Web Mining

Examples and Applications

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Dipl.-Inform. Arne Pottharst
TU Darmstadt, Germany
pottharst@gmx.de

What is Web Mining?

- Uses techniques of Data Mining to discover pattern from the internet
  - Information Retrieval, Machine Learning, Statistic, Pattern Recognition
- Extract information from the internet
  - especially world wide web

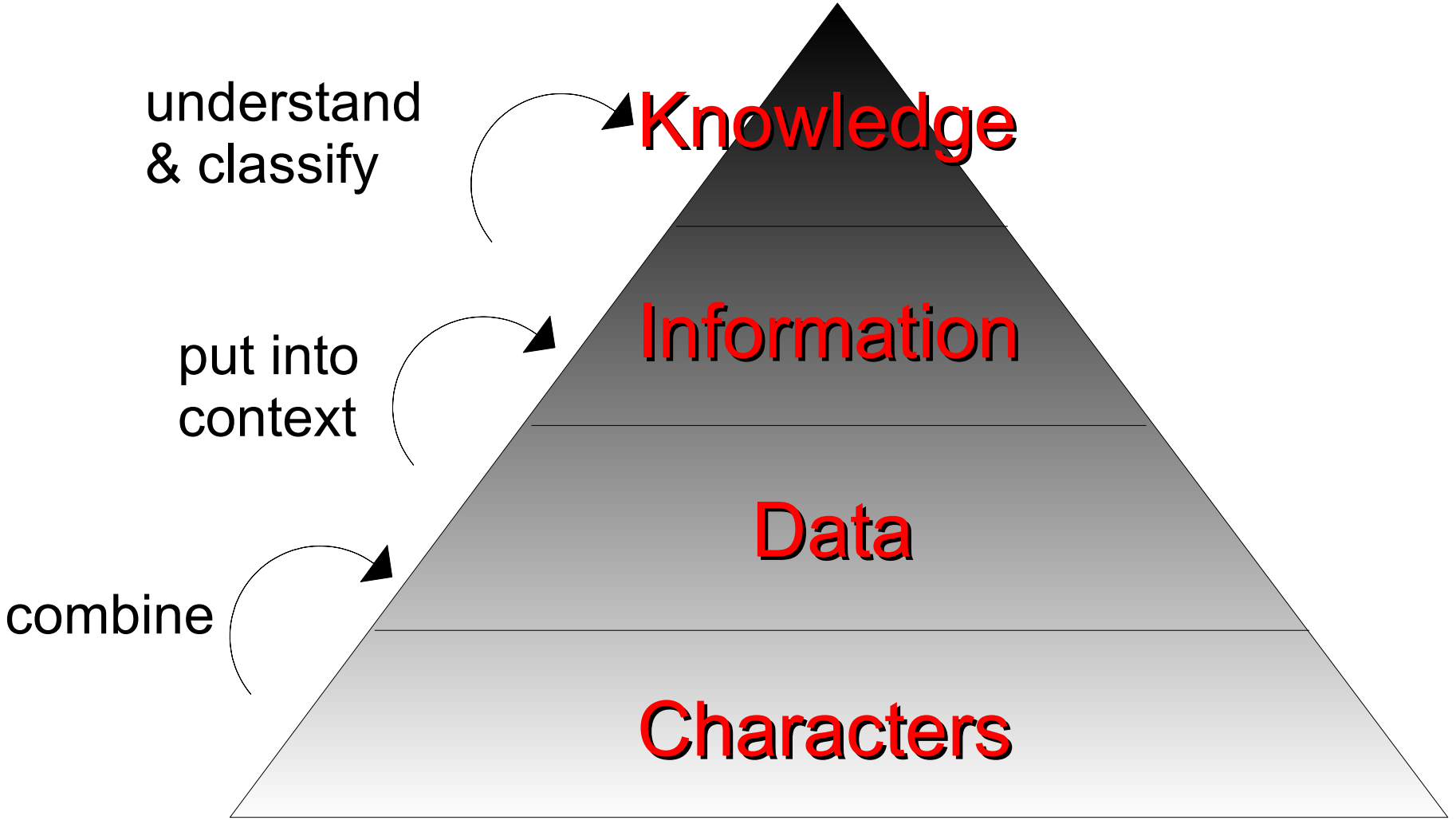
“The World Wide Web can be seen as the largest data base in the world. This huge, and ever-growing amount of data is a fertile area for data mining research.”

[Wel and Royakkers, 2004]

What is Web Mining?

- Web *Usage* Mining
  - Discover patterns in user's behaviour
    - e.g. Amazon.com recommendation system; Google Analytics
- Web *Structure* Mining
  - Use hyperlinks between websites to gain information
    - e.g. Google's Page Rank
- Web *Content* Mining
  - Discover useful information (text, image, audio, video...) in websites
    - e.g. Spam Mail Filtering, find knowledge, structure knowledge

Data – Information – Knowledge



Where do information come from?

- Business Data producing, selling, buying items
- Financial Data stock exchange, automatic trading
- Medical Data studies about diseases
- Web Usage Data web servers' log files
- Environmental Data weather forecast, climate change
- Research Data Large Hadron Collider: 300 GB/s

Web Usage Mining

- Surfing the Internet leaves a big data trail
  - Every click is logged with IP, date and URL

```
120.241.211.21 - - [10/Apr/2007:10:39:11 +0300] "GET /index.html HTTP/
227.10.1.1 - - [10/Apr/2007:10:39:11 +0300] "GET /favicon.ico HTTP/1.1
139.12.3.2 - - [10/Apr/2007:10:40:54 +0300] "GET /about.html HTTP/1.1"
139.12.12.2 - - [10/Apr/2007:10:40:54 +0300] "GET /favicon.ico HTTP/1.
267.103.53.1 - - [10/Apr/2007:10:53:10 +0300] "GET /index.html HTTP/1.
267.276.54.1 - - [10/Apr/2007:10:54:08 +0300] "GET /shop/index.html HT
187.229.163.1 - - [10/Apr/2007:10:54:08 +0300] "GET /style.css HTTP/1.
181.74.93.1 - - [10/Apr/2007:10:54:08 +0300] "GET /img/pti-round.jpg H
201.93.53.1 - - [10/Apr/2007:10:54:21 +0300] "GET /unix_sysadmin.html
213.23.22.3 - - [10/Apr/2007:10:54:51 +0300] "GET /index.html HTTP/1.1
117.2.22.3 - - [10/Apr/2007:10:54:51 +0300] "GET /favicon.ico HTTP/1.1
114.2.22.3 - - [10/Apr/2007:10:54:53 +0300] "GET /cgi/pti.pl HTTP/1.1"
217.32.22.3 - - [10/Apr/2007:10:58:27 +0300] "GET /shop/index.html HTT
```

- Companies collect these information
  - They use it to optimize their websites/sales

Web Usage Mining

- amazon.com:
  - 200M visits/month (source: compete.com)
  - 200M log file entries/month
  - Logged *every click* since 1994

- Use log files to calculate recommendations
  - “Frequently bought together ...”
  - “Customers who bought ... also bought ...”

Web Usage Mining

amazon.com

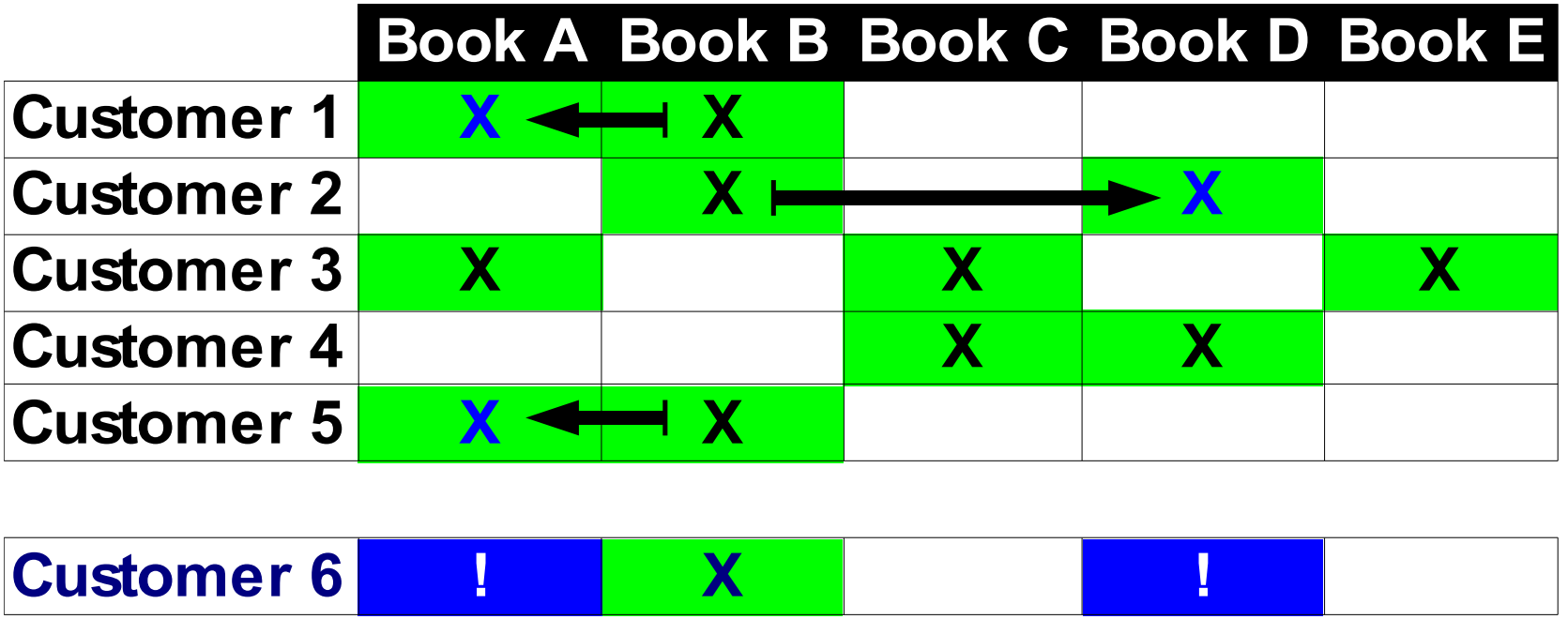
- Customer 1-5 bought some books
- New customer 6 wants to buy book **B**
- Which book(s) should be recommended?

	Book A	Book B	Book C	Book D	Book E
Customer 1	X	X			
Customer 2		X		X	
Customer 3	X		X		X
Customer 4			X	X	
Customer 5	X	X			
Customer 6	?	X	?	?	?

Web Usage Mining

amazon.com

- “Customers who bought book **B** also bought book **A** and book **D**”.
- “Frequently bought together with book **B** is book **A**.”



Web Usage Mining

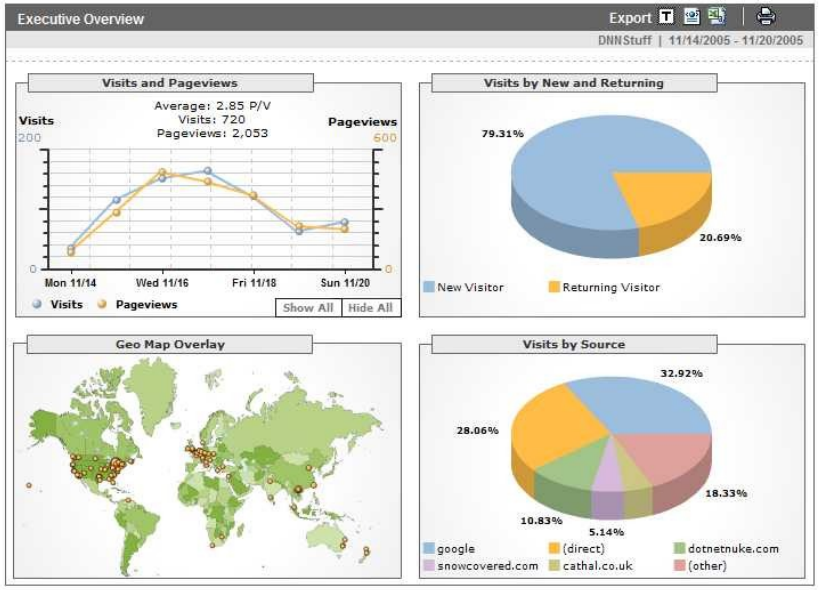
- Small example with 5 books and 6 customers
- In reality: 100'000s of articles, 100 customers/sec

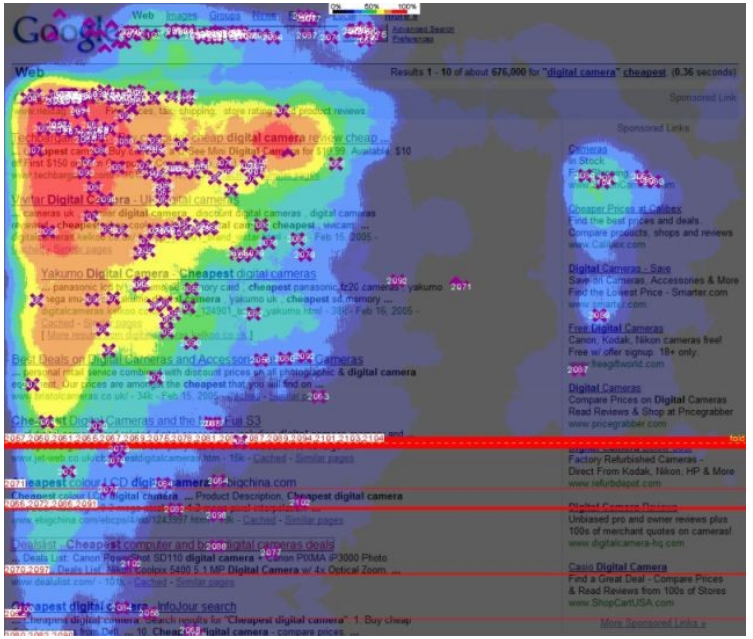
- More background data used:
  - user's age, sex, region
  - former buys, former viewed products

- Need of efficient algorithms
  - Data Warehousing
  - Data Mining
  - Web Mining

Web Usage Mining

- Google Analytics
  - Free service for website analysis
  - Detailed statistics about visitors
  - “Heatmaps”

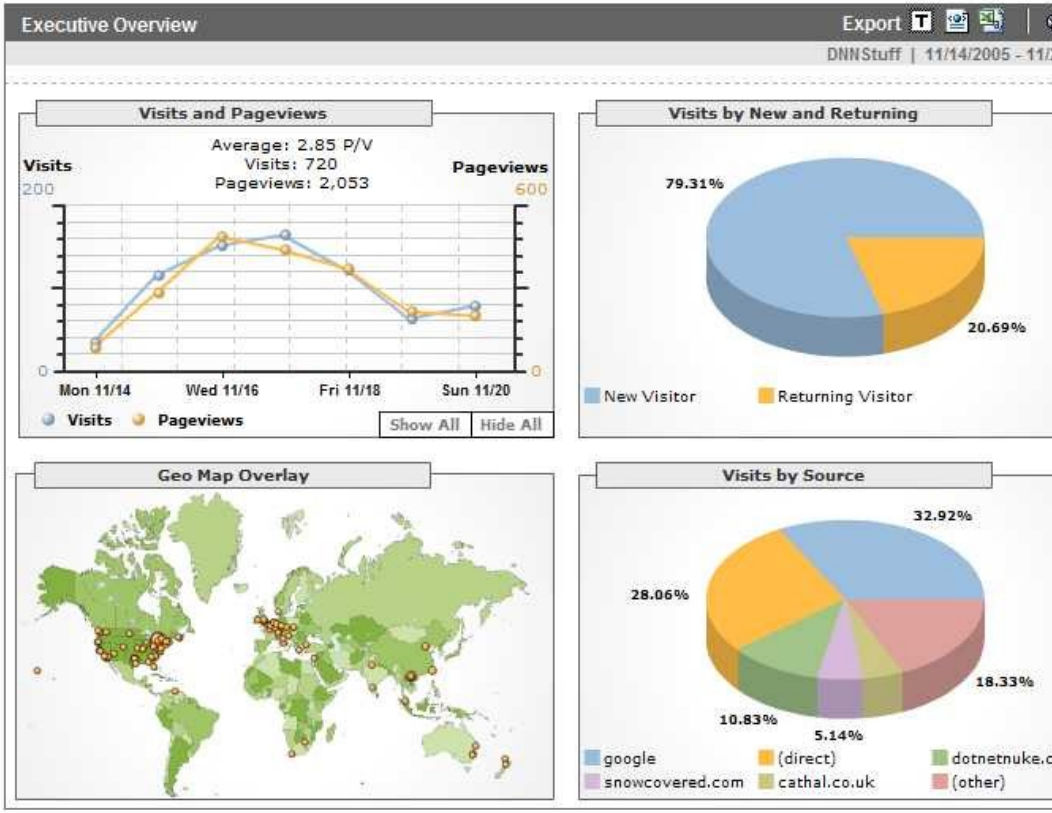




Web Usage Mining

Google Analytics

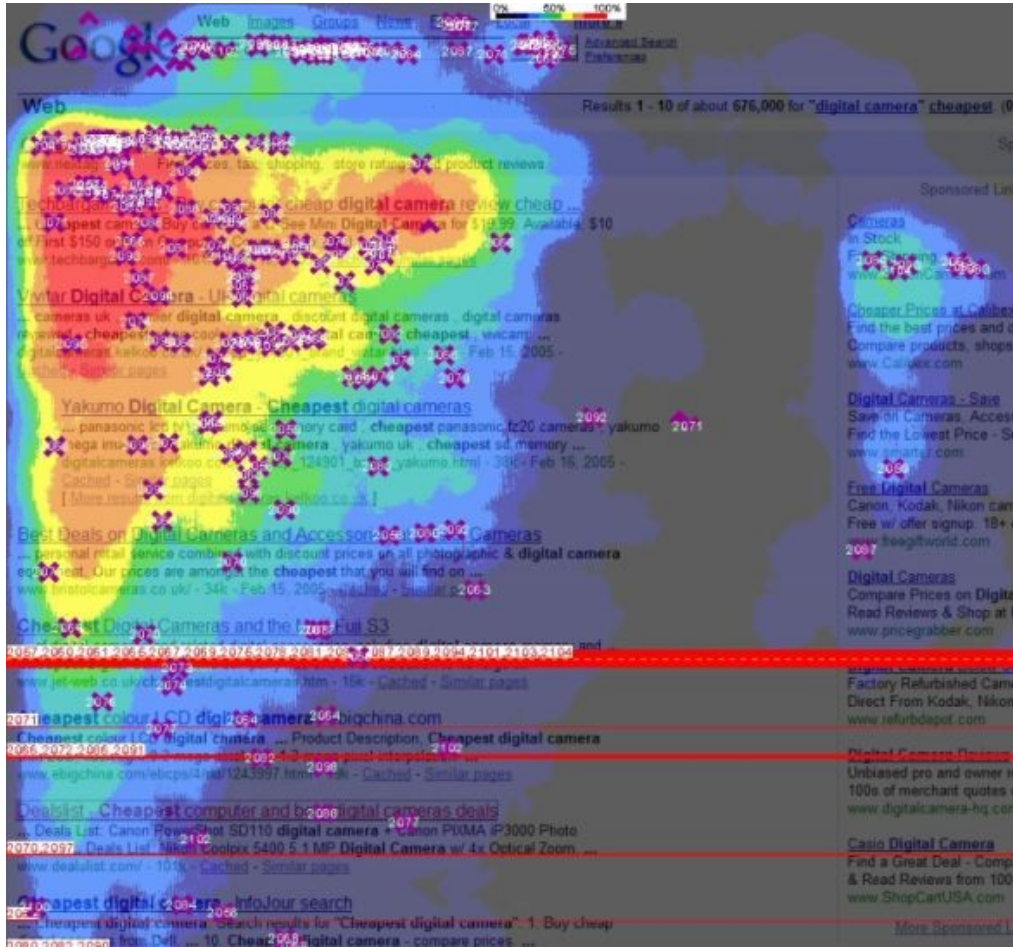
- Website statistic
  - Most visited sites
  - Times of visit
  - Clickstream
  - Where do visitors come from?
    - referrer = previous website
    - country, city
  - Browser/OS they use
- Optimize content
- Optimize campaigns



Web Usage Mining

Google | Analytics

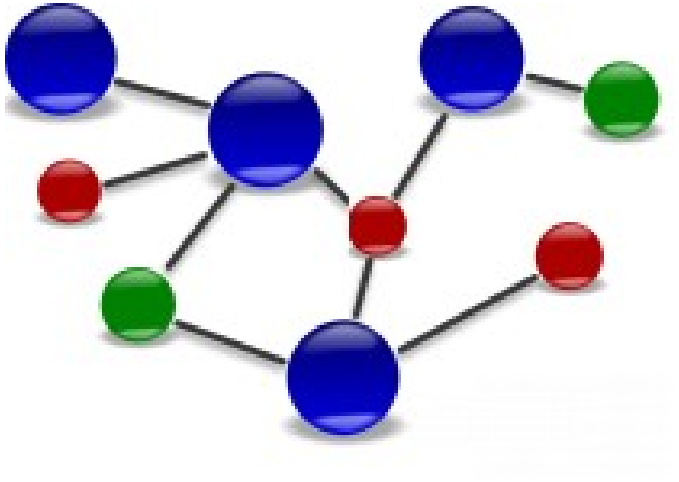
- “Headmap”
  - Where do users move their mouse?
    - Psychology: Many users move their mouse where they look/read
  - Where do they click?
- Optimize layout



Web Structure Mining

- Use the unique structure (hyperlinks!) of WWW to gain information

- Web as a graph:
  - Nodes = websites
  - Edges = links
    - Use algorithms from graph theory



- Find out popular Websites (Google Page Rank)
- Find out similar Websites

Web Structure Mining

- Google Page Rank (Brin, S.; Page, L, 1998)

- Idea: “Random Surver Model”
  - if there are few links, a specific one will be chosen with high probability
  - if there are many links, a specific one will be chosen with low probability
- Many in-links: Authority
- Many out-links: Hub

Web Structure Mining

- Google Page Rank
  - Websites link to interesting websites, so they “vote” for them
  - The more websites vote to a website, the more interesting it is
  - Also regard the votes for recommending Websites

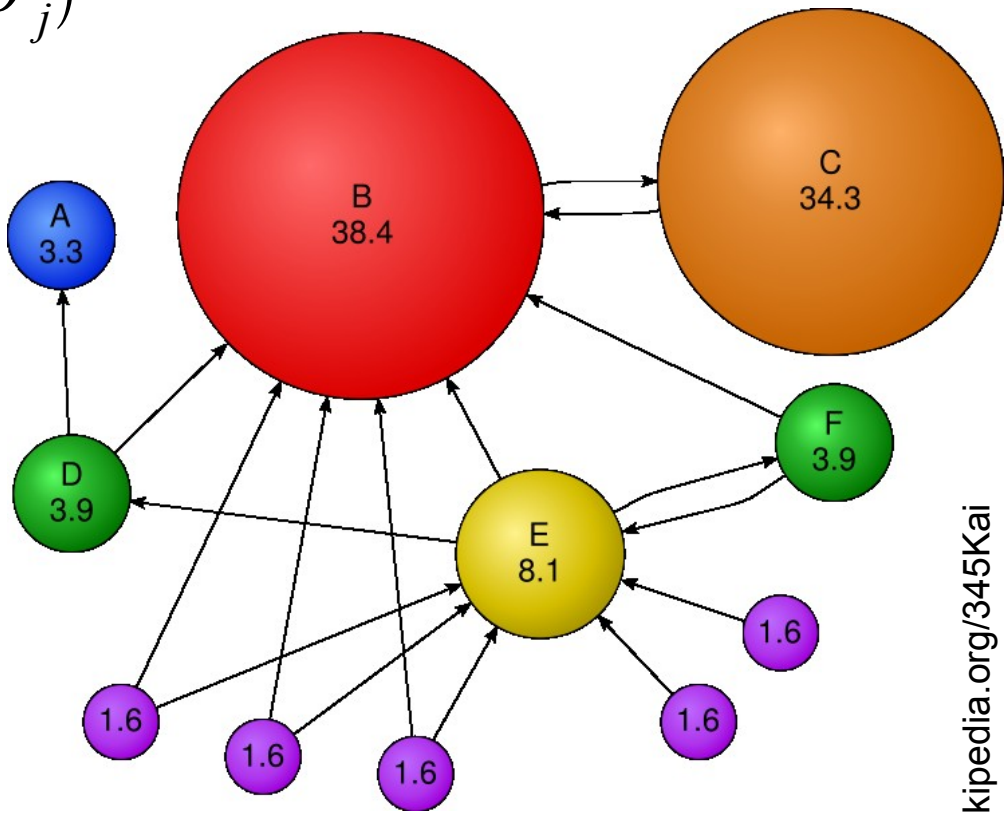
  - Every website has a starting score
  - Scores are calculated incremental

Web Structure Mining

Google

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- PR: Page Rank
- p_i: page I
- d: damping factor
- N: number of pages
- L: out-links
- M: in-links



wikipedia.org/345Kai

Web Content Mining

- Get structured information from unstructured website content

"Who was the former president of the U.S.A?"

"How will the weather be tomorrow in Jakarta?"

Web Content Mining

- Discover implicit knowledge from explicit knowledge
  - Textmining, Natural Language Processing
    - Use features of language to discover knowledge
    - Make conclusions

- List of all U.S.-Presidents } explicit knowledge
- Former U.S.-President } implicit knowledge

- “Peter can read.” } explicit knowledge
- “You have to learn reading.”

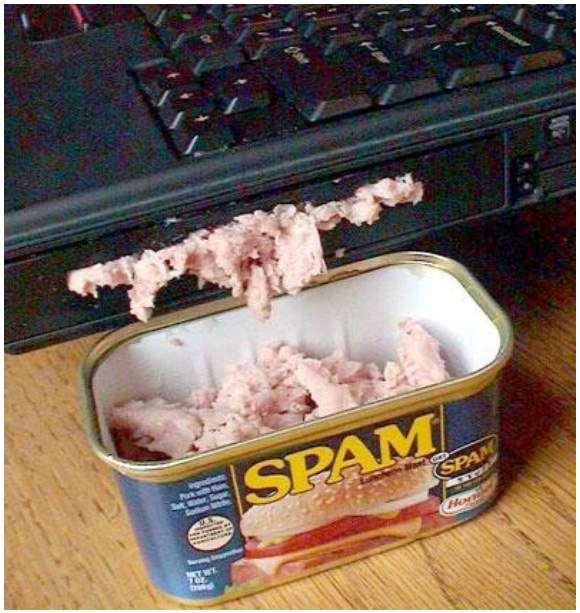
  → “Peter once learnt reading.” } implicit knowledge

Web Content Mining

- Get interesting content of website
  - navigation, ads, headers ... not interesting at all
- find out, which parts repeat on other sites and which are unique
  - analysis of written text (content)
  - structure of HTML
  - generalize it as much as possible

Web Content Mining

- Spam Mail Filtering
  - >50% of email traffic is spam
  - need efficient methods to filter spam but not throw away “ham”



- Bayes' Method:
  - train with spam and “ham”
  - calculate for each word the spam probability
  - when new mail arrive:
    - for each word X in the mail:
    - if X is often in spam and rarely in “ham”, new email is probably spam

Ethical Issues

- Processing of personal data
  - You need to give your email address and other personal data on many sites where it is not needed
- Create profiles of *every user* of the internet
  - Data profile
  - Behaviour profile
- Clustering of profiles to get a prediction of your behaviour:
  - Credit companies use this:
  - e.g. Everybody living in a certain area does not get a credit

Conclusion

- Many techniques to get Knowledge out of Information and Information out of Data
- There are more aspects than in this short presentation
- Lots of formulas and statistics (not presented here)
- Relatively new research field (~10 years), lot of work to be done :-)
- But: take care of ethics!

Advance of technology

Literature

- Baraglia and Silvestri. Dynamic personalization of web sites without user intervention. In *Communication of the ACM* 50(2): 63-67, 2007.
- Wel and Royakkers. Ethical issues in web data mining. In *Ethics and Information Technology* 6: 129–140, 2004.
- Cooley, Mobasher and Srivastave. Web Mining: Information and Pattern Discovery on the World Wide Web. In *Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence*, 1997