Further Teaching

• First time in FSS 2014: Data Mining II
  – Taught by Heiko & Robert

• Topics
  – Advanced Data Preprocessing
  – Regression
  – Anomaly Detection
  – Time Series Analysis
  – Ensemble Learning
  – Online Learning
  – Parameter Tuning

• Project
  – Participation in the Data Mining Cup 2014
Further Teaching

• One comment in the feedback:
  – ideas for improvement: “include web mining topics”, “more sentiment analysis”

• Lecture in FSS 2014: Web Mining
  – taught by Christian Bizer, Cäcilia Zirn, and Robert Meusel

• Topics:
  – Web Usage Mining
  – Recommender Systems
  – Web Structure Mining
  – Social Network Analysis
  – Web Content Mining
  – Information Extraction
  – Sentiment Analysis
Further Teaching

- Web Search and Information Retrieval
  - Focus on text mining problems

- Topics
  - Efficient text indexing
  - Boolean and vector space retrieval models
  - Evaluation of retrieval systems
  - Probabilistic Information Retrieval
  - Text classification and clustering
  - Web search, crawling and link-based algorithms.
Master Thesis Topics

• Our group offers various topics in various areas
  – Data Mining
  – Web Mining
  – Semantic Web
  – Information Extraction
  – ...

• Just scan our website
  – thesis topics are often not up to date
  – but research interests are
- References:
  - Daniel Rinser, Dustin Lange, Felix Naumann: Cross-lingual entity matching and infobox alignment in Wikipedia.

- Contact: Christian Bizer
Fast-forward Feature Generation

• RapidMiner LOD Extension
  – an extension to RapidMiner
  – acquires background knowledge from Linked Open Data

• Current approach
  – get data first
  – evaluate and filter later
  – most data is useless, but it's hard to tell before having it

• Scalable approach (your work!)
  – combine data acquisition and evaluation/filtering
Divide et Impera

• Given a table about cities
  – population, area, cinemas, restaurants, …

• Task
  – predict, e.g., quality of living

• Some “artificial” columns would help
  – in particular: quotients, e.g., cinemas/population
  – trying all quotients is too complex – $O(cols^2 \times rows)$

• Envisioned solution (your work)
  – try to predict which quotients will be insightful before computing them
John, Paul, George, and Ringo

• Who are these four? An easy task for most humans
  → John Lennon, Paul McCartney, George Harrison, and Ringo Starr

• However, an equally valid solution would be
  → John Cage, Paul Auster, George W. Bush, and Ringo Mendoza

• Goal (your work!)
  – find a set of things that are as equal as possible
  – practical relevance: automatic table fusion
DBpedia Usage Mining

• DBpedia contains millions of facts
  – but not all of them are equally interesting
  – some are more closely related than others

• Can we tell the interesting stuff from the non-interesting one?

• Approach (your work!)
  – look at DBpedia usage logs
  – find typical usage patterns
  – exploit usage patterns for information ranking and clustering
What's in a Wikipedia Category Name?

• Wikipedia categories are interesting to humans
  – e.g., *Australian Bands founded in 1990*

• ...but not interpretable by machines

• Approach (your work!)
  – translate categories into formal axioms
  – e.g., $\text{Band}(x)$, $\text{from}(x, \text{Australia})$, $\text{foundingYear}(x, 1990)$
  – contribute formal axioms to DBpedia
Hyperlink Graph – Structure of the WWW

3.5 billion URLs

128 billion links

>350GB of raw data