

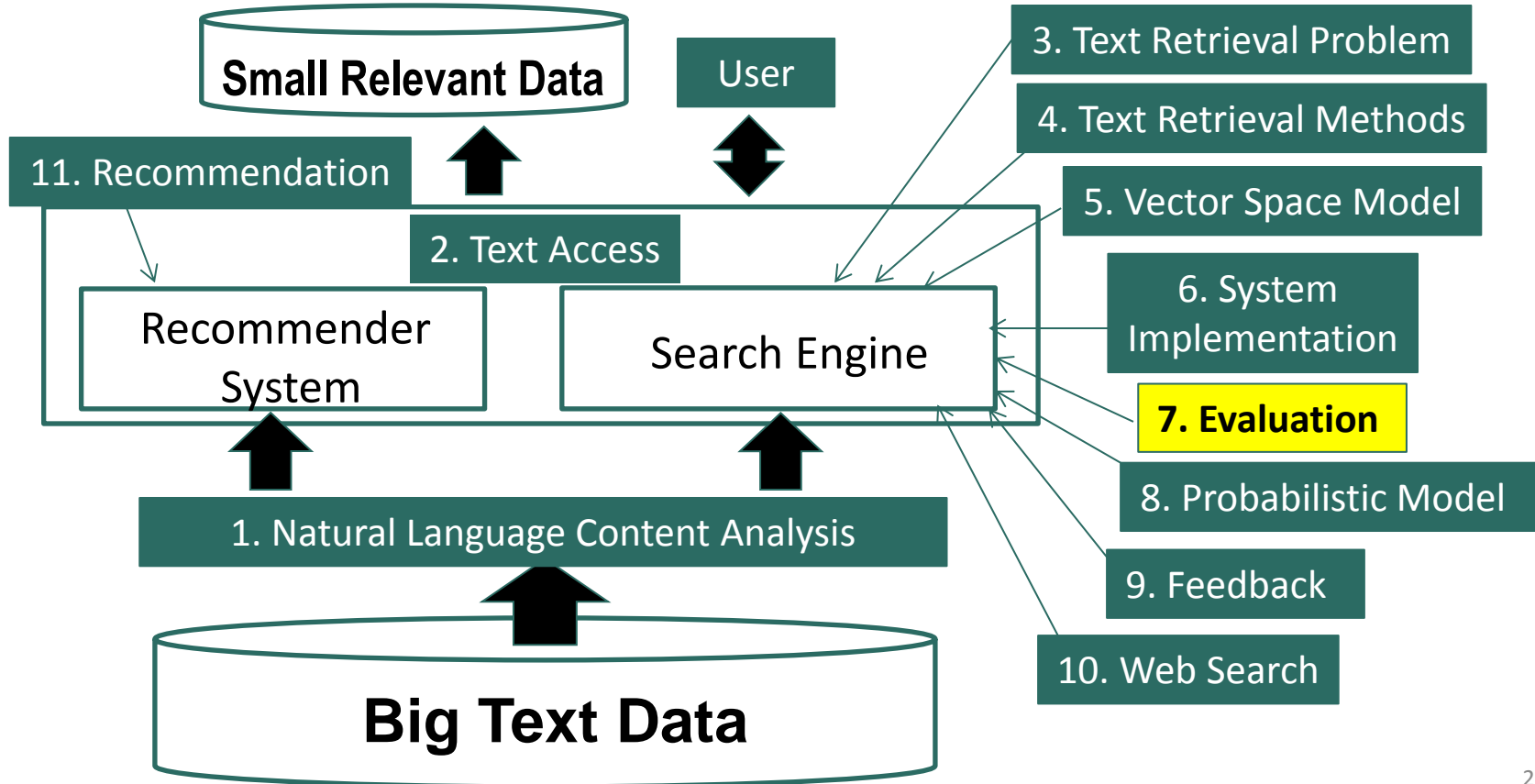


Text Retrieval and Search Engines

Evaluation of TR Systems

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

Evaluation of Text Retrieval Systems



Why Evaluation?

- Reason 1: Assess the actual utility of a TR system
 - Measures should reflect the utility to users in a real application
 - Usually done through user studies (interactive IR evaluation)
- Reason 2: Compare different systems and methods
 - Measures only need to be correlated with the utility to actual users, thus don't have to accurately reflect the exact utility to users
 - Usually done through test collections (test set IR evaluation)

What to Measure?

- Effectiveness/Accuracy: how accurate are the search results?
 - Measuring a system's ability of ranking relevant documents on top of non-relevant ones
- Efficiency: how quickly can a user get the results? How much computing resources are needed to answer a query?
 - Measuring space and time overhead
- Usability: How useful is the system for real user tasks?
 - Doing user studies

The Cranfield Evaluation Methodology

- A methodology for laboratory testing of system components developed in 1960s
- Idea: Build reusable test collections & define measures
 - A sample collection of documents (simulate real document collection)
 - A sample set of queries/topics (simulate user queries)
 - Relevance judgments (ideally made by users who formulated the queries) → Ideal ranked list
 - Measures to quantify how well a system's result matches the ideal ranked list
- A test collection can then be reused many times to compare different systems

Test Collection Evaluation

