PURDUE UNIVERSITY

CS47300: Web Information Search and Management

Prof. Chris Clifton

27 August 2019

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

PURDUE UNIVERSITY

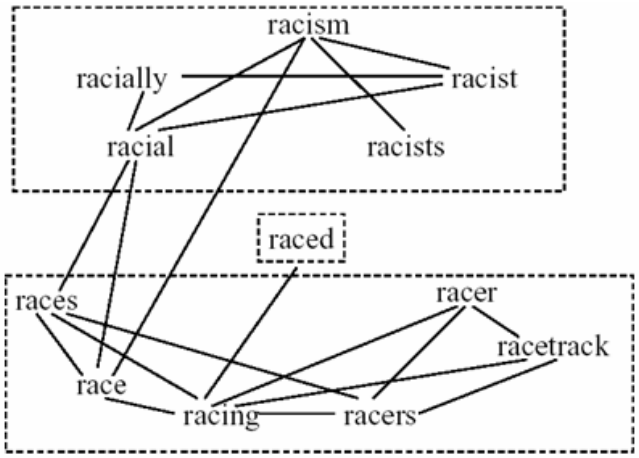
Text Representation: Word Stemming

Corpus-Based Stemming

- Hypothesis: Word variants that should be considered equally often co-occur in documents (passages or text windows) in the corpus
  - Collect the statistics of co-occurrence of words in the corpus and form the connected graph
  - Cut the graph by different methods and find the connected subgraphs to form equivalence classes

PURDUE UNIVERSITY

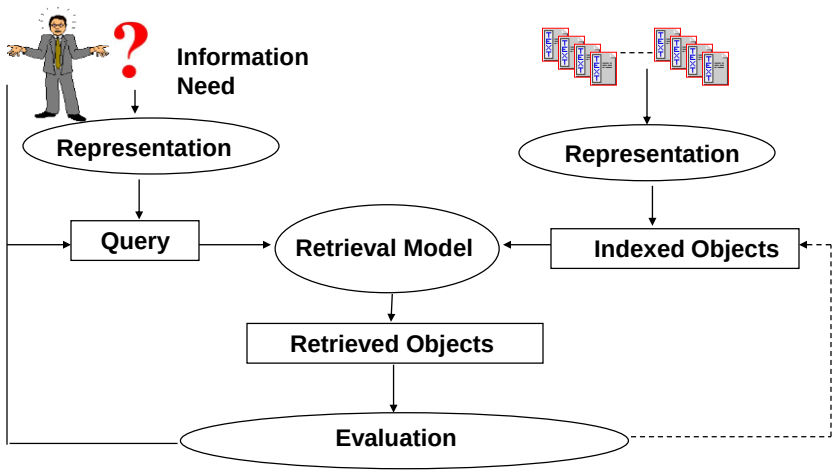
Text Representation: Word Stemming



(Xu & Croft, 1998)

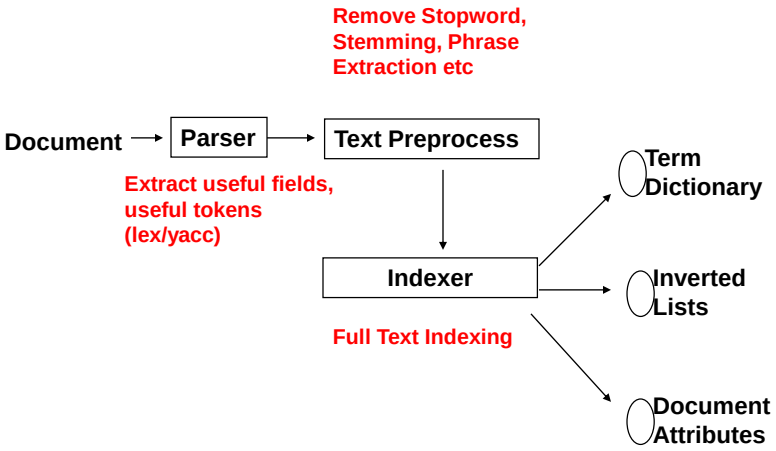
PURDUE UNIVERSITY

AD-hoc IR: Basic Process



PURDUE UNIVERSITY

Text Representation: Process of Indexing



PURDUE UNIVERSITY

Text Representation: Inverted Lists

Inverted lists are one of the most common indexing techniques

- Source file: collection organized by documents
- Inverted list file: collection organized by term
  - one record per term, the lists of documents that contain the specific term
- Possible actions with inverted lists
  - OR: the union of lists
  - AND: the intersection of lists

PURDUE UNIVERSITY

Text Representation: Inverted Lists

Doc ID	Text
1	kids question noting in 1960s
2	young man question everything in 1970s
3	kids question questions in 1980s
4	young man question nothing in 2000s

Documents

Term ID	Term	Documents
1	kids	1,3
2	question	1,2,3,4
3	nothing	1,4
4	in	1,2,3,4
5	19060s	1
6	young	2,4
7	man	2,4
8	everything	2
9	1970s	2
10	questions	3
11	1980s	3
11	2000s	4

Inverted Lists

PURDUE UNIVERSITY

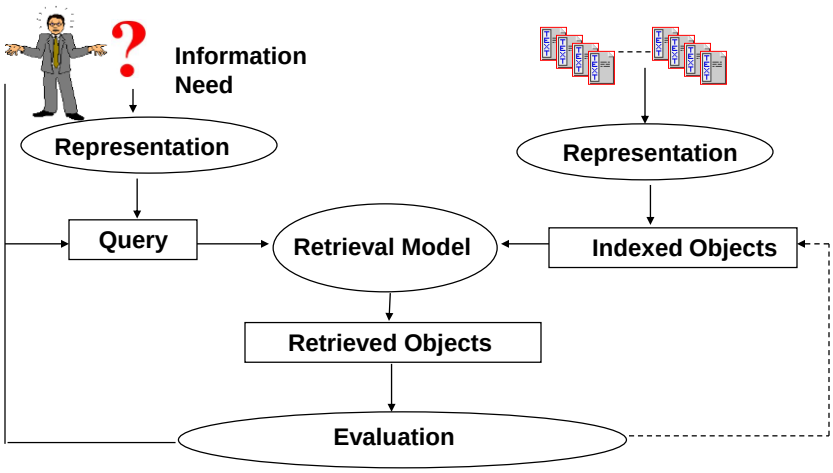
Text Representation: Inverted Lists

Many engineering details

- Update inverted lists: delete/insert a term or document
- Compression: trade off between I/O time and CPU time
- Add more information such as position information
- Take CS34800 and CS44800 for even more

PURDUE UNIVERSITY

AD-hoc IR: Basic Process



PURDUE UNIVERSITY

Evaluation: What do we Evaluate?

- Effectiveness
  - How do we define *effective*?
  - Where can we find the correct answers?
- Efficiency
  - Retrieval speed?
  - Storage space?

  Particularly important for large-scale real-world system
- Usability
  - What do real users really want?
  - Is user interface important to IR evaluation?

11

PURDUE UNIVERSITY

Evaluation Criteria

- Effectiveness
  - Favor returned document ranked lists with more relevant documents at the top
  - Objective measures
    - Recall and Precision
    - Mean-average precision
    - Rank based precision

For documents in a subset of a ranked lists, if we know the truth

	Retrieved	Not retrieved
Relevant	Relevant docs retrieved	Relevant docs not retrieved
Irrelevant	Irrelevant docs retrieved	Irrelevant docs not retrieved

$$\text{Precision} = \frac{\text{Relevant docs retrieved}}{\text{Retrieved docs}}$$

$$\text{Recall} = \frac{\text{Relevant docs retrieved}}{\text{Relevant docs}}$$

12

PURDUE UNIVERSITY

Evaluation: "Ground Truth"

	Retrieved	Not retrieved
Relevant	Relevant docs retrieved	Relevant docs not retrieved
Irrelevant	Irrelevant docs retrieved	Irrelevant docs not retrieved

Question: How to find all relevant documents?

Difficult for Web, but possible on controllable corpus

- How to find all relevant documents? (difficult to check one by one)
- Judgers may have inconsistent decisions (subjective judgment)

The Pooling process

PURDUE UNIVERSITY

Evaluation: Inconsistent Judgement

- People may not agree on the “right” answer
  - Some think document is relevant to query, others don’t
- Discussion among multiple judgers to reduce bias
- Combine judgments from multiple judgers
  - Majority vote
- *If it is hard to decide for human judges, it is likely to be hard for an automatic system*

14

PURDUE UNIVERSITY

Evaluation: Pooling Strategy

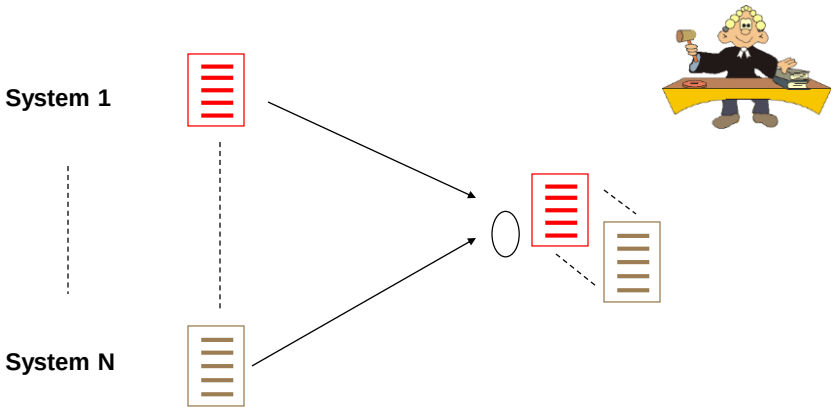
- Retrieve documents using multiple methods
- Judge top *n* documents from each method
- Whole retrieved set is the union of top retrieved documents from all methods
- Problems: the judged relevant documents may not be complete

- *It is possible to estimate the total number of relevant documents by random sampling*

15

PURDUE UNIVERSITY

Evaluation: Pooling Strategy



PURDUE UNIVERSITY

Unranked Measures:

- Precision : $\frac{\#\ Relevant\ Retrieved}{\#\ Retrieved}$
- Recall : $\frac{\#\ Relevant\ Retrieved}{\#\ Relevant}$
- F1 score : $\frac{2PR}{P+R}$

17

PURDUE UNIVERSITY

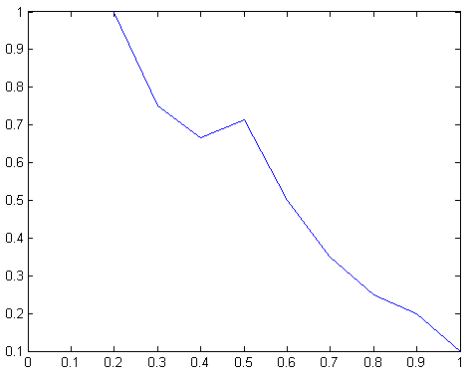
Evaluation

- Evaluate a ranked list
  - Precision at Recall
- Evaluate at every relevant document

+
+
-
+
+
-
+

Not Retrieved: +++++

Precision	Recall
1	0.1
1	0.2
0.667	0.2
0.75	0.3
0.8	0.4
0.667	0.4
0.714	0.5



18

PURDUE UNIVERSITY

Ranked Metrics
Single number

- Mean average precision
  - Calculate precision at each relevant document; average over all precision values
  - Mean average precision – average over many queries
- 11-point interpolated average precision
  - Calculate precision at standard recall points (e.g., 10%, 20%...); smooth the values; estimate 0 % by interpolation
  - Average the results

20

PURDUE UNIVERSITY

Evaluation: Single Value Metrics

- Rank based precision
  - Calculate precision at top ranked documents (e.g., 5, 10, 15…)
  - Desirable when users care more for top ranked documents
- Mean Reciprocal Rank
  - Reciprocal Rank: 1/rank (position in list) of first relevant document
  - MRR: Average Reciprocal Rank over many queries

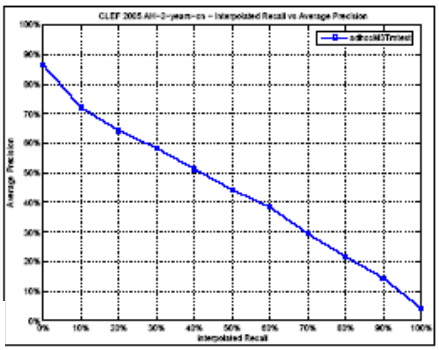
21

PURDUE UNIVERSITY

Evaluation: Example

Interpolated Recall (%)	Precision Averages (%)
0	86.49
10	72.16
20	64.25
30	58.40
40	51.33
50	44.30
60	38.43
70	29.43
80	21.68
90	14.40
100	4.15
Average precision (non-interpolated) for all relevant documents (averaged over queries)	
	43.06

Docs Cutoff Levels	Precision at DCL (%)
5 docs	72.50
10 docs	67.00
15 docs	61.83
20 docs	59.25
30 docs	55.42
100 docs	39.75
200 docs	30.92
500 docs	19.54
1000 docs	12.02
R-Precision (precision after R document retrieved, where R = Relevant retrieved)	
	44.99



22

PURDUE UNIVERSITY

Evaluation: TREC

TREC collections with queries and relevance judgment

- **TREC CDs 1-5**: 1.5 millions docs, 5GB, news and government reports (e.g., AP, WSJ, Dept of Energy abstracts)
- **TREC WT10g**: crawled from Web (open domain), 1.7 million docs, 10GB
- **TREC Terabyte**: crawled from U.S. government Web pages, 25 million docs, 426 GB
- *All have more than 100 queries with relevance judgment*

23

PURDUE UNIVERSITY

Evaluation: TREC

- TREC query example

<title> airport security

<desc> Description:
What security measures are in effect or are proposed to go into effect in airports?

<narr> Narrative:
A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.

24

PURDUE UNIVERSITY

Evaluation: TREC

- TREC relevance judgment example

```
451 WTX058-B50-85 0
451 WTX059-B06-411 0
451 WTX059-B07-154 0
451 WTX059-B09-203 0
451 WTX059-B11-245 0
451 WTX059-B30-262 1
451 WTX059-B37-11 0
451 WTX059-B37-149 1
451 WTX059-B37-217 0
451 WTX059-B37-268 0
451 WTX059-B37-27 0
```

25

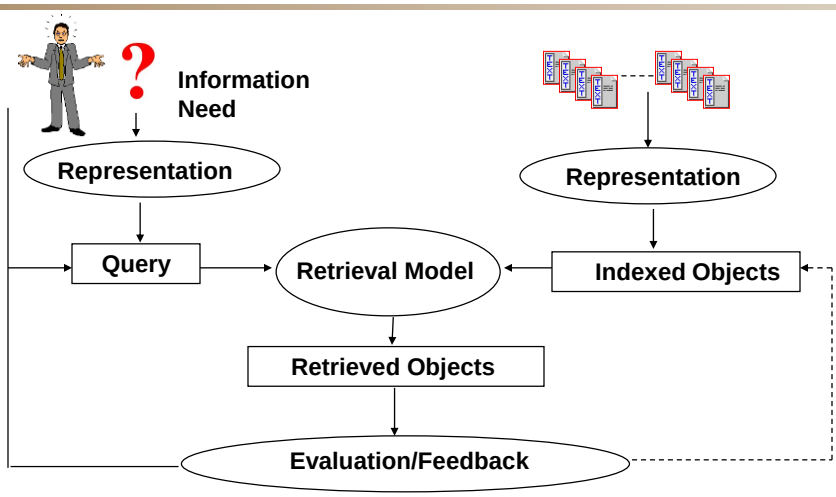
PURDUE UNIVERSITY

Review to date:

- Basic Concepts of Information Retrieval:
- Task Definition of Ad-hoc IR
  - Terminologies and Concepts
  - Overview of Retrieval Models
- Text representation
  - Indexing
  - Text preprocessing
- Evaluation
  - Evaluation methodology
  - Evaluation metrics

PURDUE UNIVERSITY

Retrieval Models



PURDUE UNIVERSITY

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

30

PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

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)…

PURDUE UNIVERSITY

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))

PURDUE UNIVERSITY

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)

PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

Retrieval Models: Ranked Boolean

Ranked Boolean: Exact match

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

Retrieve docs from Wall Street Journal Collection

Query: (Thailand AND stock AND market)

Reflect importance of document by its words

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

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

Query: (Thailand AND stock AND market)