Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

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

TableID	ArticleID	Content

ArticleID	Title

- New Indices
  - Inverted Table Index
  - Linked Table Index
  - Attribute Table Index

TableID	ArticleID	Content

TableID	LinkedTables	BacklinkedTables

Attribute	Tables

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

Oscar Mathisen Award winners

Year	Winner	Nationality
1959	Knut Johannesen	Norway
1960	Boris Stenin	Soviet Union
1961	Henk van der Grift	Netherlands
1962	Jonny Nilsson	Sweden
1963	Nils Aaness	Norway
1964	Ants Antson	Soviet Union
1965	Per Ivar Moe	Norway

1. Extract Links

via Bliki API

article_titles

id	title
10	AccessibleComputing
15	AfghanistanPeople
20	AfghanistanMilitary
25	Autism
30	AsWeMayThink
35	AlbaniaGovernment

lookup

tables

tableId	articleId	content
1	39	{| class="wikitable" border="1&quot...
2	290	{| class="wikitable" |- style="back...
3	290	{| cellspacing="10" cellpadding="0&...
4	303	{| class="wikitable" border="1&quot...
5	303	{| class="wikitable" border="1&quot...
6	305	{| class="wikitable" border="1&quot...
7	324	{| class="sortable wikitable" style=&quo...

Opt: Backlinking

tableId	linkedTables	backLinkedTables
4	332059,17500	*NULL*
7	14064,218503,218504,218505,218506	*NULL*
10	101533,101534	*NULL*
12	4354,4355,4356,4464,4465,4466,4467	*NULL*

Demo & Questions