Web Technologies
(Technologien für das Internet I)
Foundations of Information Retrieval

http://www2.kbs.uni-hannover.de/internet1.html

Introduction

Prof. Wolfgang Nejdl, Elena Demidova

Institut für Verteilte Systeme
L3S Researh Center
Leibniz Universität Hannover

13 October 2011

Plan for today

- Organizational issues
- Course overview

Outline

Information for the ITIS students

Transmission will be available to the Internet Technologies and Information Systems (ITIS) students located outside Hannover upon request at `https://webconf.vc.dfn.de/foundations_of_ir/`. If this applies to you, please ask for a password via email from Elena Demidova, demidova (at) L3S.de

Lecture and exam dates

- We meet every Thursday 11:30 - (ca.) 14:00. The exercise sessions follow the lectures (do not be late!).
- The lectures start next week (October 20).
- The exercise sessions start the week after (October 27).
- StudIP: please register for the exercise sessions (mailing list).
- Exam: 90 minutes written exam on the March 5, 2012.
- Prerequisites for the ITIS-students will be clarified later.

Outline

Literature

Selected Chapters of the IIR Book: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.
`http://www-nlp.stanford.edu/IR-book/`

Outlook for the lectures

- We will look at the algorithms and technologies used in the modern search engines to satisfy informational needs of the users from within large document collections (usually stored on computers).

Outlook for the lectures

- We will look at the algorithms and technologies used in the modern search engines to satisfy informational needs of the users from within large document collections (usually stored on computers).

Outlook for the lectures

- We will look at the algorithms and technologies used in the modern search engines to satisfy informational needs of the users from within large document collections (usually stored on computers).

Outlook for the lectures

- We will look at the algorithms and technologies used in the modern search engines to satisfy informational needs of the users from within large document collections (usually stored on computers).

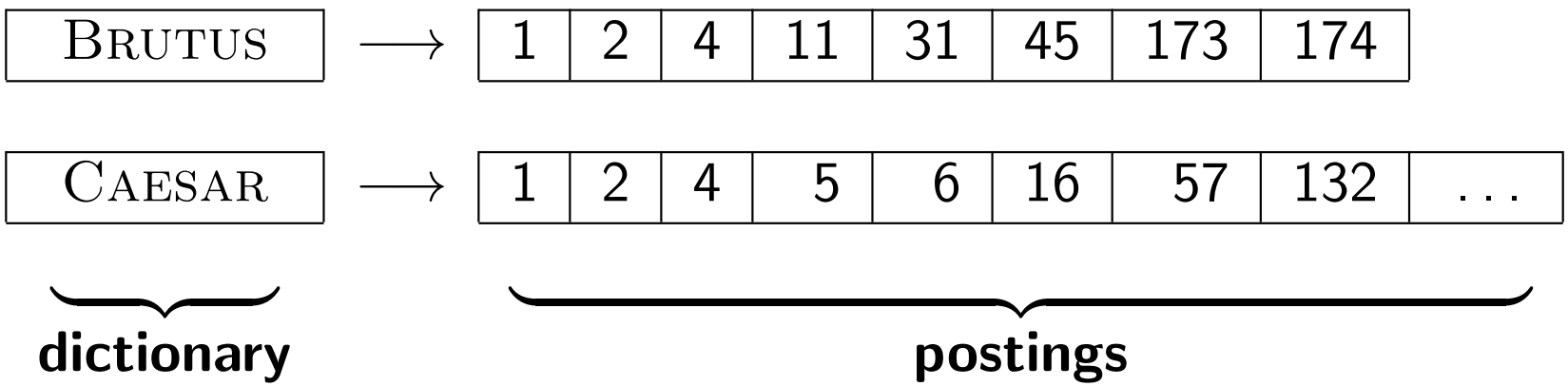
Outlook for the lectures

- We will look at the algorithms and technologies used in the modern search engines to satisfy informational needs of the users from within large document collections (usually stored on computers).

IIR 01: Boolean retrieval

- Design and data structures of a simple information retrieval system
- Queries are Boolean expressions, e.g., CAESAR AND BRUTUS
- The seach engine returns all documents that satisfy the Boolean expression.

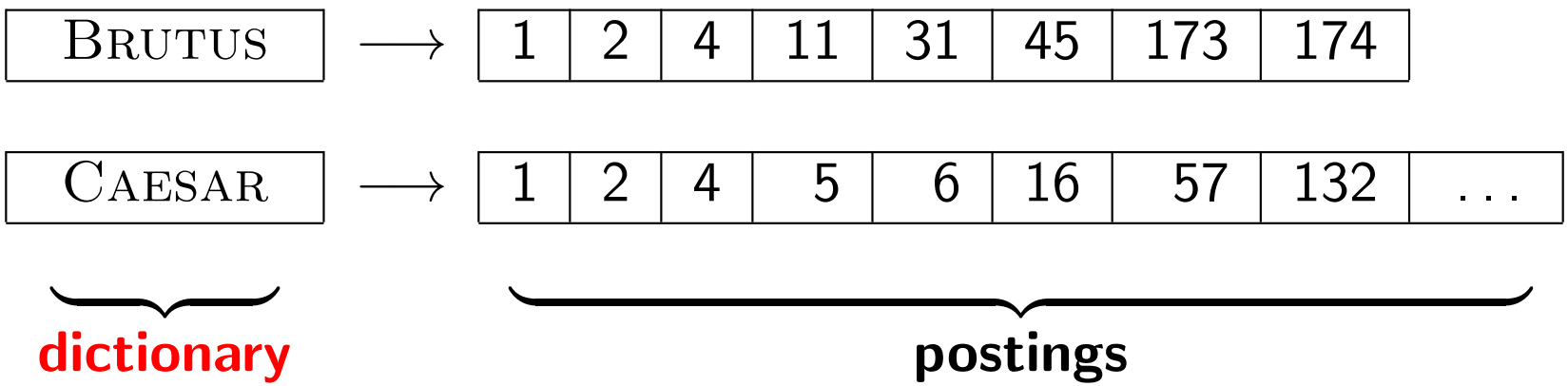
Inverted index:



IIR 01: Boolean retrieval

- Design and data structures of a simple information retrieval system
- Queries are Boolean expressions, e.g., Caesar and Brutus
- The seach engine returns all documents that satisfy the Boolean expression.

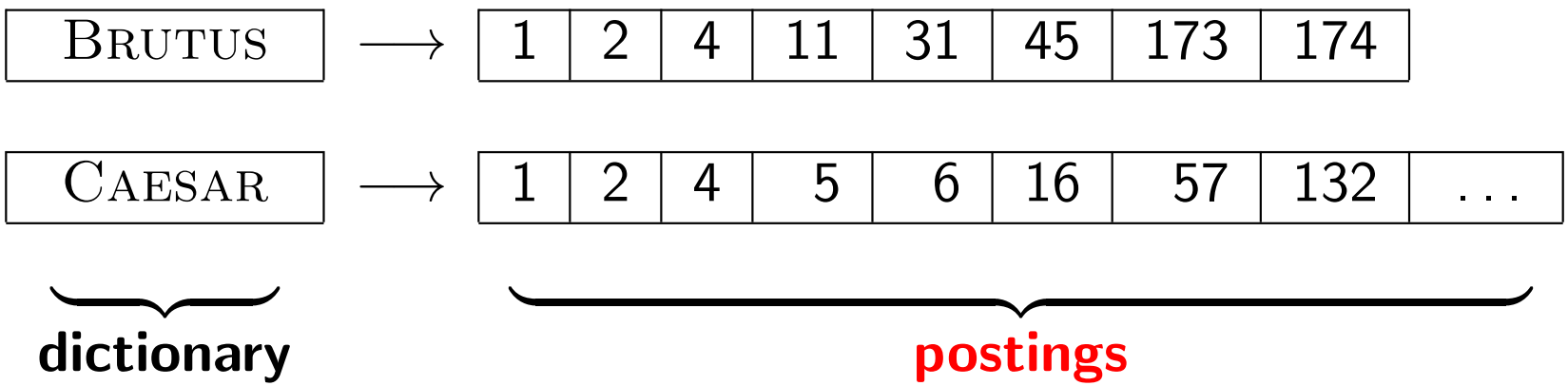
Inverted index:



IIR 01: Boolean retrieval

- Design and data structures of a simple information retrieval system
- Queries are Boolean expressions, e.g., CAESAR AND BRUTUS
- The seach engine returns all documents that satisfy the Boolean expression.

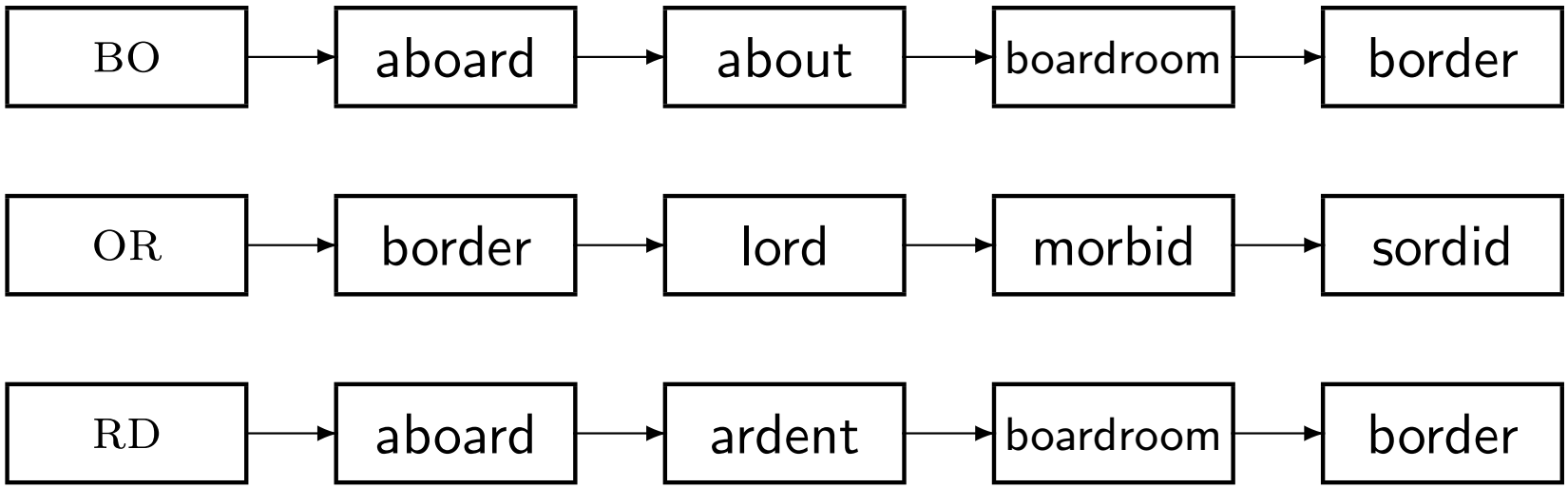
Inverted index:



IIR 02: The term vocabulary and postings lists

- Phrase queries: Stanford University
- Proximity: find Gates near Microsoft
- We need an index that captures position information for phrase queries and proximity queries.

IIR 03: Dictionaries and tolerant retrieval



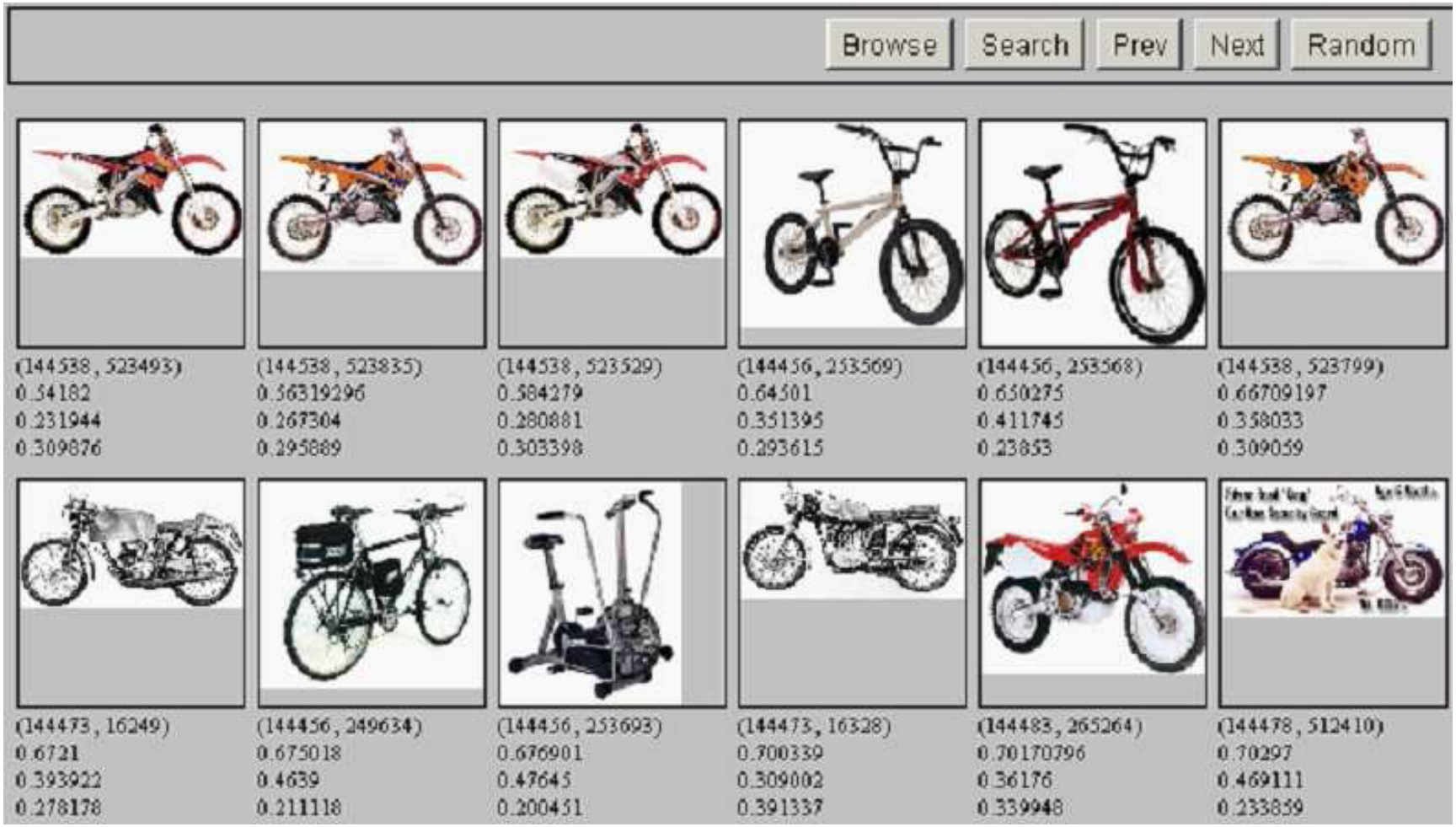
IIR 06: Scoring, term weighting and the vector space model

- Ranking search results
- Boolean queries only give inclusion or exclusion of documents.
- For ranked retrieval, we measure the proximity from query to each document.
- One formalism for doing this: the vector space model

IIR 08: Evaluation and dynamic summaries

- Benchmarks (e.g. TREC = Text Retrieval Conference)
- Measures (Precision, Recall, etc.)

IIR 09: Relevance feedback and query expansion



IIR 10: XML retrieval

- Semi-structured / structured documents vs. unstructured documents
- Can we utilize the structure of the data in IR systems?
- Databases support search for numerical range and exact match, e.g., salary $<$ 60,000 and manager=Smith.
- If your data is structured and you only need precise queries like this (numerical, exact match etc), do not use an IR system.

IIR 13: Text classification and Naive Bayes

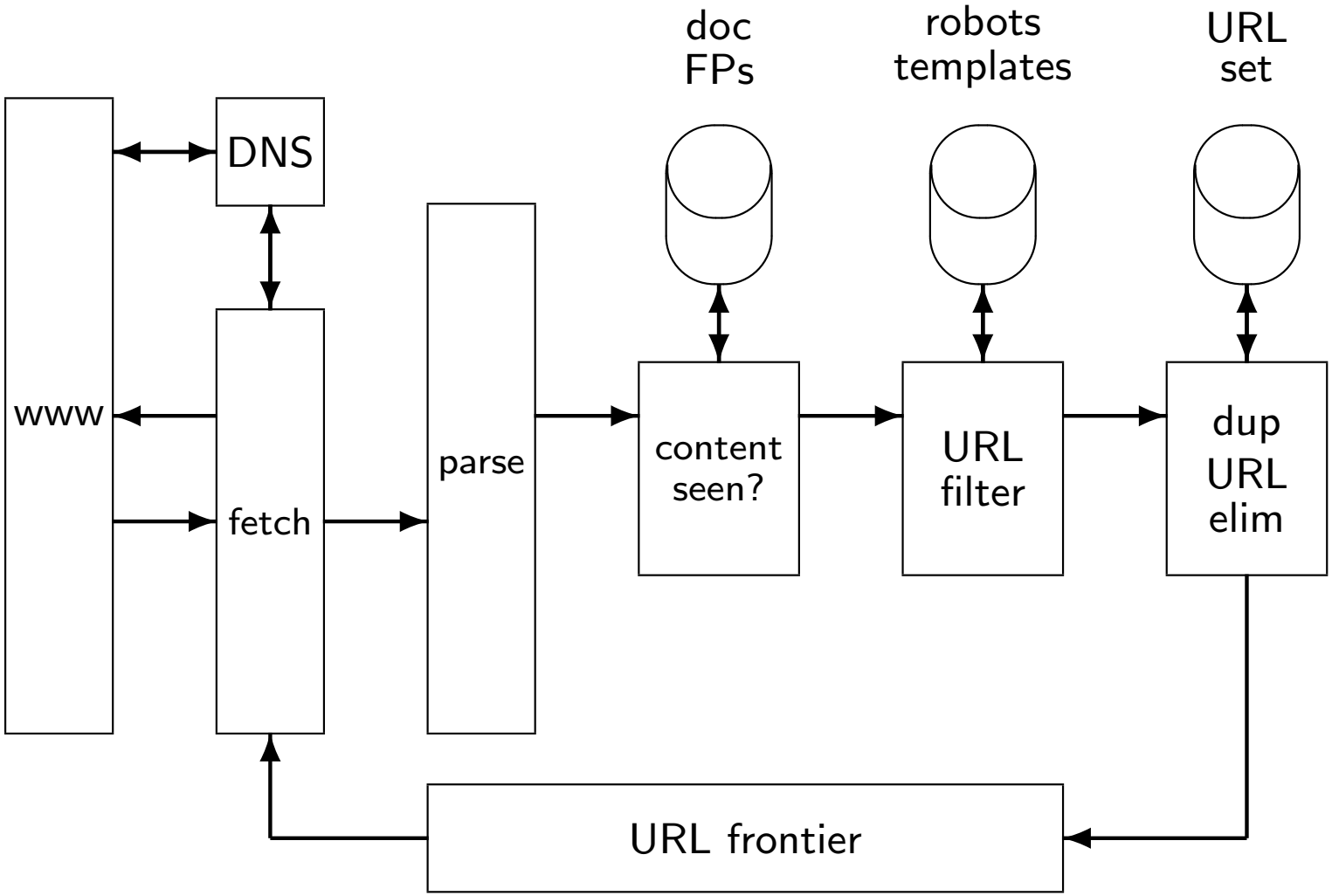
- Text classification = assigning documents automatically to predefined classes
- Examples:
  a. Language (English vs. French)
  b. Location

IIR 16: Flat clustering

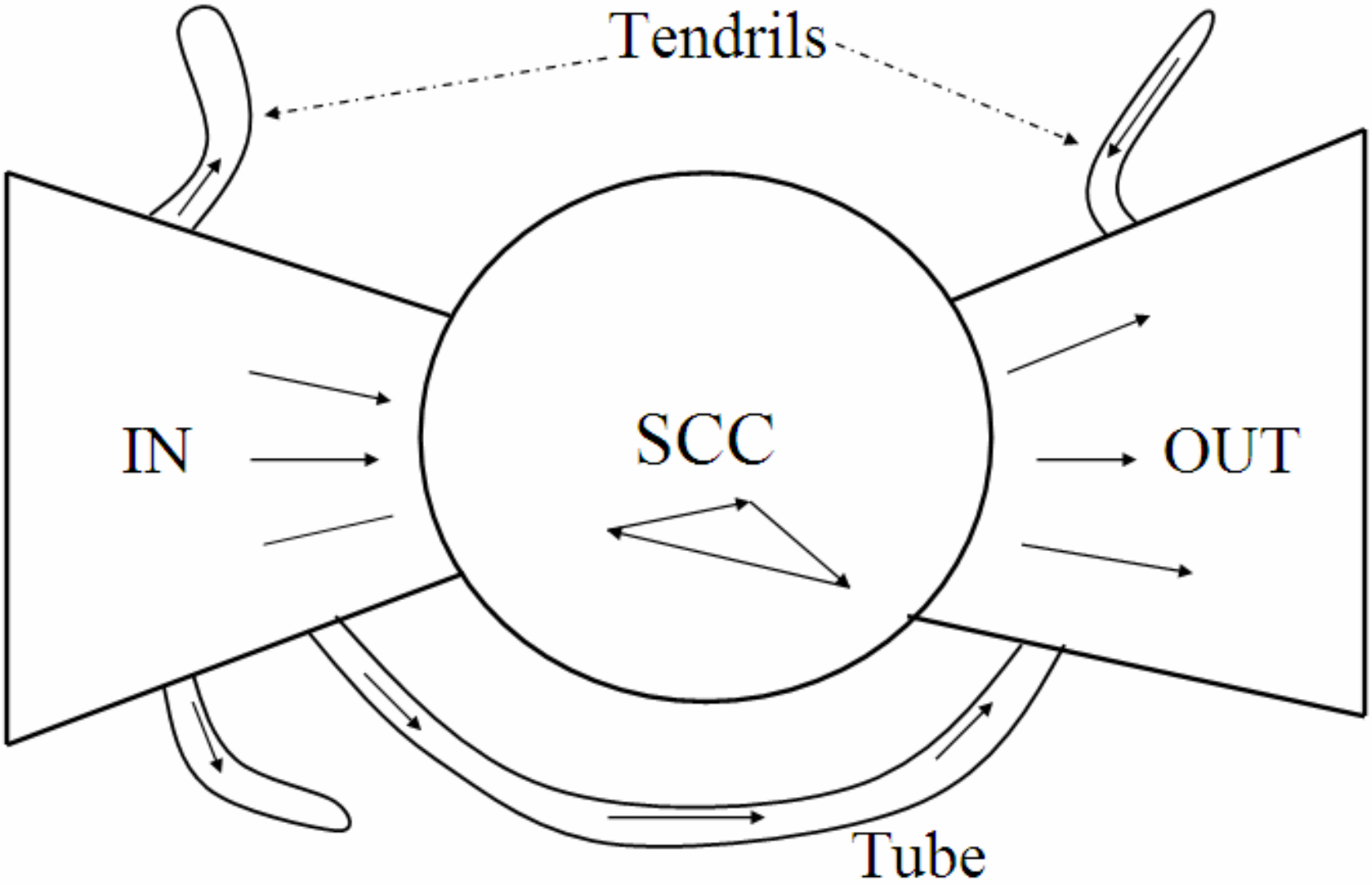
IIR 19: Web search

- Unusual and diverse documents
- Unusual and diverse users and information needs
- Beyond terms and text: exploit link analysis, user data
- How do web search engines work?
- How can we make them better?

IIR 20: Crawling



IIR 21: Link analysis / PageRank



Questions?

Thank you for your attention!