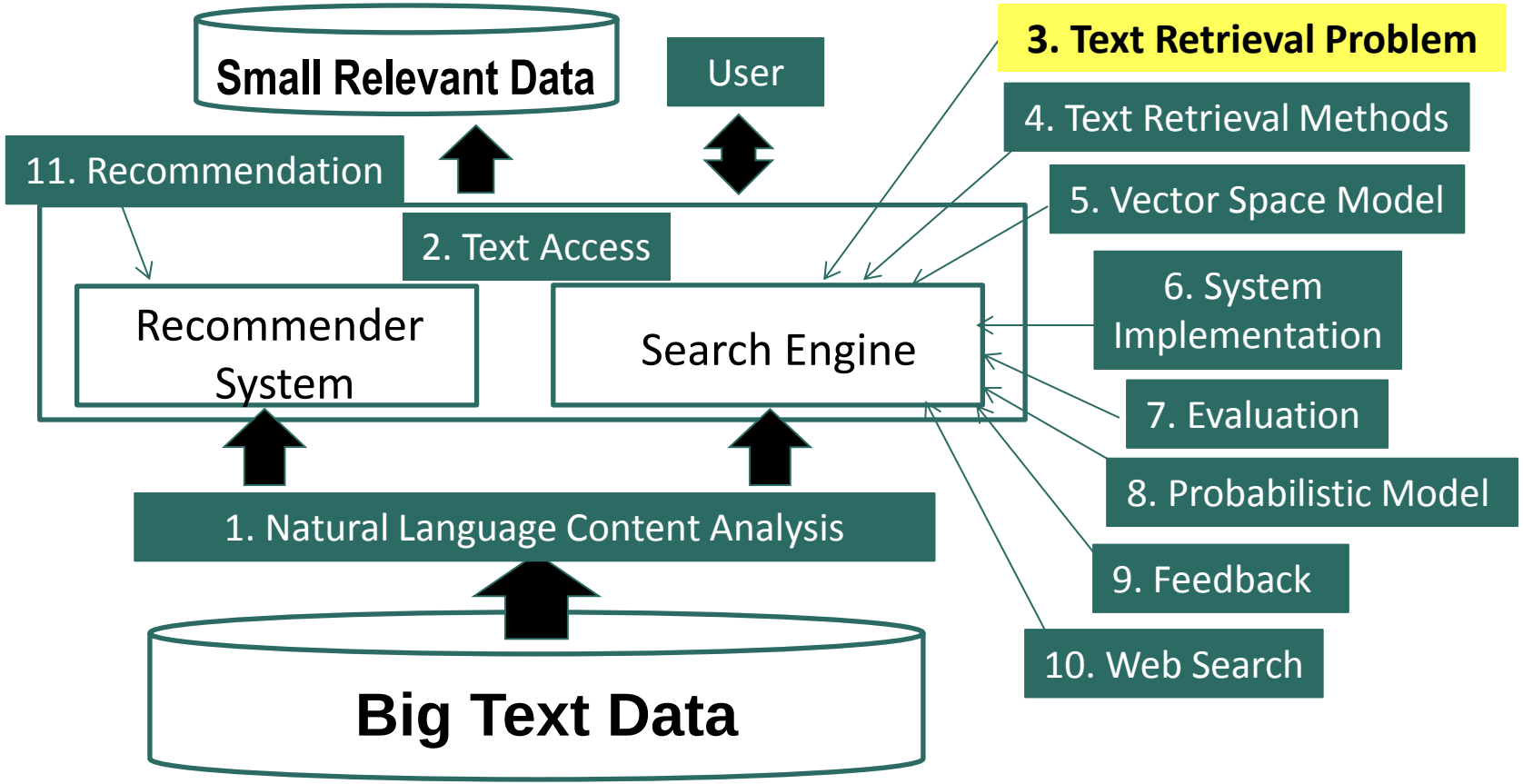
Text Retrieval and Search Engines

Text Retrieval Problem

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Course Schedule



Overview

- What is Text Retrieval?
- Text Retrieval vs. Database Retrieval
- Document Selection vs. Document Ranking

What Is Text Retrieval (TR)?

- Collection of text documents exists
- User gives a query to express the information need
- Search engine system returns relevant documents to users
- Often called “information retrieval” (IR), but IR is actually much broader
- Known as “search technology” in industry

TR vs. Database Retrieval

- Information
  - Unstructured/free text vs. structured data
  - Ambiguous vs. well-defined semantics
- Query
  - Ambiguous vs. well-defined semantics
  - Incomplete vs. complete specification
- Answers
  - Relevant documents vs. matched records
- TR is an empirically defined problem
  - Can’t mathematically prove one method is better than another
  - Must rely on **empirical evaluation** involving users!

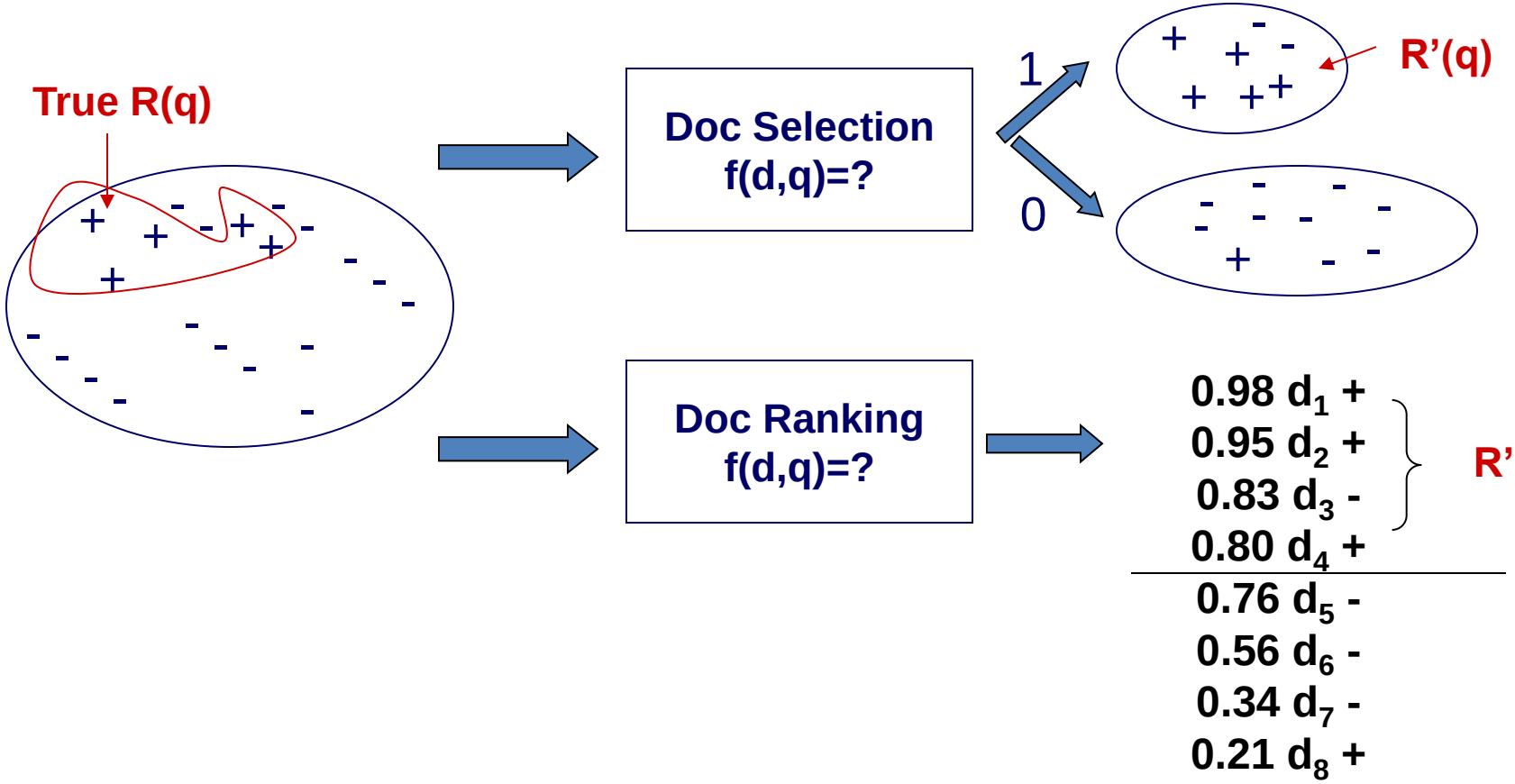
Formal Formulation of TR

- **Vocabulary**: $V=\{w_1, w_2, \ldots, w_N\}$ of language
- **Query**: $q = q_1,\ldots,q_{m,}$ where $q_i \in V$
- **Document**: $d_i = d_{i1},\ldots,d_{im_i,}$ where $d_{ij} \in V$
- **Collection**: $C= \{d_1, \ldots, d_M\}$
- **Set of relevant documents**: $R(q) \subseteq C$
  - Generally unknown and user-dependent
  - Query is a "hint" on which doc is in R(q)
- **Task** = compute R'(q), an approximation of R(q)

How to Compute R'(q)

- Strategy 1: Document selection
  - R'(q)={d∈C|f(d,q)=1}, where f(d,q) ∈{0,1} is an indicator function or binary classifier
  - System must decide if a doc is relevant or not (**absolute relevance**)
- Strategy 2: Document ranking
  - R'(q) = {d∈C|f(d,q)>θ}, where f(d,q) ∈ℜ is a relevance measure function; θ is a cutoff determined by the user
  - System only needs to decide if one doc is more likely relevant than another (**relative relevance**)

Document Selection vs. Ranking



Problems of Document Selection

- The classifier is unlikely accurate
  - "Over-constrained" query ➔ no relevant documents to return
  - "Under-constrained" query ➔ over delivery
  - Hard to find the right position between these two extremes
- Even if it is accurate, all relevant documents are not equally relevant (relevance is a matter of degree!)
  - Prioritization is needed
- Thus, ranking is generally preferred

Theoretical Justification for Ranking

- **Probability Ranking Principle** [Robertson 77]: Returning a ranked list of documents in descending order of probability that a document is relevant to the query is the optimal strategy under the following two assumptions:
  - The utility of a document (to a user) is **independent** of the utility of any other document
  - A user would browse the results **sequentially**
- Do these two assumptions hold?

Summary

- Text retrieval is an empirically defined problem
  - Which algorithm is better must be judged by users
- Document ranking is generally preferred to
  - Help users prioritize examination of search results
  - Bypass the difficulty in determining absolute relevance (users help decide the cutoff on the ranked list)
- Main challenge: design an effective ranking function **f(q,d) =?**

Additional Readings

- S.E. Robertson, The probability ranking principle in IR. *Journal of Documentation* **33**, 294-304, 1977
- C. J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworth-Heinemann, Newton, MA, USA, 1979
  - A must-read for anyone doing research in information retrieval. Chapter 6 has an in-depth discussion of PRP.