VL Search Engines

Lab course: WikiSearch

Agenda

- **Projektbeschreibung**
- Topics
- Prototyp
- Organisatorisches

Search

search engine Go

Page: **1** 2 3 4 5 6 7 8 9 10

Apollo 11
Snippet for article #662

Algorithm
Snippet for article #775

Aircraft
Snippet for article #849

Anna Kournikova
Snippet for article #890

Aluminium

Kournikova played at the Medibank International in Syd
round. She then reached the third round of the Australi
(6–4, 4–6, 6–4). She lost in the second round of the P

Kournikova reached the semifinals in Hannover. She lo
to Novotná in the quarterfinals in Linz, and to Conchita
reached her first WTA Tour final in Miami, where she lo
won the first set (2–6, 6–4, 6–1).

She then played at Amelia Island, when she reached th
lost in the quarter finals of the Italian Open to Martina H
when she lost to Conchita Martínez. During this tournar
Hingis.

Kournikova then played at the French Open. She lost t
the semifinals at Eastbourne (lost to Sánchez Vicario). She then lost in the first, second, thir
tournament was the WTA Championships, where she lost in the first round to Monica Seles.

In 1998, Kournikova reached her first doubles final, partnering with Larisa Neiland. That was
Appelmans and Miriam Oremans in three sets 1–6, 6–3, 7–6(3). They also lost to Nathalie T
Partnering Monica Seles, she won the Tokyo title. They defeated Mary Joe Fernandez and A
Sánchez Vicario, she lost to Lindsay Davenport and Natasha Zvereva in the finals at Filderst
doubles.[5]

1999

At the end of the season, Anna Kournikova was ranked #12 in singles and #1 in doubles.[5] A
athlete in the world on Yahoo!, the premier search engine of the day.[6]

search engin

Page: 1 2 3 4 5

Apollo 11
Snippet for art

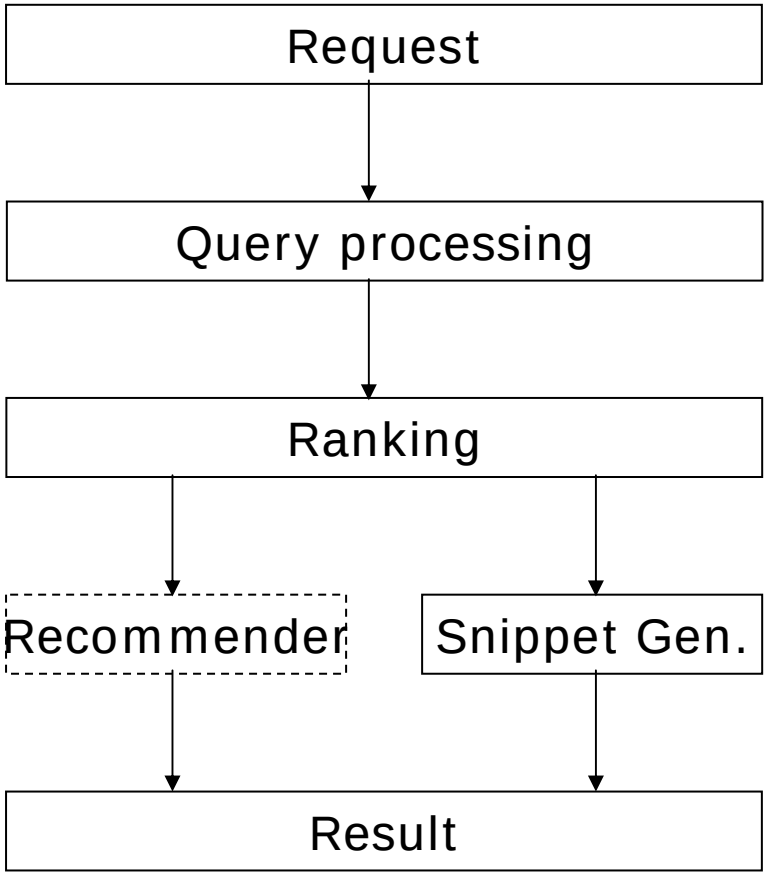
Algorithm
Snippet for art

Aircraft
Snippet for art

Anna Kourniko
Snippet for art

Aluminium

WikiSearch


Request
Query processing
Ranking
Recommender
Snippet Gen.
Result

WikiSearch

- CREATE TABLE **articles** (
  id int(11) NOT NULL,
  title text NOT NULL,
  text text NOT NULL,
  PRIMARY KEY (id)
)
  - 8,227,433 Artikel

id	title	text
887	MessagePad	[[Image:Apple Newton MP100.jpg|thumb|250px|The App...
888	A. E. van Vogt	<!-- Unsourced image removed: [[Image:A.E._van_...
890	Anna Kournikova	{{Infobox tennis player |playername= Anna Kourniko...
891	Accounting	#REDIRECT [[Accountancy]]{{R from related word}}
892	Alfons Maria Jakob	{{Infobox Scientist |name = {{PAGENAM...

WikiSearch

- CREATE TABLE **inverted_index** (
  word varchar(100) NOT NULL,
  articles text NOT NULL,
  PRIMARY KEY (word)
)
  - 7,841,856 Terme

word	
search	303,305,307,308,332,339,569,573,576,584,594,621,628,662,664,681,689,700,736,737,752,775,777,782,791,800,808,824,849,876,890,8
engine	657,662,663,698,775,780,803,848,849,887,890,898,904,974,1009,1014,1022,1027,1132,1146,1175,1202,1208,1271,1311,1336,1358,136

Team

Prof. Dr. Felix Naumann

Mohammed

Alexander

Jens

Christoph

Agenda

- Projektbeschreibung
- **Topics**
- Prototyp
- Organisatorisches

Themen

- **Core Topics**
  - Index building
  - Query processing
  - Snippet Generation

Index Construction

- **What is an index?**
  - Map from terms to the parts of a document where they occur
- **Why is it important?**
  - Term, document frequency -> weighting
- **Important aspects**
  - Memory, Storage, Processor resources
  - Dynamic indexing
  - Positional indexes
- **The task**
  - Build an index for the Wikipedia collection
- **References**
  - Manning et al., Introduction to Information Retrieval, Chapter 4.
  - Croft et al., Search Engines, Chapter 5

Index Construction

- Relational-Database index.
- Frequencies of each term in each article (term-article basis).
- **Fist Task**:
  - Read related material.
  - Model a relational schema for required tables.
  - Design required classes / modules to scan through all articles and create the index.

Query processing

- Processing keyword queries
  - Stemming keywords
  - Using an index
  - Query expansion
- Result page with relevant articles
  - Simple ranking
  - Paging
  - Spell checking & suggestions
- **First Task**
  - Read Croft et al., Search Engines, Section 5.7 & Chapter 6

Snippet Generation

- **What are Snippets?**
  - A short summary of the document, which is designed to allow the user to decide document's relevance.
- **Its importance?**
  - Help the user determine which result item to choose.
- **Important Aspects:**
  - Static vs. Dynamic snippet generation
- **The task?**
  - Build a snippet generation engine
- **References**
  - Manning et al., Introduction to Information Retrieval, Section 8.7
  - Croft et al., Search Engines, Section 6.2.4

Snippets Generation

- Query-biased snippet generation.
- Caching mechanism?

- **First Task**:
  - Read related material.
  - Evaluate document summarization techniques.

Themen

- **Advanced Topics**
  - Index compression
  - Smart Extract
  - Information extraction
  - Ranking by Paths through the Web
  - Recommendation
  - Search Engine Optimization
  - Search Engine Evaluation
  - Wikify! Learning to Link with Wikipedia

Index Compression

- **Why is compression important?**
  - Less disk space
  - Increased use of caching (frequent terms)
  - Faster transfer of data from disk to memory
- **Important aspects**
  - Fast compression and decompression algorithms
  - Lossless vs. Lossy compression
- **The task**
  - Compress the built index for the Wikipedia collection
- **References**
  - Manning et al., Introduction to Information Retrieval, Chapter 4.
  - Croft et al., Search Engines, Chapter 5

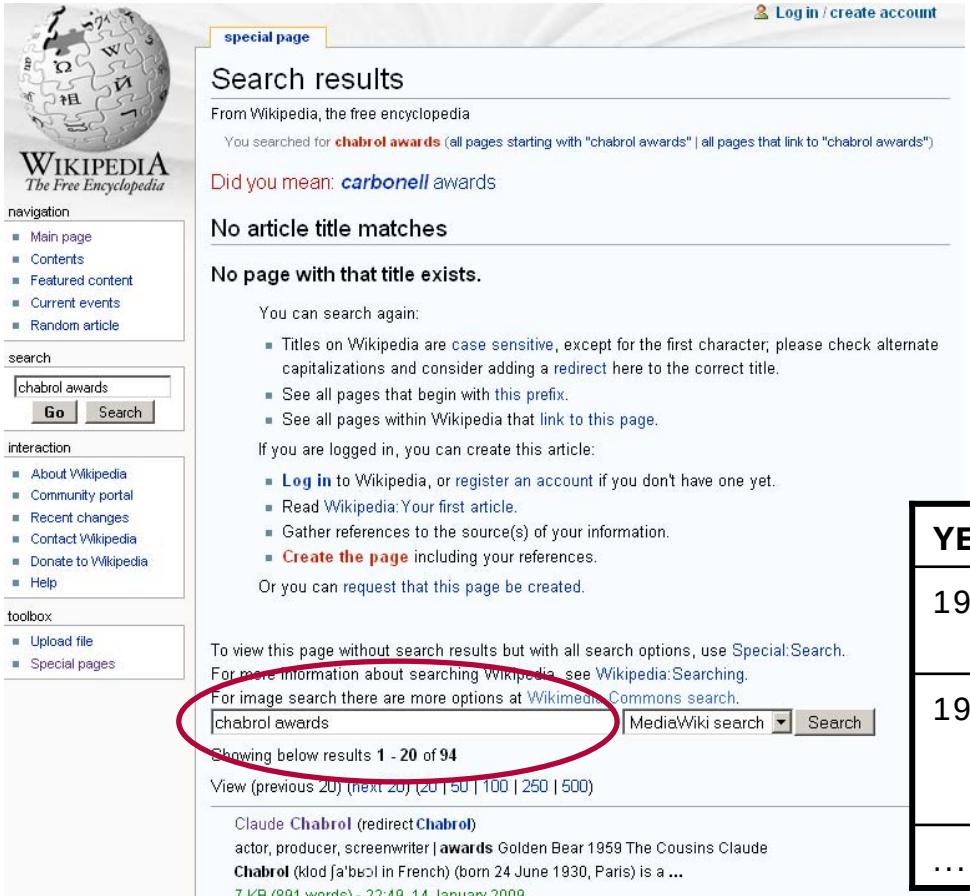
Index Compression

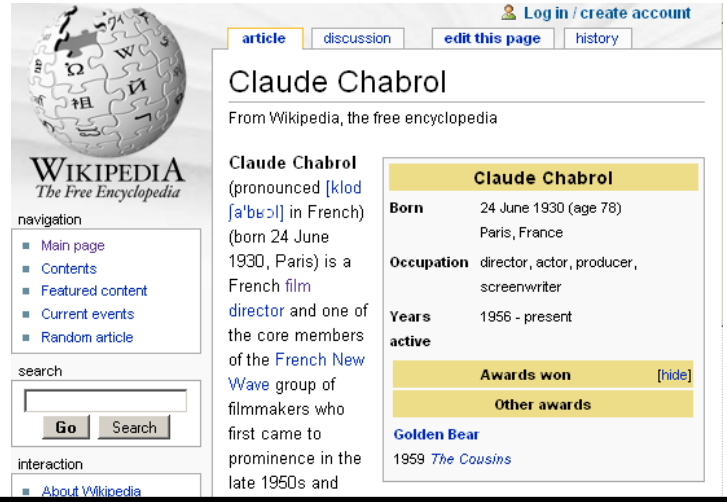
- Decrease the size of the index by:
  - Stop-words removal.
  - Removal of rare terms, i.e. occurrence (term) < threshold.
  - Removing redundant data.

- **First Task**:
  - Read related material.
  - Compare database- vs. file- based index compression alternatives.
  - Evaluate encoding schemes and choose one.

Smart Extract (1/3) - Task

- **Extract structured data** from Wikipedia articles (tables) and **find dependencies between them**
- Improved keyword search by using extracted structured data





YEAR	AWARD	MOVIE	...
1959	Berlin International Film Festival	The Cousins (Les cousins)	...
1989	New York Film Critics Circle Award for Best Foreign Language Film	Story of Women	...
...	...	...	...

Smart Extract (2/3)

- Example: Keyword query *chabrol awards* results in *Claude Chabrol*, *César Awards 1996*, *Madame Bovary (1991 film)*, ...
  - Incomplete award list in article of Claude Chabrol (only Golden Bear, 1959)
  - Goal: Offer relevant tables, i.e., awarded movies of Claude Chabrol, next to the well-known result page with the relevant Wikipedia articles
- Wikipedia documents contain lots of structured data
  - Sample Wikipedia articles, containing corresponding relational tables about films: *New York Film Critics Circle Award for Best Foreign Language Film*, *French films of 1959*, *San Sebastián International Film Festival*, *List of crime films:1980s*, *List of films set in Las Vegas*, ...

Smart Extract (3/3)

- **First Task**
  - Read *Uncovering the Relational Web* (WebDB 2008)
  - Start implementing a parser for extracting tables from wikipedia articles

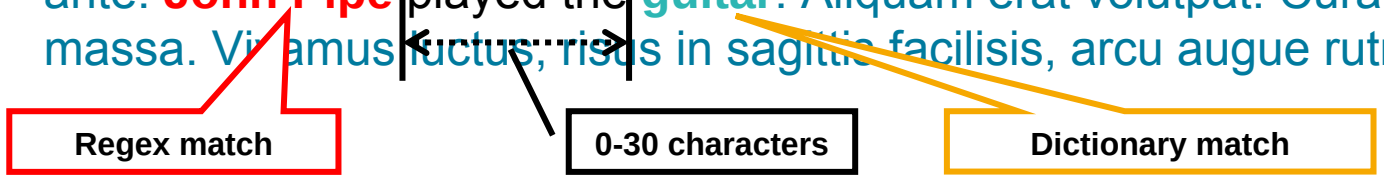
Information Extraction – IE (1/3)

- Traditional approach
  - Regular expression grammars (e-mails, phone numbers, …)
  - Drawbacks: Scaling to large data sets and large numbers of rules
- New approach
  - IBM System T
  - An Algebraic Approach to Rule-Based Information Extraction (ICDE 2008)

- **Task 1**: Choose three concepts to extract (companies, persons, …) from Wikipedia articles and write AQL scripts for each of these concepts. Evaluate your solution.
- **Task 2**: Extend the keyword search with a semantic autocompletion feature/advanced query language using your extracted concepts.

Information Extraction – IE (2/3)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Proin elentum non ante. **John Pipe** played the **guitar**. Aliquam erat volutpat. Curabitur a massa. Vivamus luctus, risus in sagittis facilisis, arcu augue rutrum ve



<ProperNoun>	<within 30 characters>	<Instrument>
Regular Expression Match		Dictionary Match

```
select
    CombineSpans(name.match, instrument.match)
        as annot
from
    Regex(/[A-Z]\w+(\s[A-Z]\w+)?/, DocScan.text) name,
    Dictionary("instr.dict", DocScan.text) instrument
where
    Follows(0, 30, name.match, instrument.match);
```

(c) Declarative Information Extraction, The Avatar Group IBM Almaden Research Center, 2008

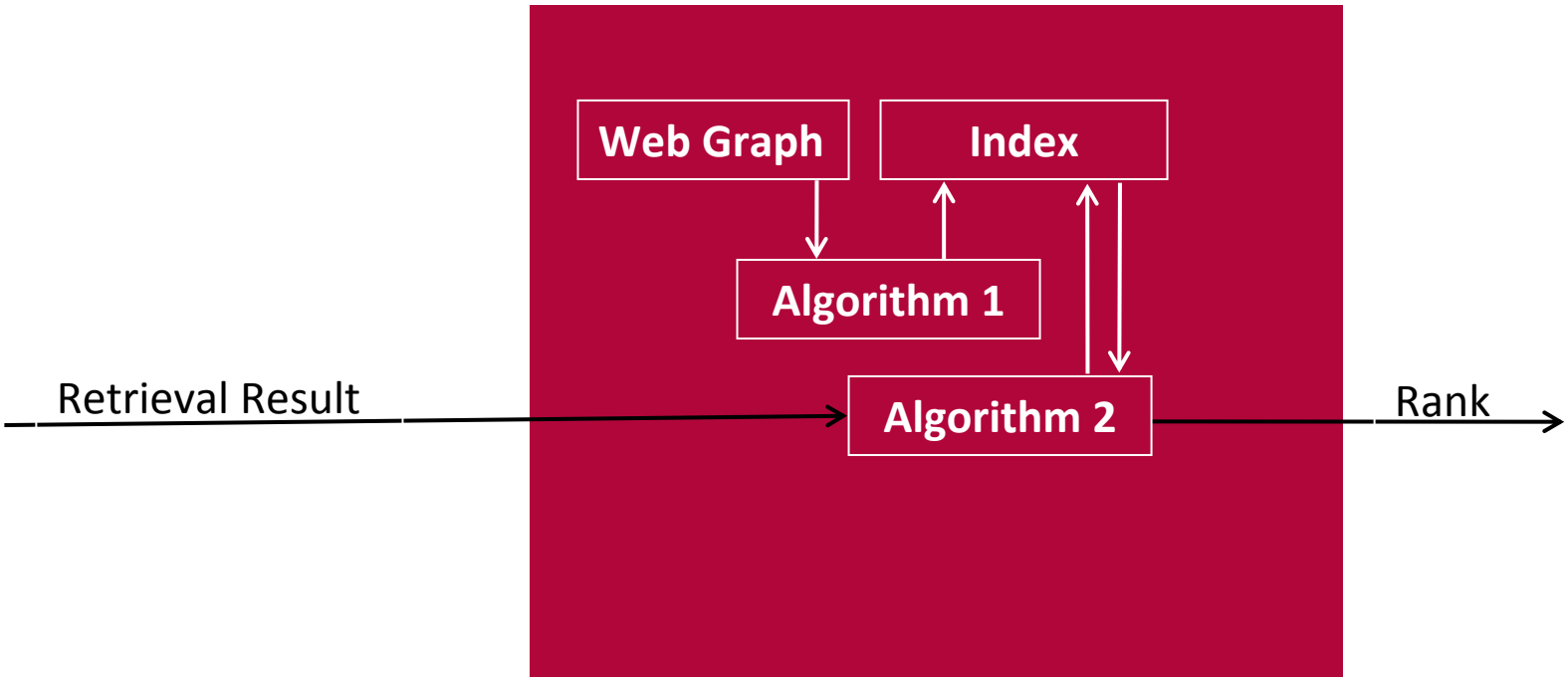
Information Extraction – IE (3/3)

- **First Task**
  - Read *An Algebraic Approach to Rule-Based Information Extraction* (ICDE 2008)
  - Install IBM System T
  - Start writing AQL scripts for wikipedia articles

Ranking by Paths through the Web (1/2)

- SPRINT: Search by Path Ranks on the INTranet
- Idea:
  - Imagine you were a computer scientist
  - You search for ‚genetic' on Wikipedia
  - ‚genetic programming' should be ranked higher than ‚genetics' or ‚genetic diversity'
  - Unless you are a biologist
- Implementation: Rank retrieval results according to the (link) distance from a user defined central webpage
- Prerequisite:
  - We know who you are or a webpage that fits your characteristics best
- **First task**: Read *Efficient search ranking in social networks*, Vieira et al., CIKIM 2007

Ranking by Paths through the Web (2/2)



Recommendation

- Expand the results of a search by attaching recommendations
  - Find one or more “recommended item” to every search
  - Items are other Wikipedia Pages
  - Recommendation should be based on interactions with pages, or other data that is not already used in a ranking
- Two possible ways of recommending items:
  - Use authorship information: authors that have written parts of this page have also written these other pages
  - Use usage/view information: users that have seen (clicked on) this results have also seen these other pages

Recommendation – first steps

- Get familiar with recommendation systems:
  - Read two easy articles (get a print copy from Jens, next week)
- Think about page interactions, and answer the questions:
  - What interactions are possible? What data for which interactions is available? What data do you want to use?
  - How do you want to gather and store the data? Think about a schema, plan gathering the data.
- Collect first small amounts of interaction data (a sample):
  - Download data
  - Crawl data, screen scrape data
  - Create or collect data by hand
- Meet me at May, 14th, with that data and the plan

Search Engine Evaluation

- Which search engine answers Wikipedia keyword queries best?
  - Wikipedia
  - Google
  - Yahoo
  - …
- Compare the quality of results and the ranking over a defined set of queries
  - Precision & recall
- Identify and evaluate applied techniques, such as stemming, query expansion, …
- Repeat your experiments within certain periods (weeks) and determine possible changes
- **First task:** Read *Croft et al., Search Engines, Chapter 8*

Wikify! Learning to Link with Wikipedia

- Automatically recognize topics mentioned in unstructured text and link them to the appropriate Wikipedia articles
- Machine learning approach
  - Can be used to identify significant terms within unstructured text
  - Millions of manually-defined links found within Wikipedia articles are used as training set
- **First task:** Read *Learning to Link with Wikipedia*, CIKM 2008 & *Wikify! Linking Documents to Encyclopedic Knowledge*, CIKM 2007

Iranian POW negotiator holds talks with Iraqi ministers

The head of **Iran**'s **prisoner of war** commission met with two **Iraqi** Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister **Mohammed Saeed al-Sahhaf** told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file,'' INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the **1980**-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The **Associated Press** that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 fied as civil law detainees in the largest exchange since **1990**.

More than 1 million people w

Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

open in wikipedia

Agenda

- Projektbeschreibung
- Topics
- **Prototyp**
- Organisatorisches

Prototyp

R:\SearchEngines09_Naumann\Materialien_nosync\

- MySQL installieren
- Wikipedia-Datenbank anlegen und laden (siehe HowTo)
- Apache Tomcat installieren & anpassen (siehe HowTo)
- WikiSearch Anwendung entpacken und starten

Agenda

- Projektbeschreibung
- Topics
- Prototyp
- **Organisatorisches**

Organisatorisches

- Liste mit 3 Themenwünschen per Email an Alexander Albrecht (mindestens eine Core Topic).
- E-Mail bis zum 26.04.2009, 23:59 Uhr
- Vergabe der Themen und Zuordnung der Teams am 27.04.
- Teams vereinbaren individuelles Treffen mit ihren Betreuern bis spätestens 14.05. (Terminvorschläge)
- Bewertung der implementierten Lösung