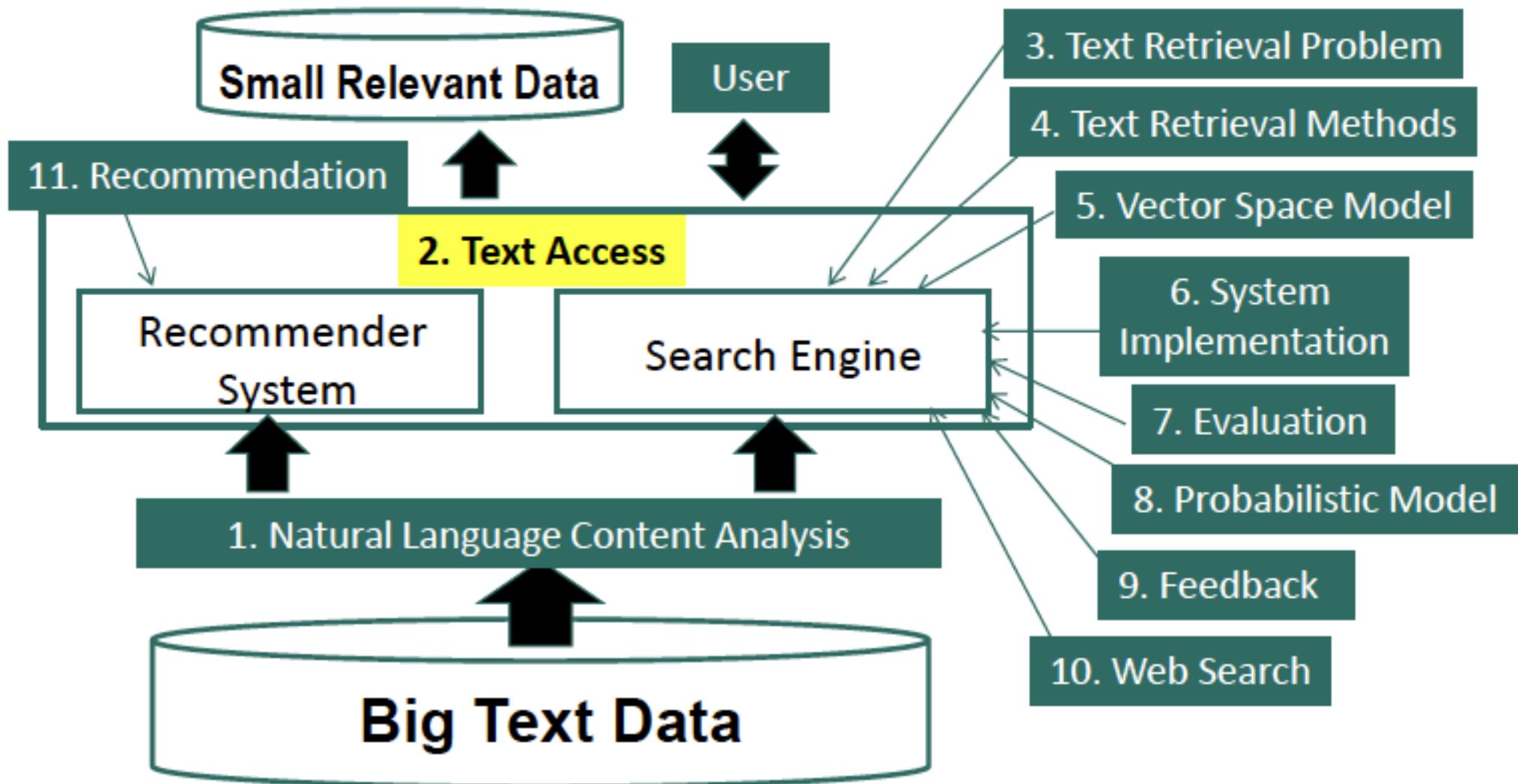


Information Retrieval & Text Mining

Text Access

Dr. Saeed UI Hassan
Information Technology University

Course Schedule



Access to Relevant Text Data

How can a text information system help users get access to the relevant text data?

- Push vs. Pull
- Querying vs. Browsing

Two Modes of Text Access: Pull vs. Push

- **Pull Mode (search engines)**
 - Users take initiative
 - Ad hoc information need
- **Push Mode (recommender systems)**
 - Systems take initiative
 - Stable information need or system has good knowledge about a user's need

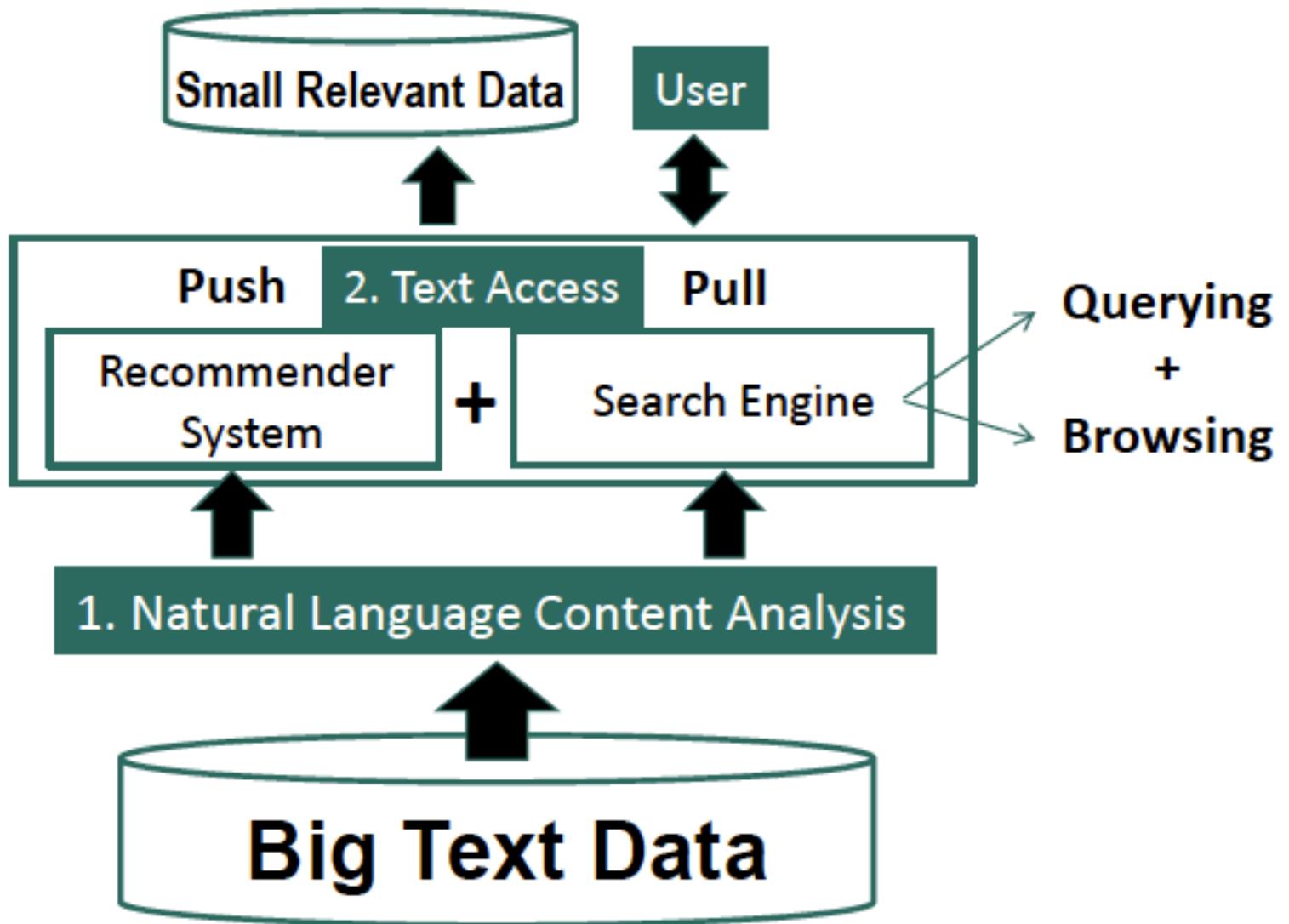
Pull Mode: Querying vs. Browsing

- **Querying**
 - User enters a (keyword) query
 - System returns relevant documents
 - Works well when the user knows what keywords to use
- **Browsing**
 - User navigates into relevant information by following a path enabled by the structures on the documents
 - Works well when the user wants to explore information, doesn't know what keywords to use, or can't conveniently enter a query

Information Seeking as Sightseeing

- **Sightseeing:** Know address of an attraction?
 - Yes: take a taxi and go directly to the site
 - No: walk around or take a taxi to a nearby place then walk
- **Information seeking:** Know exactly what you want to find?
 - Yes: use the right keywords as a query and find the information directly
 - No: browse the information space or start with a rough query and then browse

Summary



Additional Reading

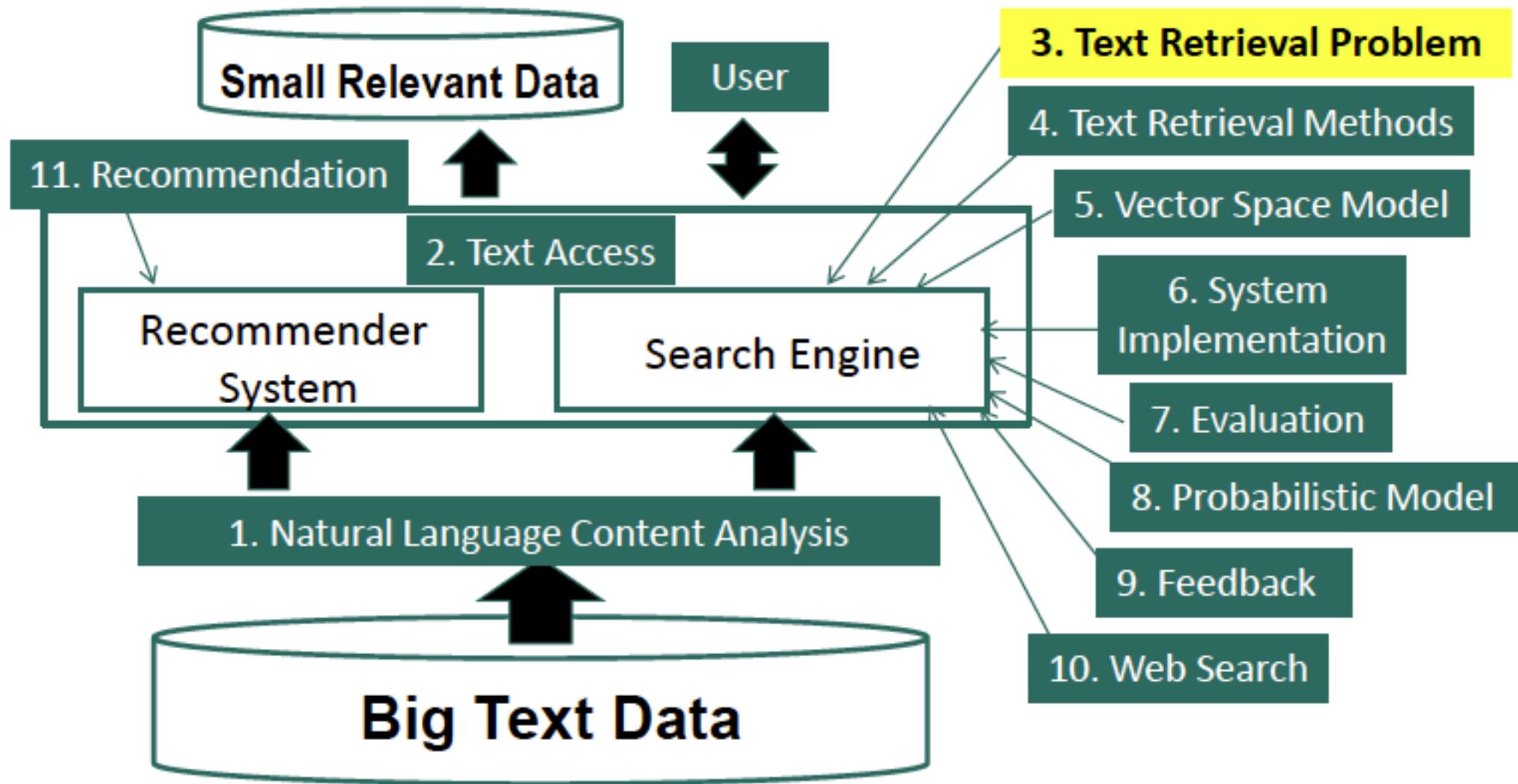
N. J. Belkin and W. B. Croft. 1992. Information filtering and information retrieval: two sides of the same coin?. *Commun. ACM* 35, 12 (Dec. 1992), 29-38.

Information Retrieval & Text Mining

Text Retrieval Problem

Dr. Saeed Ul Hassan
Information Technology University

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!

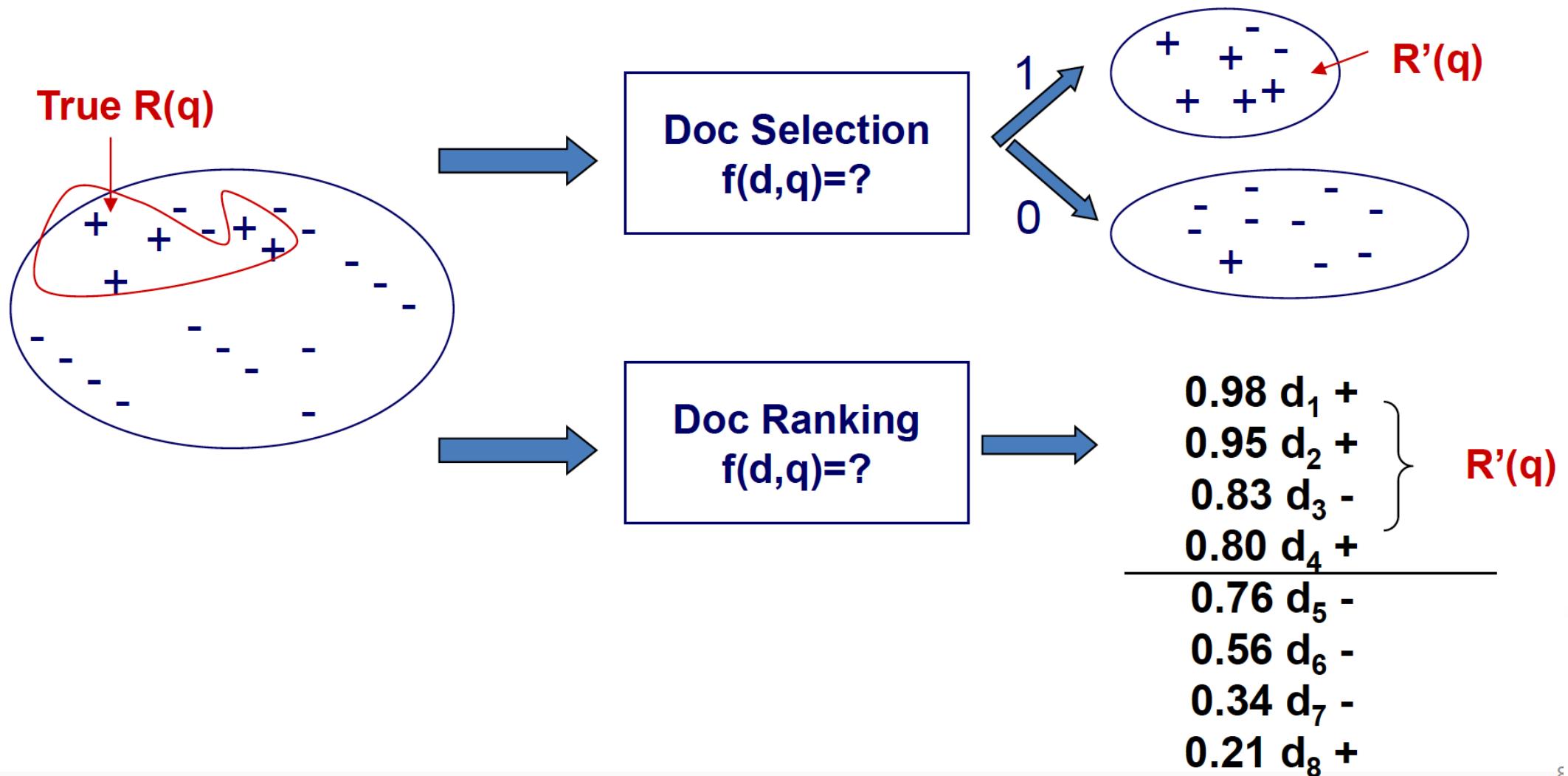
Formal Formulation of TR

- **Vocabulary:** $V = \{w_1, w_2, \dots, w_N\}$ of language
- **Query:** $q = q_1, \dots, q_m$, where $q_i \in V$
- **Document:** $d_i = d_{i1}, \dots, d_{im_i}$, where $d_{ij} \in V$
- **Collection:** $C = \{d_1, \dots, d_M\}$
- **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
 - $R'(q) = \{d \in C \mid f(d, q) = 1\}$, where $f(d, q) \in \{0, 1\}$ is an indicator function or binary classifier
 - System must decide if a doc is relevant or not (**absolute relevance**)
- Strategy 2: Document ranking
 - $R'(q) = \{d \in C \mid f(d, q) > \theta\}$, where $f(d, q) \in \mathbb{R}$ 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.