universität koblenz
weiter:denken

Web Information Retrieval
Introduction

Frank Hopfgartner, Stefania Zourlidou
Institute for Web Science and Technologies

Objectives of the lecture

universität koblenz
weiter:denken

- Basic concepts of Information Retrieval (IR)
- Web IR (WIR) examples
- IR history

Let’s talk about Information Retrieval

Definition: Data Retrieval

universität koblenz
weiter:denken

- Obtaining data from a database
  - using a query language (e.g. SQL)
- Data is structured and free of ambiguity

Definition: Information

universität koblenz
weiter:denken

- Information (merriam-webster.com)
  - “a collection of factual knowledge about something”
  - “a report of recent events or facts not previously known”

Definition: Information Retrieval

universität koblenz
weiter:denken

- It is very broad. E.g. getting the price of a product in a supermarket is a kind of information retrieval
- Obtaining information from **unstructured data**

"Information retrieval as a field of study is finding a **relevant information resource** that satisfies the **information need** from within a **collection of resources**."

Unstructured data

What is unstructured data?

- “it refers to data which does not have clear, semantically overt, easy-for-a-computer structure.” [5]

Structured vs. Unstructured Data

universität koblenz
weiter:denken

	Structured data	Unstructured data
Retrieval system	Data retrieval	Information retrieval
Chanonical example	Relational database	Collection of documents
Result	Very precise and always correct.	Relevance varies
Interaction	One shot query	Interaction is important
Type	Text only	Not limited to one type
Amount [6]	20% of enterprise data (2017)	80% of enterprise data (2017)
Volume	Less storage is required	More storage is required
Retrieval velocity	Relatively high	Relatively low
Scalability	Difficult	Highly scalable
Accessibility	Easy	Hard
Schema	Dependent	Free of

Information Retrieval (IR)

universität koblenz
weiter:denken

*"Information retrieval as a field of study is finding a **relevant information resource** that satisfies the **information need** from within a **collection of resources**."*

- The elements of an information retrieval system
  - Information need
  - Relevant information resource
  - Collection of resources

IR system elements

universität koblenz
weiter:denken

- **Information need**
  - is the topic about which the user desires to obtain information that satisfies conscious or unconscious need
  - is differentiated from (but expressed as) a *query*

- **Query**
  - is what the user communicates with the computer in an attempt to express the information need in words (or other format)

IR system elements

universität koblenz
weiter:denken

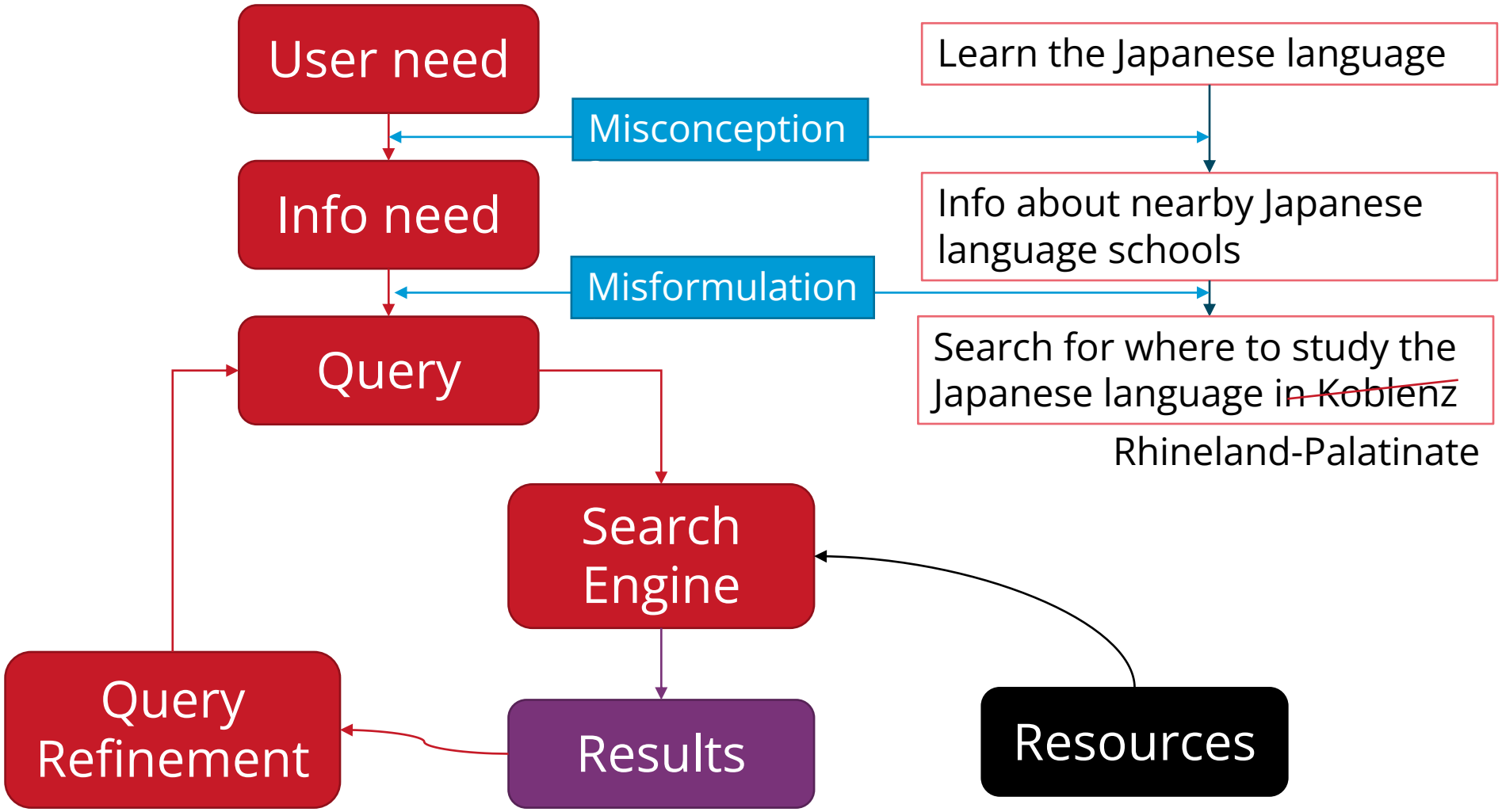
- **Relevant information resource**
  - Is the retrieved information that the user perceives valuable with respect to his/her information need
- Collection of resources
  - In case of text documents, it is referred to as corpus, but it can refer to a collection of any sort of unstructured data (text, images, videos, audio, etc.)
  - Often the resources themselves are not kept or stored directly in the IR system, but are instead represented in the system by other surrogates or metadata

Some examples of IR

universität koblenz
weiter:denken

- Web search
- Email search
- Question Answering

IR Model


User need
Misconception
Learn the Japanese language
Info need
Info about nearby Japanese language schools
Misformulation
Query
Search for where to study the Japanese language in ~~Koblenz~~ Rhineland-Palatinate
Search Engine
Query Refinement
Results
Resources

universität koblenz
weiter:denken

High Level View

An IR example

Web Retrieval

A collection of
text documents

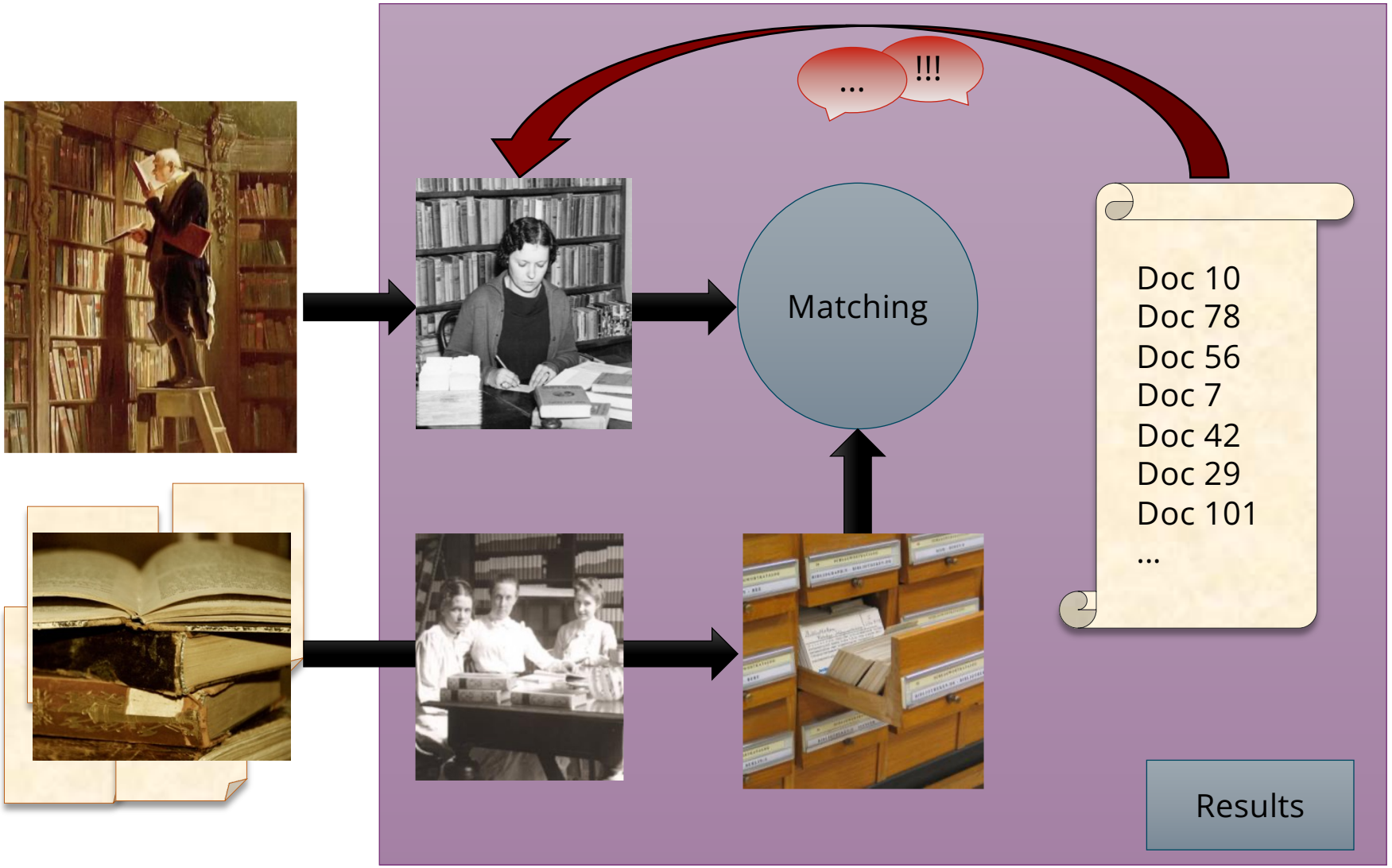
- One solution is to go through all documents and read through all the text searching for the query string
  - the computer (e.g. *grep* command in Linux) can perform this process
  - *grep* stands for *global regular expression print*
  - this process is very effective

An IR example

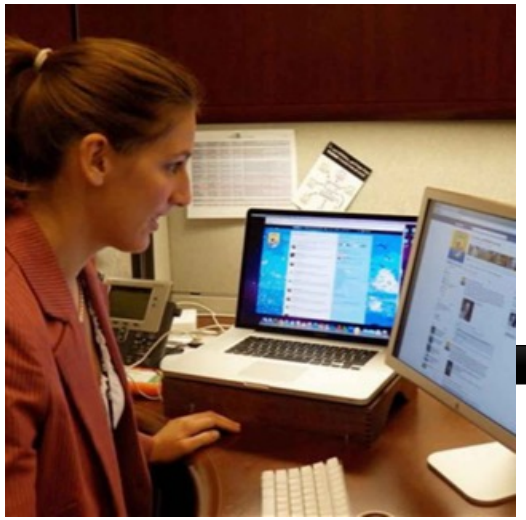
universität koblenz
weiter:denken

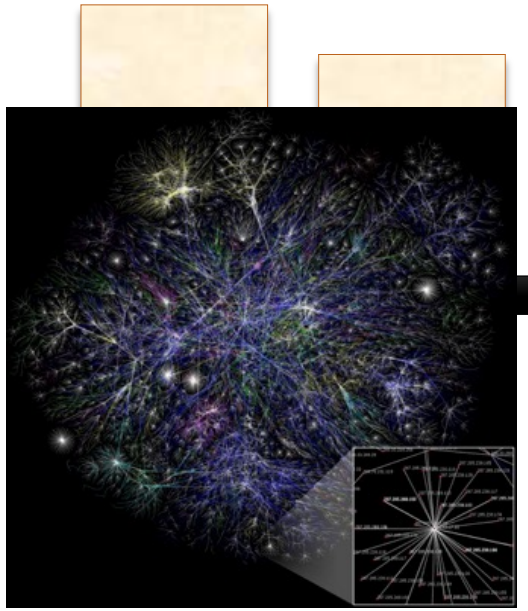
- But
  - the volume of online data compared to the speed of today's computers does not allow all data to be processed in a reasonable time
  - this process retrieves only exact matches and does not allow flexible queries
  - this process can retrieve a lot of irrelevant results that do contain the query string
  - this process cannot rank the retrieval results

IR System: General Layout



IR System: Web





Clickthrough data

web retrieval

Query Processing

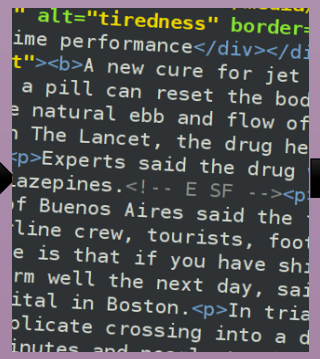
Filter

Split

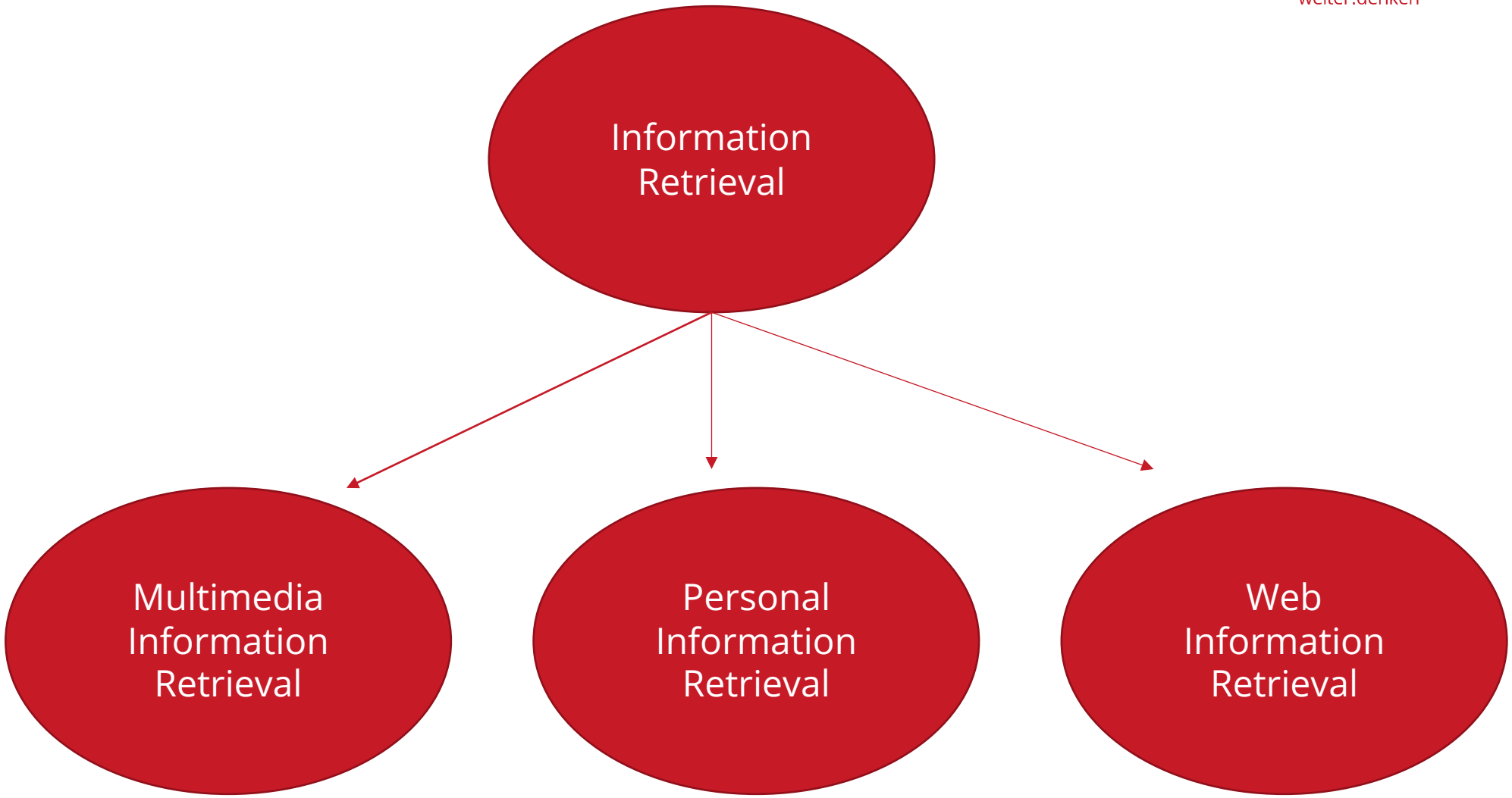
Normalize

Matching

Doc 10
Doc 78
Doc 56
Doc 7
Doc 42
Doc 29
Doc 101
...



Results

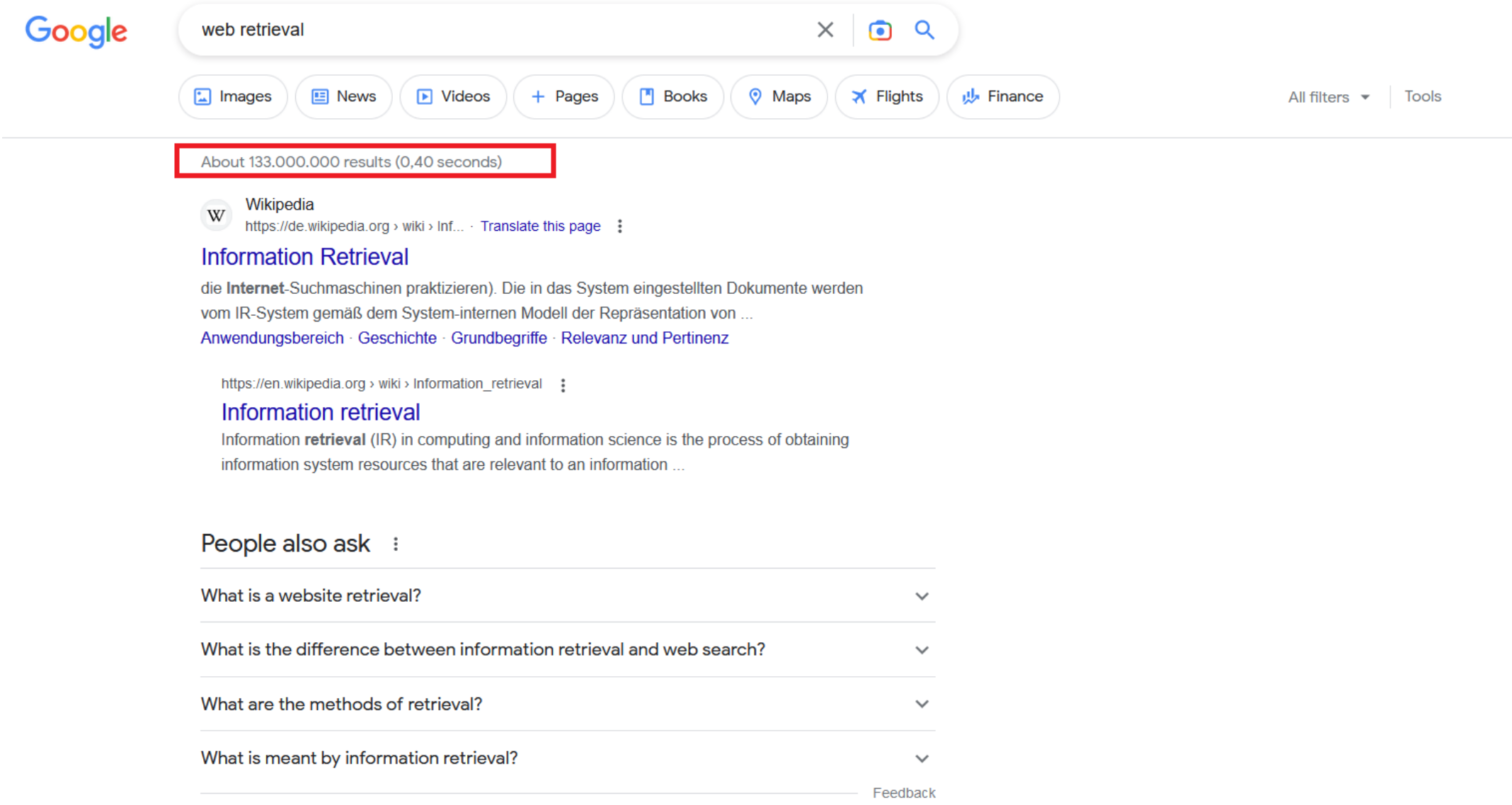

Information Retrieval
Multimedia Information Retrieval
Personal Information Retrieval
Web Information Retrieval

universität koblenz
weiter:denken

Examples

Example: google.com

universität koblenz
weiter:denken



Example: yahoo.com

universität koblenz
weiter:denken

web retrieval

Anmelden yahoo!

Alle Bilder Videos Mehr Alle Treffer

Etwa 3.080.000 Suchergebnisse

Suchergebnisse:

de.wikipedia.org › wiki › Information_Retrieval

Information Retrieval – Wikipedia

Das Wort **retrieval** bedeutet auf Deutsch Abruf bzw. Wiederauffinden. Beim IR geht es also darum, bestehende Informationen wieder aufzufinden. Etwas anderes wäre das Entdecken neuer…

→ **Relevanz:** → Pertinenz **Objektiver Informationsbedarf:** Subjektiv…
Zur Vorbereitung einer Entscheidung di…

west.uni-koblenz.de › de › studying

Web Information Retrieval | Institute WeST - uni-koblenz.de

Web Information Retrieval refers to methods and technologies for search, analysis, and automatic organization of data collections in the World Wide Web: text documents, multimedia contents,…

www.bui.haw-hamburg.de › doc › Web_Information_Retrieval_Buch

herausgegeben von Marlies Ockenfeld - HAW Hamburg

Lewandowskis **„Web Information Retrieval“** widmet sich den Grundlagen der Suchmaschinen im Internet. Gegenstände seiner Arbeit sind erstens die bei den Web-Suchmaschinen derzeit…

Example: bing.com

universität koblenz
weiter:denken

Microsoft Bing | web retrieval | Deutsch | Sign in | 3

ALL | CHAT | IMAGES | VIDEOS | MAPS | NEWS | MORE

About 3.070.000 results | Open links in new tab

From Wikipedia

Content

- Übersicht
- Anwendungsb…
- Geschichte
- Grundbegriffe
- Relevanz und …
- Dokumente
- Typologie von …

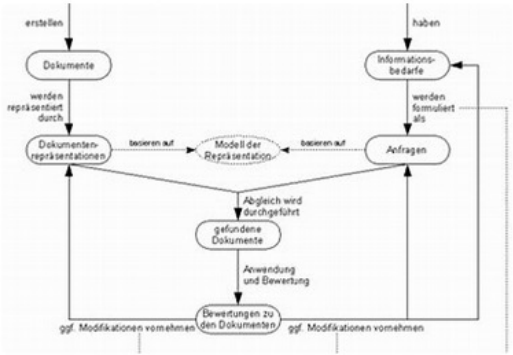
Information Retrieval – Wikipedia

https://de.wikipedia.org/wiki/Information_Retrieval

Das Wort retrieval bedeutet auf Deutsch Abruf bzw. Wiederauffinden. Beim IR geht es also darum, bestehende Informationen wieder aufzufinden. Etwas anderes wäre das Entdecken neuer Strukturen: Das gehört zur Knowledge Discovery in Databases mit Data-Mining und Text Mining. See more

Übersicht

Information Retrieval [ˌɪnfəˈmeɪʃən ɹɪˈtɹiːvəl] (IR) betrifft das Wiederauffinden von Information, meist durch Abruf aus Datenbanken. Das Fachgebiet beschäftigt sich mit computergestütztem Suchen nach komplexen Inhalten … See more

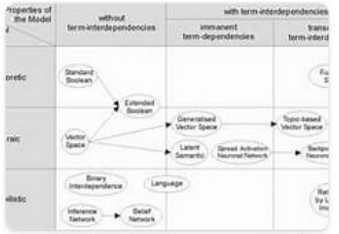


Anwendungsbereich

IR-Methoden werden beispielsweise in Internet-Suchmaschinen (wie Google) verwendet. Man nutzt sie auch in digitalen Bibliotheken (z. … See more

Geschichte

Der Begriff „Information Retrieval“ wurde erstmals 1950 von Calvin N. Mooers verwendet. Vannevar Bush beschrieb



Information retrieval

Searching for information

Information retrieval in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searc…

Related people


Susan Dumais


C. J. van Rijsbergen


Ricardo Baeza-Yates


Karen Spärck…


Rajeev Motwani

Online Information Retrieval was by far the most important course I took in library school. As a librarian, I am required to retrieve

universität koblenz
weiter:denken

IR History

History of IR

universität koblenz
weiter:denken

- **1960-70's**
  - Initial exploration of text retrieval systems for "small" corpora of scientific abstracts, and law and business documents
  - Development of the basic Boolean and vector-space models of retrieval
- **1980's**
  - Large document database systems, many run by companies
    - Lexis-Nexis
    - Dialog
    - MEDLINE

History of IR

universität koblenz
weiter:denken

- **1990's**
  - Searching FTPable documents on the Internet
    - Archie
    - WAIS
  - Searching the World Wide Web
    - Lycos
    - Yahoo
    - Altavista

IR History Continued

universität koblenz
weiter:denken

- **1990's continued**
  - Organized Competitions
    - NIST TREC
  - Recommender Systems
    - Ringo
    - Amazon
    - NetPerceptions
  - Automated Text Categorization & Clustering

Recent IR History

universität koblenz
weiter:denken

- **2000's**
  - Link analysis for Web Search
    - Google
  - Automated Information Extraction
    - Fetch
    - Burning Glass
  - Question Answering
    - TREC Q/A track

Recent IR History

universität koblenz
weiter:denken

- **2000's continued**
  - Multimedia IR
    - Image
    - Video
    - Audio and music
  - Cross-Language IR
    - DARPA Tides
  - Document Summarization
  - Learning to Rank

Related areas

universität koblenz
weiter:denken

- Database Management
- Library and Information Science
- Artificial Intelligence
- Natural Language Processing
- Machine Learning

Database Management

universität koblenz
weiter:denken

- Focused on *structured* data stored in relational tables rather than free-form text
- Focused on efficient processing of well-defined queries in a formal language (SQL)
- Clearer semantics for both data and queries
- Recent move towards *semi-structured* data (XML) brings it closer to IR

Library and Information Science

universität koblenz
weiter:denken

- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization)
- Concerned with effective categorization of human knowledge
- Concerned with citation analysis and *bibliometrics* (structure of information)
- Recent work on *digital libraries* brings it closer to CS & IR

Artificial Intelligence

universität koblenz
weiter:denken

- Focused on the representation of knowledge, reasoning, and intelligent action
- Formalisms for representing knowledge and queries
  - First-order Predicate Logic
  - Bayesian Networks
- Recent work on web ontologies and intelligent information agents brings it closer to IR

Natural Language Processing

universität koblenz
weiter:denken

- Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse
- Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on *meaning* rather than keywords
- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*)
- Methods for identifying specific pieces of information in a document (*information extraction*)
- Methods for answering specific NL questions from document corpora or structured data

Machine Learning

universität koblenz
weiter:denken

- Focused on the development of computational systems that improve their performance with experience
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*)
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*)

Machine Learning

universität koblenz
weiter:denken

- Text Categorization
  - Automatic hierarchical classification (Yahoo)
  - Adaptive filtering/routing/recommending
  - Automated spam filtering
- Text Clustering
  - Clustering of IR query results
  - Automatic formation of hierarchies (Yahoo)
- Learning for Information Extraction
- Text Mining
- Learning to Rank

universität koblenz
weiter:denken

Summary

Summary

universität koblenz
weiter:denken

- At the end of this lecture, you are expected to
  - have obtained a wide overview of WIR
  - know the elements of Web Information Retrieval
  - understand the difference between structured and unstructured data
  - understand the layout of IR system