

# Applied Data Mining

Ingo Lütkebohle, Julia Lüning

21. Chaos Communication Congress

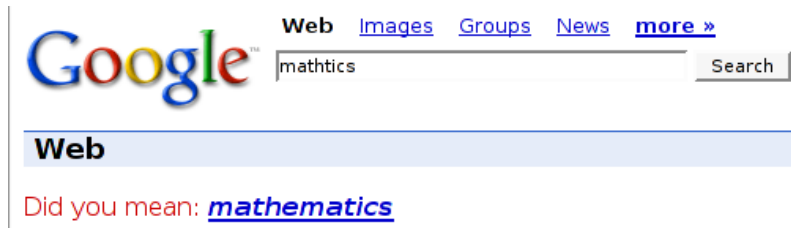
27.12.2004

# Outline

- 1 introduction
  - motivation
  - process of mining data
- 2 features
  - visualisation
- 3 exploration
  - statistics
  - clustering
- 4 Association Rule Mining
  - algorithm
  - tool
  - example

# Google Query Suggestion

Find *similar* words with more *hits*.



The image shows a screenshot of the Google search homepage. The Google logo is on the left. To its right are navigation links: [Web](#), [Images](#), [Groups](#), [News](#), and [more »](#). Below these is a search input field containing the text "mathctics" and a "Search" button. A horizontal line separates the search area from the results area. Below the line, the word "Web" is displayed in a light blue bar. Underneath, the text "Did you mean: [mathematics](#)" is shown, where "mathematics" is underlined and blue.

# Amazon Recommender System

Item shown: Holy Bible, King James Version

## Customers who bought this book also bought:

- [Holy Bible King James Version Study Bible \(Burgundy\)](#) by [Not Applicable \(Na \)](#)
- [The Holy Quran: An English Translation](#) by [Allamah Nooruddin](#)
- [The Torah](#) by [Rodney, Rabbi Mariner](#)
- [The Qur'an Translation](#) by [Abdullah Yusuf Ali](#)
- [The Holy Bible: King James Version](#) by [Not Applicable \(Na \)](#)

Association Rule:  $A \leftarrow B$

# process of mining data

first of all: define your *objective*  
then:

- 1 data collection
- 2 feature extraction
- 3 data cleaning
- 4 exploration — summaries, clustering
- 5 rule mining and/or classification

# types of attributes

very simple world view:

**binary** true, false; present, not present

**nominal** blue, red, green

**ordinal** drizzle < rain < torrent

**numeric** 4.45, 5.76, 19.33

# features

## data mining on eMail:

- bag of words
- length of the mail (number of words)
- number of recipients
- date — epoch, week number, daytime, ...
- ...

# text mining

- many, infrequently occurring features (words)
- one word, many meanings
- one meaning, many words
- → extensive preprocessing necessary



# merging

- aggregating more of the same  
example:
- joining different feature spaces  
example: pgp signature data and event data → who met  
who at which key signing party

```
> DAYS <- data.frame(day=c("Monday", "Tuesday",  
...), num=c(1, 2, ...))  
> SCHEDULE <- data.frame(SPK=("Sven", "Mitch",  
...), daynum=c(2, 2, ...))  
> merge(SCHEDULE,DAYS, by.x="daynum", by.y="num")  
num day SPK  
1 2 Tuesday Sven  
2 2 Tuesday Mitch
```

# data formats

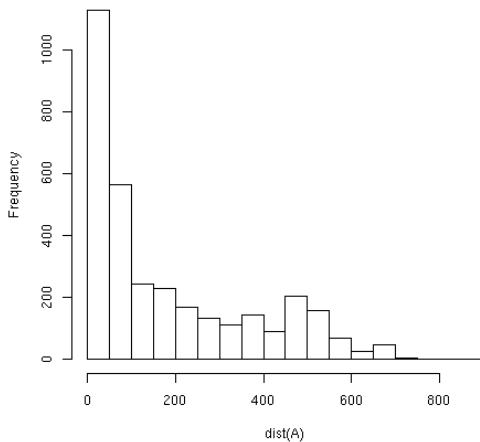
simple whitespace separated table

	label 1	label 2	label 3	...
1	3	2	1	
2	5	2	3	
3	7	3	5	
4	8	9	2	
5	...			

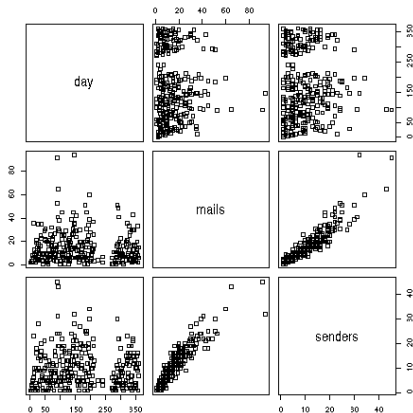
labels are optional

# histograms

Histogram of dist(A)



# scatter plots



# tool we use

R is a language and environment for statistical computing and graphics.

`http://www.r-project.org/`

FreeBSD: `/usr/ports/math/R/`

Debian:

# statistics in R

```
> data <- c(2, 2, 3, 3, 5, 5, 5, 6, 6, 7)
> data
2 2 3 3 5 5 5 6 6 7
> range(data)
2 7
> mean(data)
4.4
> median(data)
5
> summary(data)
  Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
 2.00   3.00   5.00  4.40   5.75   7.00
```

# statistics in R

```
> data
2 2 3 3 5 5 5 6 6 7
> var(data)
3.155556
> duta <- c(5, 5, 3, 6, 7, 9, 7, 4, 2, 3)
> cov( data, duta )
-0.6
> cor (data, duta )
-0.1547056
```

# clustering

## Idea

eMails with similar subject lines are about similar topics

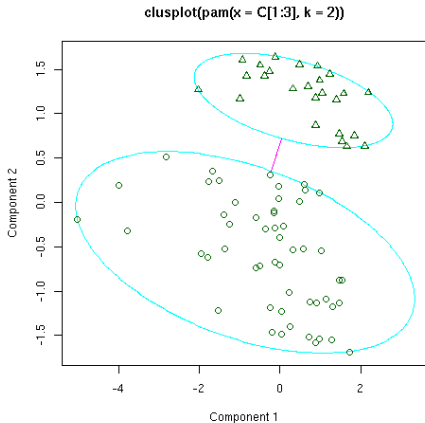
for each list

- 1 get all subject lines
- 2 for all words: count how often the word occurs in the subject lines
- 3 clean the lists from words, that carry no information

compare the lists of the word counts → clustering

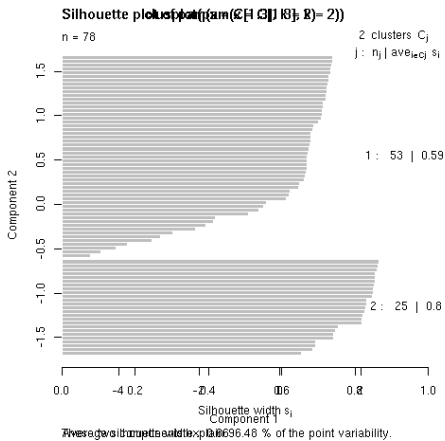


## cluster plots



These two components explain 96.48 % of the point variability.

# silhouette plots



# Association Rule Mining

## Idea

mailing lists with many equal writers are somehow related

**item** mailing list

**transaction** all the mailing lists someone writes to  
within a week

we used the mailing list archive of the ietf

- 171 items (mailing lists)
- 2084 transactions (writers who write to two different mailing lists within a week)

# Association Rule Mining

association rule:

$\text{dhc} \leftarrow \text{dhcwg dhcipv6} (10.9, 99.6)$

## support

proportion of transactions which contain all items from the rule

## confidence

accuracy — proportions of all transactions which contain right part of the rule that also contain the left part of the rule

# Apriori

- rules with enough support are called frequent
- each subset of a frequent itemset has to be frequent
- so the algorithm starts with small itemsets, checks if they are frequent and goes on to supersets of frequent itemsets

# tool we use

## Apriori-Implementation by Christian Borgelt

<http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html>

# example

```
imrg asrg-announce  
ipngwg ipv6  
ipngwg ipv6  
atommib rohc  
ipngwg ipv6  
...
```

```
./apriori -s2 -c90 writers rules.rul
```

```
dhc <- dhcwg (11.1, 97.8)  
dhcwg <- dhc (11.5, 95.0)  
dhcipv6 <- dhcwg (11.1, 98.3)  
dhcwg <- dhcipv6 (11.6, 94.6)  
...
```