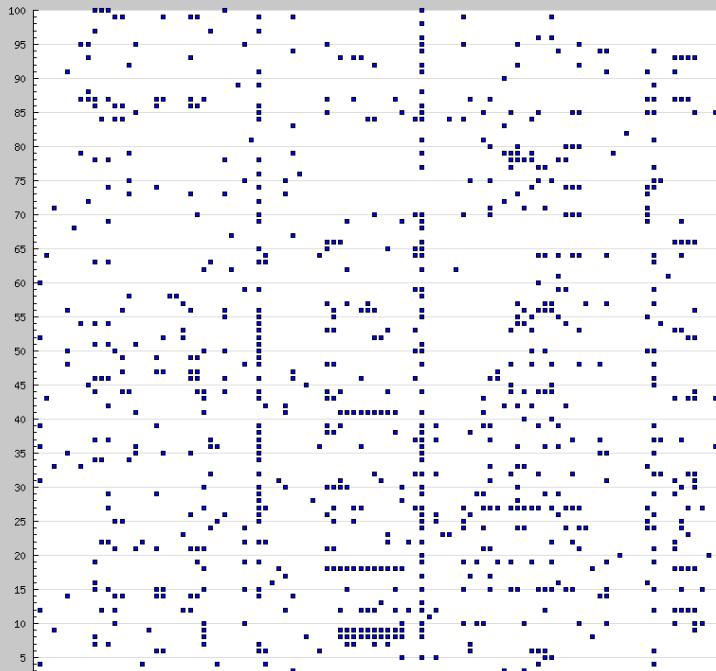
# Introduction of the project

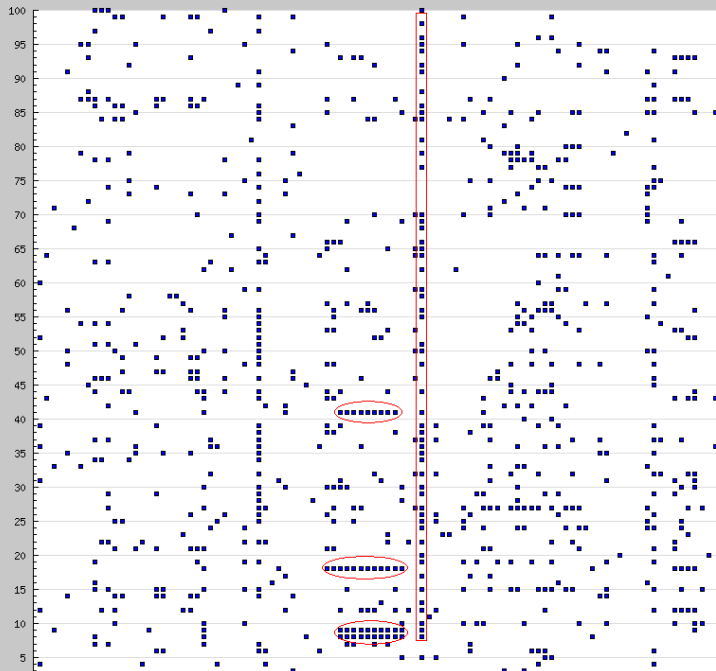
- User behaviour model (context model)
  - ▶ User profiling based on their previous behaviour (context)
  - ▶ PATS graph model for modeling context information
- Representative behavioral patterns
  - ▶ Identification of groups with the same behavioral characteristics
  - ▶ Try to identify user(s) based on their behavioral patterns only
- Impact on **users' privacy** (ISPs have huge traffic databases available)
  - ▶ e.g. AOL search query data set
- Techniques for finding behavioral characteristics
  - ▶ Input data restriction and optimization
  - ▶ Processing data (appropriate input information; data mining techniques)
  - ▶ Results evaluation → impacts on users' privacy
- Input data – Netflow MU (traffic log)

# Introduction of the project – cont.

- Input restriction – selected part of a network; selected ports (Faculty of informatics and college; port 80, 22)
  - ▶ find most frequently visited destination IPs
    - ★ best ratio between source and destination IPs?
    - ★ techniques that help to clear the data
  - ▶ for every source IP find the number of hits to a particular destination
- Output is the matrix source vs. destination IPs and hits
  - ▶ we have vectors describing “behaviour” of source IPs
  - ▶ input data for the clustering process
  - ▶ matrix is very sparse :-)
- Approaches to limit the number of context information and entities
  - ▶ omit very frequently visited destinations
  - ▶ omit commonly visited destinations
  - ▶ omit very active source IPs
  - ▶ restriction of IP addresses (src/dest) and port
- Input data visualization
  - ▶ to visually detect some characteristics

Scatter plot (X=dest IPs; Y=source IPs), 819 points





# Ways to filter input data

## How to find relevant source and destination IPs?

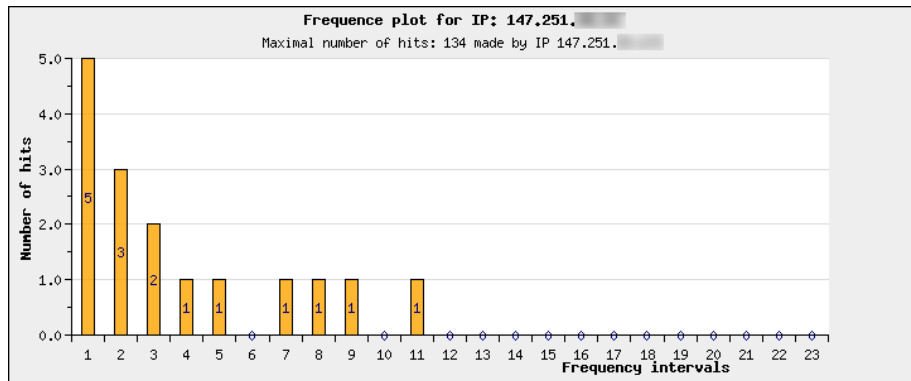
We need more dense matrix for the clustering process

- Destination IPs restrictions
  - ▶ accessed only once within a given period
  - ▶ accessed by at least a half of sources
  - ▶ different levels of entropies – number of unique sources
  - ▶ TF-IDF (text mining field), PrefixSpan (sequence based mining)
- Usage-based vs. frequency-based approach
  - ▶ usage-based – to optimize destinations
  - ▶ frequency-based – to optimize sources
- Visualization of the matrix of vectors
  - ▶ scatter plot (usage-based)
  - ▶ balloon plot (frequency-based)
- Source IPs restrictions
  - ▶ only “active” sources may help in clustering (profiling)
  - ▶ behaviour of passive sources is difficult to predict
  - ▶ differentiate between different levels of “activity”

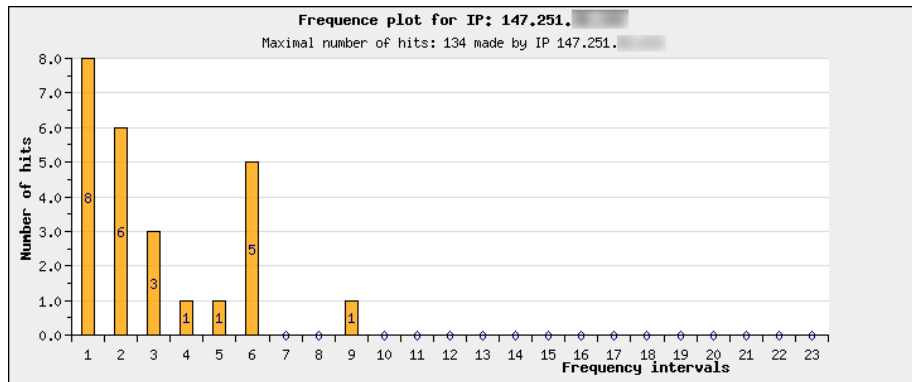
# Frequency histograms clustering

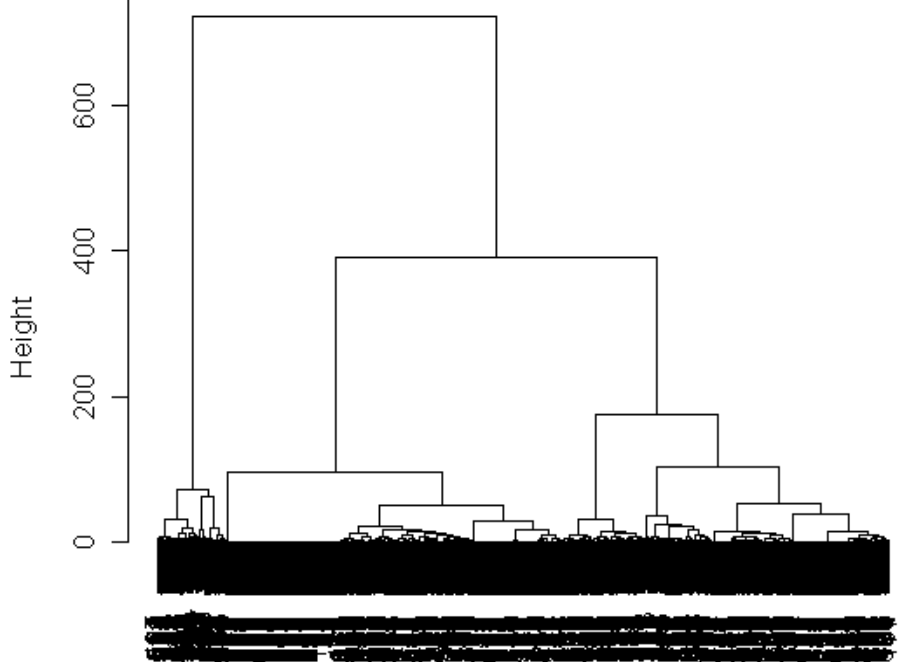
- Frequencies of source IPs activities
  - ▶ levels of frequencies and number of accessed destinations
  - ▶ 1 to 10 individually and then aggregations of tens
  - ▶ most records fall into these individual categories
- Helps to find different levels of activity
- Helps to decrease the matrix dimensions
  - ▶ process of clustering is partially automatic
    - ★ find histograms
    - ★ save vectors into arff file
    - ★ use R to perform clustering and cut clusters to sets
  - ▶ Ward's clustering method
    - ★ minimizes the 'information loss' associated with each grouping
    - ★ strong tendency to split data in groups of roughly equal size
    - ★ no clusters with only one or a few elements
    - ★ output levels of activity are used as a restriction

# Histogram visualization and processing

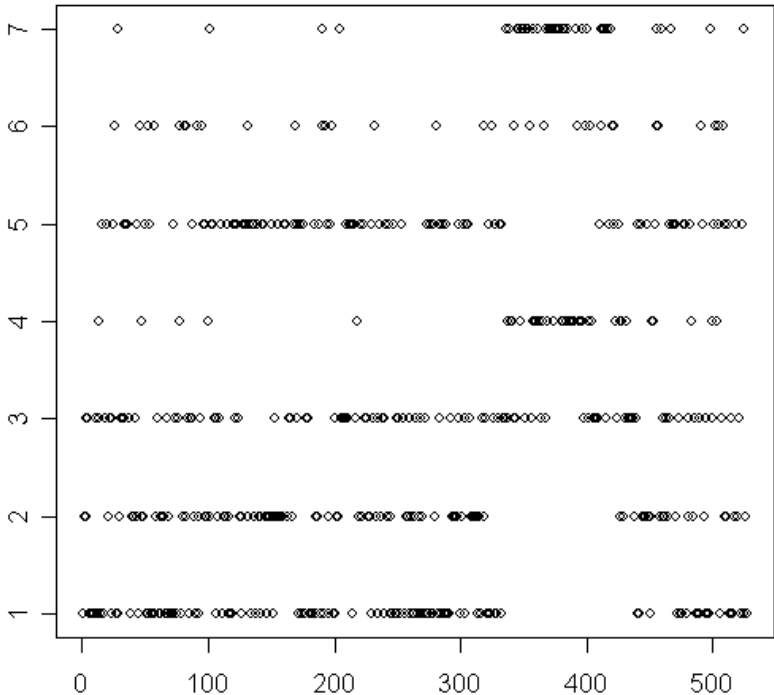


# Histogram visualization and processing





Cluster



# TF-IDF (Term Frequency - Inverse Document Frequency)

- TF-IDF – how important a word is to a document in a collection or corpus
- IDF – how important a destination is to a set of source IPs ( $\log_2(n/d_{fi})$ ) – dest. IP is a term, source IP is a document
- Negative correlation with entropy of destinations
- Sorting destinations according to IDF and entropy
- IDF for destinations with same number sources is the same
- Entropy varies according to the number of visits
- Entropy is higher if the distribution of visits is more even
- Combination of IDF and Entropy helps to select destinations
- ... with “more evenly distributed” hits
- This approach helps to distinguish different “levels of interest”

# PrefixSpan

- Sequence mining algorithm
- Searching for frequent sequences of destinations
- Sequences can contain gaps (how long?)
- Destinations ordering – IP value
- Input: sequences of destinations for each source
- Output: frequent sequences w.r.t prefixspan settings
- Frequent sequences can be processed individually
- ... to find corresponding sources
- Sources can be analyzed with more data
- Problems with proxies and very active sources

```
./prefixspan -m 2 -M 5 <sequences.txt >output.txt
-m NUM:      set minimum support
-M NUM:      set minimum pattern length
-L NUM:      set maximum pattern length
-a:          print ALL patterns (default: print longest pattern)
```

# Ideas for the future

- Graph mining algorithms
- BiClustering (or two-mode clustering)
  - ▶ simultaneous clustering of rows and columns
  - ▶ bicluster – a subset of rows that exhibit similar behavior across a subset of columns
- Implementation of “maxGap” restriction to the PrefixSpan
- Build a global profile and count distances of individual users
- Further automatization of the process
- Combinations of approaches together
- Evaluate successfulness of the model
  - ▶ use the filtered data with the PATS graph model

Any questions?

Thanks for your attention!

kumpost@fi.muni.cz