Search Engines – Exercise

Smart Data Extraction

Franz Goerke
Thomas Berger
Our guideline

■ Requirements:
  - Extraction of tables / structured data
  - Analyze table structures
  - Enable search for tables
  - Tableindex creation

■ Enhancements:
  - Get relation among tables
  - Link resolving
  - Statistics about table contents
New Database structure

- **New Tables**
  - WikiTables
  - ArticleTitles

- **New Indices**
  - Inverted Table Index
  - Linked Table Index
  - Attribute Table Index
Relations among tables

- Approaches
  - Link analysis
    - Tables contain links to articles
      » Not very useful
      » Links mostly refer to just one single term
  - Content analysis
    - For every term query inverted index
    - Count occurrences of tables for every word
    - Use top results
    - Very useful
  - Attribute analysis
    - Not useful
    - Most attribute names are equal (name, year, language)
Linked tables

Opt: Backlinking
Demo & Questions