Monsoon 2021

Boolean Retrieval - 2

- Boolean Query Processing and Optimization

Dr. Rajendra Prasath

Indian Institute of Information Technology Sri City, Chittoor

> Topics to be covered

- Recap:
  - Inverted Index Construction
  - Term - Document Matrix

- Boolean Operators
- Boolean Retrieval
- Boolean Queries

- Text Collection / Corpora
- Evaluation Strategy

  - More topics to come up … Stay tuned …!!

Overview

Recap: Information Retrieval

- **Information Retrieval (IR)** is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- These days we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval
  - and so on . . .

Recap: Look at 3 documents

d_1- **Darjeeling** is a city and a municipality in the Indian state of West Bengal. It is located in the Lesser Himalayas at an elevation of 6,700 feet

d_2- **Darjeeling** is noted for its tea industry, its views of Kangchenjunga, the world's third-highest mountain, and the **Darjeeling** Himalayan Railway, a UNESCO World Heritage Site

d_3- **Darjeeling** is the headquarters of the **Darjeeling** District which has a partially autonomous status within the state of West Bengal. It is also a tourist destination in India

Boolean Incidence Matrix

Terms	d_1	d_2	d_3	. . .	d_n
the	1	1	1	. . .	0
a	1	1	1	. . .	1
Darjeeling	1	1	1	. . .	0
is	1	1	1	. . .	0
of	1	1	1	. . .	0
in	1	0	0	. . .	1
and	1	1	0	. . .	0
Bengal	1	0	1	. . .	0
It	1	0	1	. . .	0
Its	0	1	0	. . .	1
state	1	0	1	. . .	0
West	1	0	1	. . .	1

Boolean queries: Exact match

- The Boolean retrieval model is being able to ask a query that is a Boolean expression:
  - Boolean Queries are queries using *AND*, *OR* and *NOT* to join query terms
    - Views each document as a set of words
    - Is precise: document matches condition or not
  - Perhaps the simplest model to build an IR system on
- Primary commercial retrieval tool for 3 decades
- Many search systems you still use are Boolean:
  - Email, library catalog, Mac OS X Spotlight

Example: WestLaw http://www.westlaw.com/

- Largest commercial (paying subscribers) legal search service
- started in 1975; ranking added in 1992; new federated search added 2010)
- Tens of terabytes of data; ~700,000 users
- Majority of users *still* use **boolean queries**

- Example query:
  - What is the statute of limitations in cases involving the federal tort claims act?
  - LIMIT! /3 STATUTE ACTION /S FEDERAL /2 TORT /3 CLAIM
    - /3 = within 3 words, /S = in same sentence

Example: WestLaw http://www.westlaw.com/

- Another example query:
  - Requirements for disabled people to be able to access a workplace
- Note that SPACE is disjunction, not conjunction!
- Long, precise queries; proximity operators; incrementally developed; not like web search
- Many professional searchers still like Boolean search
  - You know exactly what you are getting

Boolean queries: More general merges

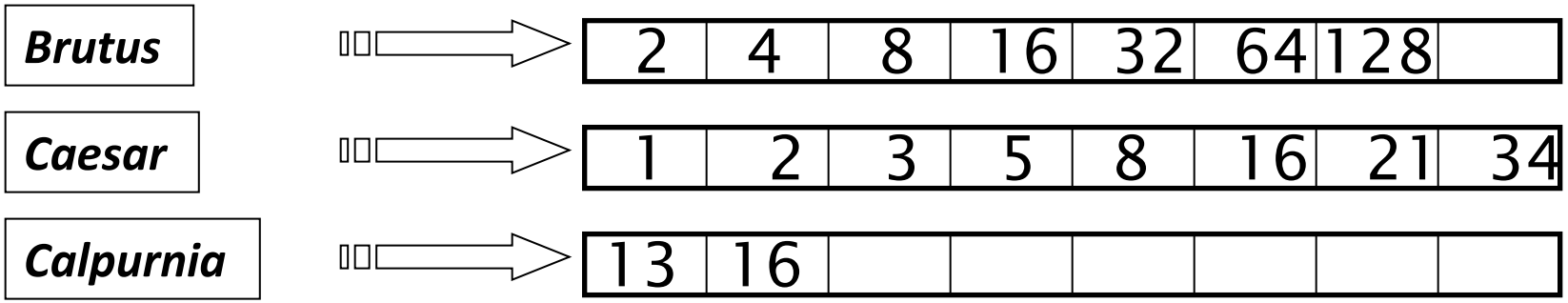
- Exercise: Adapt the merge for the queries:
  - Brutus AND NOT Caesar
  - Brutus OR NOT Caesar
- Can we still run through the merge in time $\Theta(x+y)$?
  - Linear time?
  - What can we achieve?

Merging

- What about an arbitrary Boolean formula?
- (Brutus OR Caesar) AND NOT
- (Antony OR Cleopatra)
- Can we always merge in "linear" time?
  - Linear in what?
- Can we do better?

Query optimization

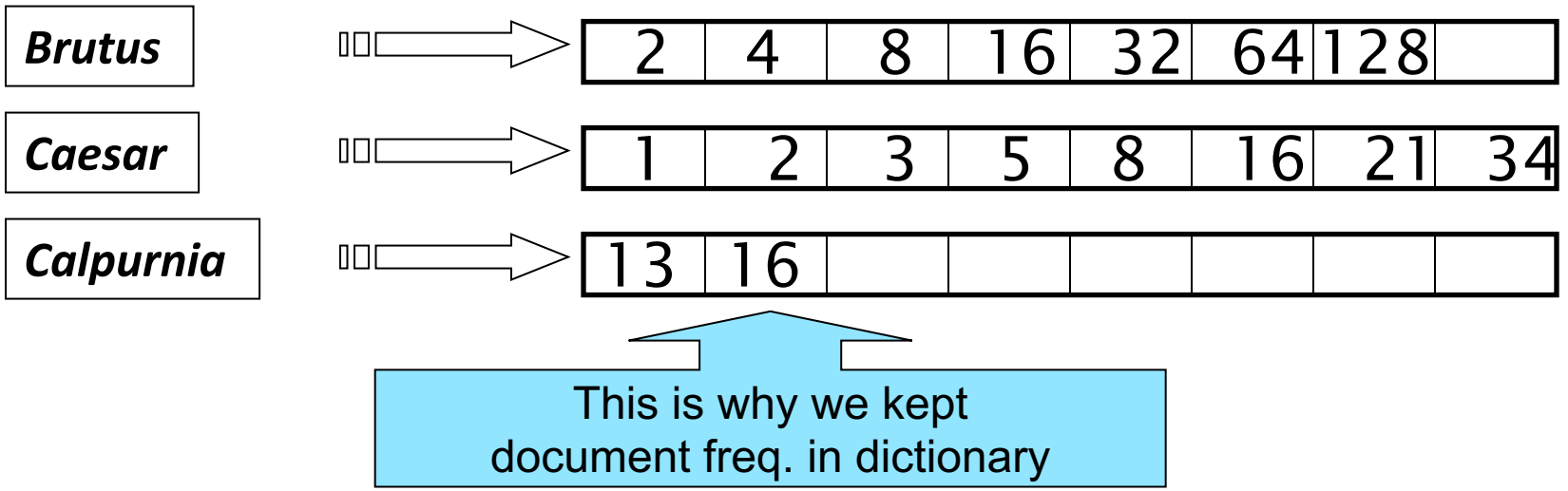
- What is the best order for query processing?
- Consider a query that is an AND of n terms.
- For each of the n terms, get its postings, then AND them together.



Query: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

Query optimization example

- Process in order of increasing frequencies:
  - *start with smallest set, then keep cutting further*



Execute the query as (***Calpurnia*** *AND* ***Brutus)*** *AND* ***Caesar***

More general optimization

- e.g., *(**madding** OR **crowd**) AND (**ignoble** OR **strife**)*
- Get doc. freq.'s for all terms
- Estimate the size of each *OR* by the sum of its doc. freq.'s (conservative)
- Process in increasing order of *OR* sizes

Exercise

- Recommend a query processing order for

(tangerine** OR **trees)** AND **(marmalade** OR **skies)** AND **(kaleidoscope** OR **eyes)

Term	Freq
eyes	213312
kaleidosco	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

- Which two terms should we process first?

Query Processing - Exercises

- Exercise: If the query is ***friends*** *AND* ***romans*** *AND (NOT* ***countrymen****)*, how could we use the freq of ***countrymen***?
- Exercise: Extend the merge to an arbitrary Boolean query. Can we always guarantee execution in time linear in the total postings size?
- Hint: Begin with the case of a Boolean *formula* query: in this, each query term appears only once in the query

Exercise

- Try the search feature at http://www.rhymezone.com/shakespeare/
- Write down five search features you think it could do better

Summary

In this class, we focused on:

(a) Boolean Index Creation
(b) Boolean Operators
(c) Boolean Queries: AND, OR and NOT
(d) Boolean Term – Document Matrix
(e) Boolean IR
    i. Document Retrieval
    ii. Evaluation of Boolean Retrieval
(f) Boolean Query Processing
(g) Query Optimization

Questions
It's Your Time

Contact Information:

Dr. Rajendra Prasath
IIIT Sri City, Chittoor