

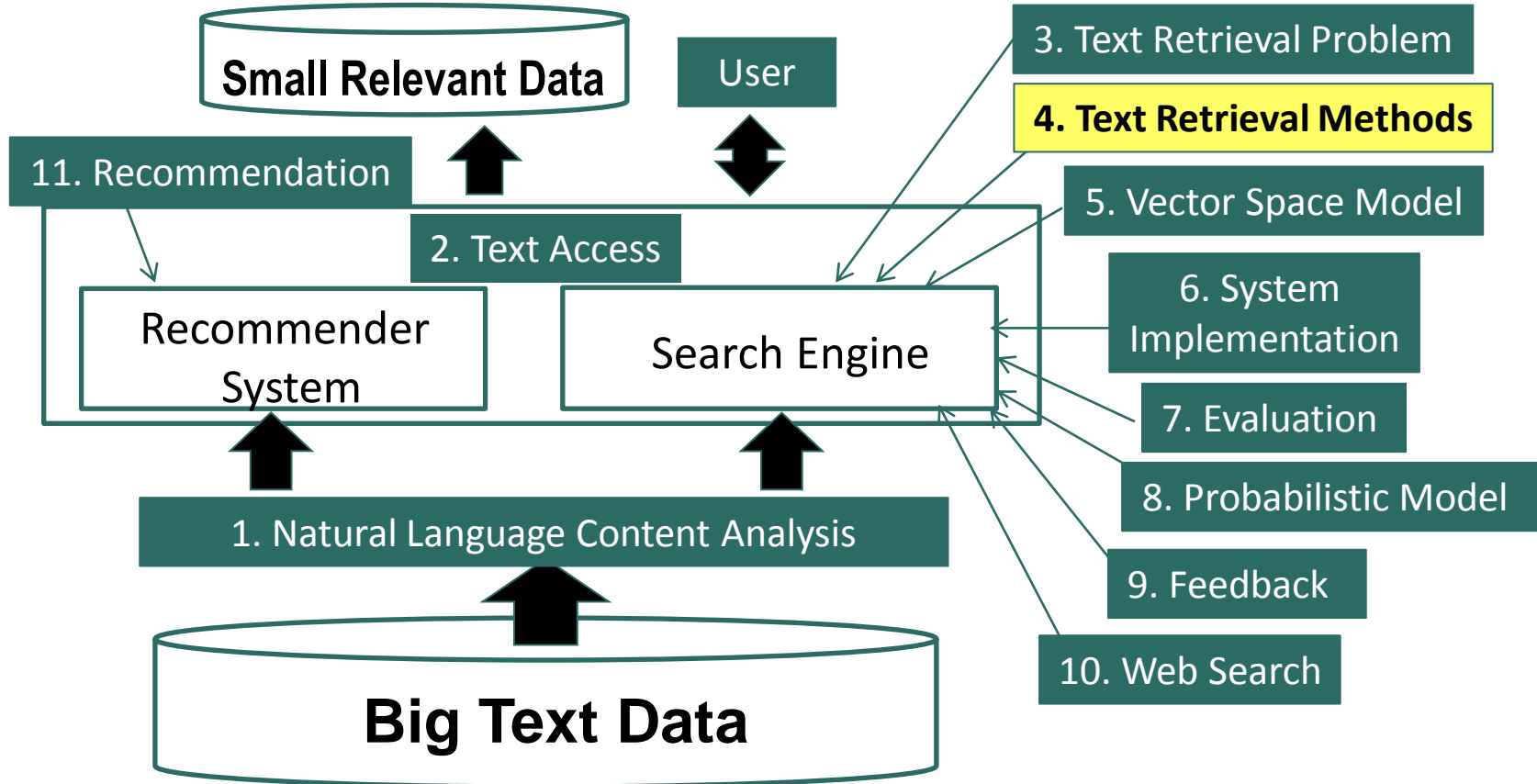


Text Retrieval and Search Engines

Overview of Text Retrieval Methods

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

Course Schedule



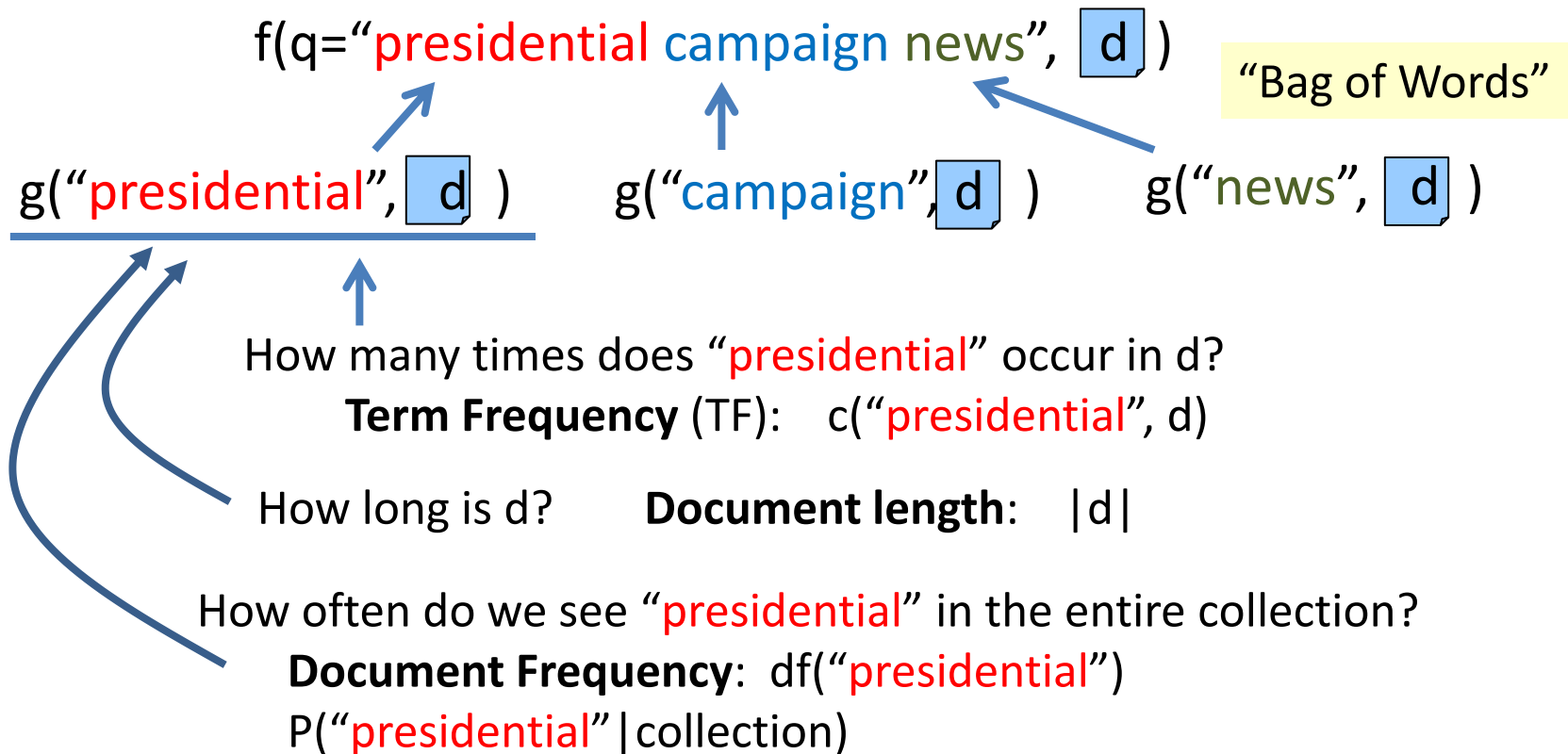
How to Design a Ranking Function

- **Query:** $q = q_1, \dots, q_m$, where $q_i \in V$
- **Document:** $d = d_1, \dots, d_n$, where $d_i \in V$
- **Ranking function:** $f(q, d) \in \mathbb{R}$
- A good ranking function should rank relevant documents on top of non-relevant ones
- Key challenge: how to measure the likelihood that document d is relevant to query q
- **Retrieval model** = formalization of relevance (give a computational definition of relevance)

Many Different Retrieval Models

- **Similarity-based models:** $f(q,d) = \text{similarity}(q,d)$
 - Vector space model
- **Probabilistic models:** $f(d,q) = p(R=1 \mid d,q)$, where $R \in \{0,1\}$
 - Classic probabilistic model
 - Language model
 - Divergence-from-randomness model
- **Probabilistic inference model:** $f(q,d) = p(d \rightarrow q)$
- **Axiomatic model:** $f(q,d)$ must satisfy a set of constraints
- These different models tend to result in similar ranking functions involving similar variables

Common Ideas in State of the Art Retrieval Models



Which Model Works the Best?

- When optimized, the following models tend to perform equally well [Fang et al. 11]:
 - **Pivoted length normalization**
 - **BM25**
 - **Query likelihood**
 - **PL2**
- BM25 is most popular

Summary

- Design of ranking function $f(q,d)$ pre-requires a computational definition of relevance (retrieval model)
- Many models are equally effective with no single winner
- State of the art ranking functions tend to rely on
 - Bag of words representation
 - Term Frequency (TF) and Document Frequency (DF) of words
 - Document length

Additional Readings

- Detailed discussion and comparison of state of the art models
 - Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2, Article 7 (April 2011)
- Broad review of different retrieval models
 - ChengXiang Zhai, *Statistical Language Models for Information Retrieval*, Morgan & Claypool Publishers, 2008. (Chapter 2)