PURDUE UNIVERSITY Department of Computer Science

CS47300: Web Information Search and Management

Prof. Chris Clifton

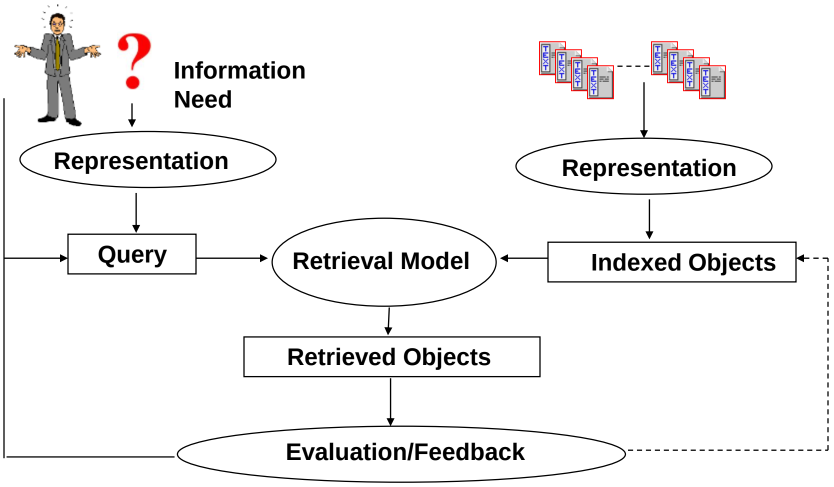
2 September 2020

Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

Indiana Center for Database Systems ™

PURDUE UNIVERSITY Department of Computer Science

Retrieval Models


Information Need

Representation

Query

Retrieval Model

Representation

Indexed Objects

Retrieved Objects

Evaluation/Feedback

PURDUE UNIVERSITY
Department of Computer Science

Overview of Retrieval Models

Retrieval Models	
• Boolean	
• Vector space	
– Basic vector space	SMART, LUCENE
– Extended Boolean	
• Probabilistic models	
– Statistical language models	Lemur Project (Indri, Galago)
– Two Possion model	Okapi
– Bayesian inference networks	Inquery
• Citation/Link analysis models	
– Page rank	Google
– Hub & authorities	Clever

PURDUE UNIVERSITY
Department of Computer Science

Retrieval Models: Outline

Retrieval Models

- Exact-match retrieval method
  - Unranked Boolean retrieval method
  - Ranked Boolean retrieval method
- Best-match retrieval method
  - Vector space retrieval method
  - Latent semantic indexing

Retrieval Models: Unranked Boolean

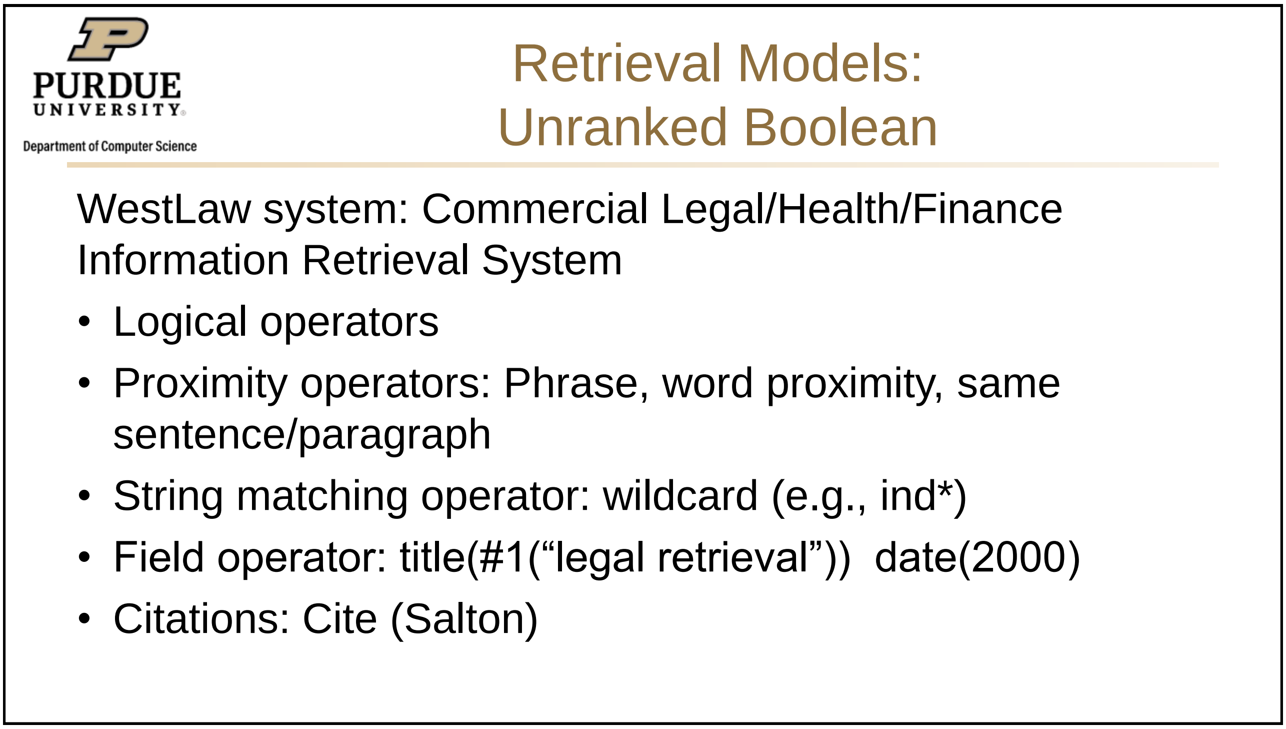
Unranked Boolean: Exact match method

- Selection Model
  - Retrieve a document iff it matches the precise query
  - Often return unranked documents (or with chronological order)
- Operators
  - Logical Operators: AND OR, NOT
  - Proximity operators:
    - #1(white house) (i.e., within one word distance, phrase)
    - #sen(Iraq weapon) (i.e., within a sentence)
  - String matching operators: Wildcard (e.g., ind* for india and indonesia)
  - Field operators: title(information and retrieval)…

Retrieval Models: Unranked Boolean

Unranked Boolean: Exact match method

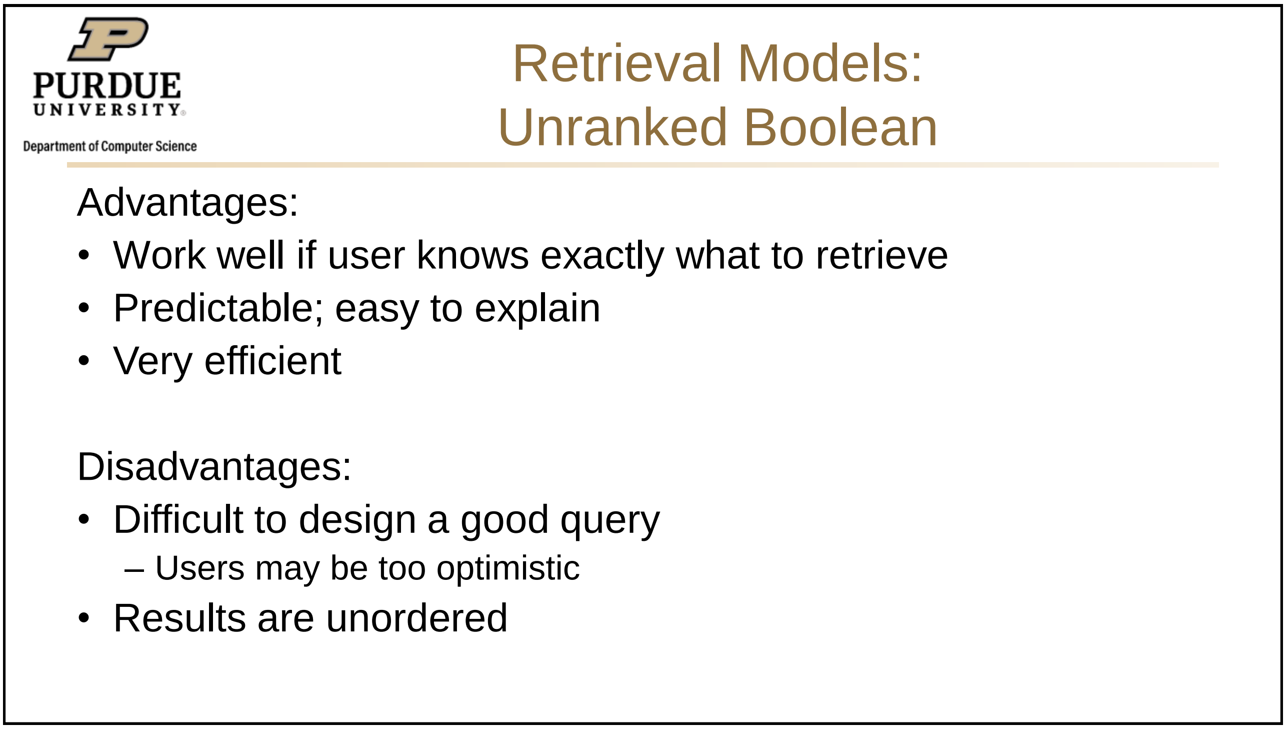
- A query example
  (#2(distributed information retrieval) OR (#1 (federated search)) AND author(#1(Jamie Callan) AND NOT (Steve))



Retrieval Models: Unranked Boolean

WestLaw system: Commercial Legal/Health/Finance Information Retrieval System

- Logical operators
- Proximity operators: Phrase, word proximity, same sentence/paragraph
- String matching operator: wildcard (e.g., ind*)
- Field operator: title(#1(“legal retrieval”)) date(2000)
- Citations: Cite (Salton)



Retrieval Models: Unranked Boolean

Advantages:

- Work well if user knows exactly what to retrieve
- Predictable; easy to explain
- Very efficient

Disadvantages:

- Difficult to design a good query
  - Users may be too optimistic
- Results are unordered

Retrieval Models: Unranked Boolean

Disadvantages:

- It is difficult to design the query
  - “Loose” query (information OR retrieval): Low precision
  - “Strict” query (information AND retrieval): Low recall
    - *Users may assume most/all relevant documents found*
- Results are unordered
  - Low precision queries not very useful

Retrieval Models: Ranked Boolean

Ranked Boolean: Exact match

● Similar to unranked Boolean but documents are ordered by some criterion

Reflect importance of document by its words

Retrieve docs from Wall Street Journal Collection

Query: (Thailand AND stock AND market)

Which word is more important?

Many “stock” and “market”, but fewer “Thailand”. Fewer may be more indicative

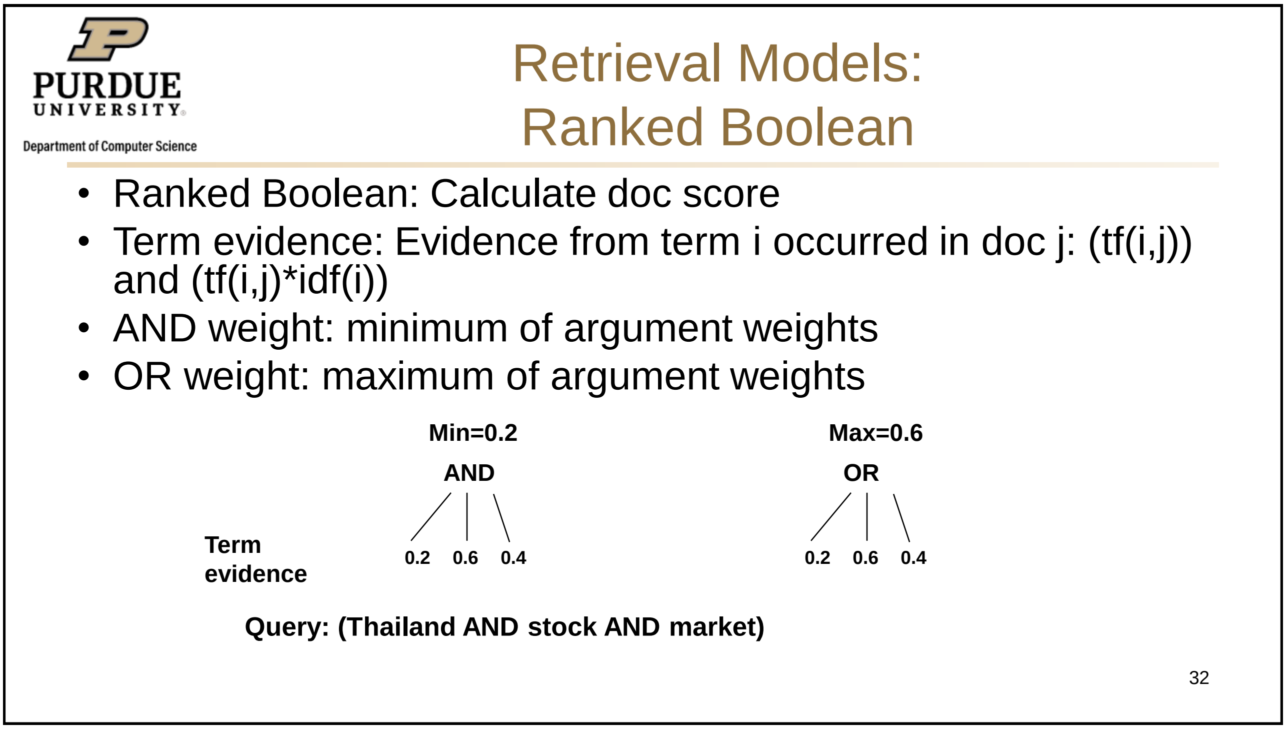
Term Frequency (TF): Number of occurrence in query/doc; larger number means more important

Inversed Document Frequency (IDF):

$$\frac{\text{Total number of docs}}{\text{Number of docs contain a term}}$$

Larger means more important

There are many variants of TF, IDF: e.g., consider document length



PURDUE UNIVERSITY

Department of Computer Science

Retrieval Models: Ranked Boolean

- Ranked Boolean: Calculate doc score
- Term evidence: Evidence from term i occurred in doc j: (tf(i,j)) and (tf(i,j)*idf(i))
- AND weight: minimum of argument weights
- OR weight: maximum of argument weights

Min=0.2

AND

0.2 0.6 0.4

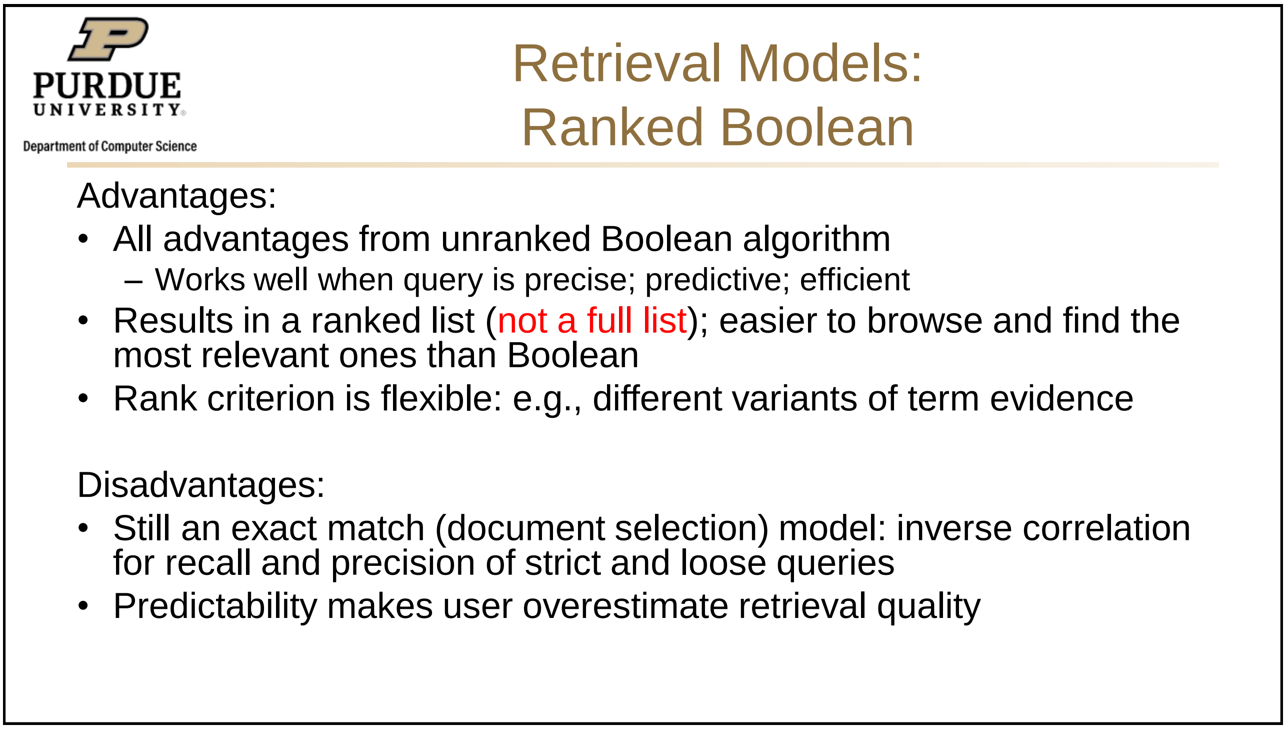
Max=0.6

OR

0.2 0.6 0.4

Term evidence

Query: (Thailand AND stock AND market)



PURDUE UNIVERSITY

Department of Computer Science

Retrieval Models: Ranked Boolean

Advantages:

- All advantages from unranked Boolean algorithm
  - Works well when query is precise; predictive; efficient
- Results in a ranked list (not a full list); easier to browse and find the most relevant ones than Boolean
- Rank criterion is flexible: e.g., different variants of term evidence

Disadvantages:

- Still an exact match (document selection) model: inverse correlation for recall and precision of strict and loose queries
- Predictability makes user overestimate retrieval quality