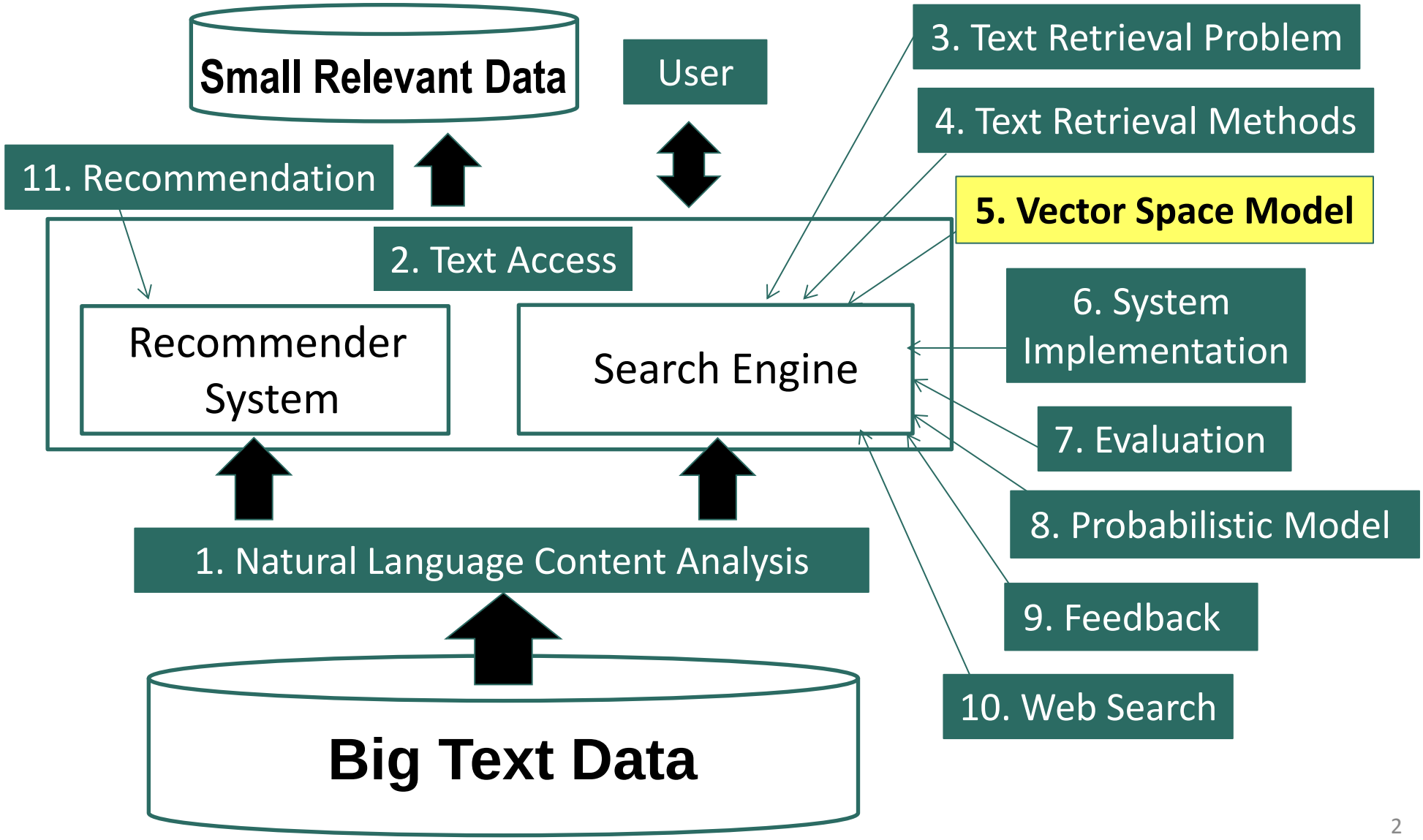
Text Retrieval and Search Engines

Doc Length Normalization

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

Course Schedule



What about Document Length?

Query = "news about presidential campaign"

d4

… **news** of **presidential campaign** …
… **presidential** candidate …

100 words

d6 > d4?

d6

… **campaign**…….. **campaign**…….. **5000 words** …..
……
……**news**……
……
…… **news**…..
……
…… ……
…… **presidential** …… **presidential**……

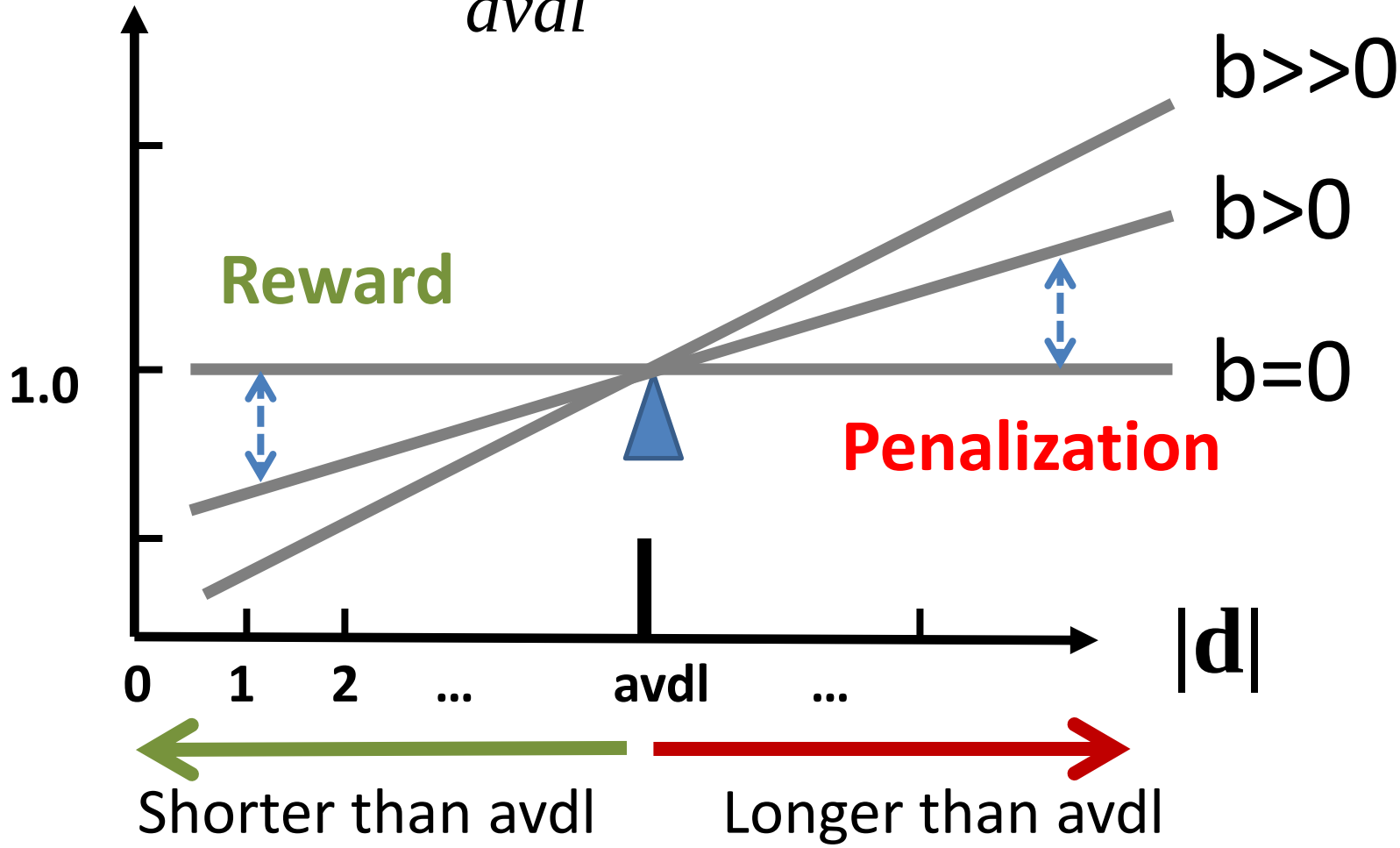
Document Length Normalization

- Penalize a long doc with a doc length nomalizer
  - Long doc has a better chance to match any query
  - Need to avoid over-penalization
- A document is long because
  - it uses more words → more penalization
  - it has more contents → less penalization
- Pivoted length normalizer: average doc length as “pivot”
  - Normalizer = 1 if |d| =average doc length (avdl)

Pivoted Length Normalization

$$normalizer = 1 - b + b\frac{|d|}{avdl}$$

$$b \in [0,1]$$



State of the Art VSM Ranking Functions

- Pivoted Length Normalization VSM [Singhal et al 96]

$$f(q,d)=\sum_{w\in q\cap d} c(w,q)\frac{\ln[1+\ln[1+c(w,d)]]}{1-b+b\frac{|d|}{avdl}}\log\frac{M+1}{df(w)}$$

- BM25/Okapi [Robertson & Walker 94]

$$b\in[0,1]$$
$$k_1,k_3\in[0,+\infty)$$

$$f(q,d)=\sum_{w\in q\cap d} c(w,q)\frac{(k+1)c(w,d)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})}\log\frac{M+1}{df(w)}$$

Further Improvement of VSM?

- Improved instantiation of **dimension?**
  - stemmed words, stop word removal, phrases, latent semantic indexing (word clusters), character n-grams, …
  - bag-of-words with phrases is often sufficient in practice
  - Language-specific and domain-specific tokenization is important to ensure “normalization of terms”
- Improved instantiation of **similarity function?**
  - cosine of angle between two vectors?
  - Euclidean?
  - dot product seems still the best (sufficiently general especially with appropriate term weighting)

Summary of Vector Space Model

- Relevance(q,d) = similarity(q,d)
- Query and documents are represented as vectors
- Heuristic design of ranking function
- Major term weighting heuristics
  - TF weighting and transformation
  - IDF weighting
  - Document length normalization
- BM25 and Pivoted normalization seem to be most effective

Additional Readings

- A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of ACM SIGIR 1996.* *http://singhal.info/pivoted-dln.pdf*
- S. E. Robertson , S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, *Proceedings of ACM SIGIR 1994*.

http://staff.city.ac.uk/~sb317/papers/robertson_walker_sigir94.pdf