Extending
classic information retrieval
for today’s possibilities

1

Ranking

- What intuitive criteria?

2

Enhanced document model

- First model: set of terms
  - term in/not in document
- Next: bag of terms
  - know frequency of terms in document
- Now: sequence of terms + additional properties of terms
  - sequence gives you where term in doc
    - derive relative position of multiple query terms
  - Special use? (e.g. in title, font, …)
    - most require “mark-up”: tags, meta-data, etc.

3

HTML mark-up example

<h2> <font color="#A52A2A"> Communication </font></h2>
This course will be essentially ``paperless''. All assignments will be posted <i>only</i> on the course Web site. ``Handouts'' and copies of any transparencies used in class will be posted on the course Web site as well. Important announcements on all aspects of the course will be made on the <a href="announce.html"> Announcements</a> page. <b>Students are responsible for monitoring the postings under ``Announcements''. </b> Schedule changes will be made on the on-line <a href="schedule.html"> schedule page</a>. and announced under ``Announcements''. The only paper we will exchange is your solutions to the problem sets, which we will grade and hand back, the exam questions and your responses, and your project reports.

4

yields

Communication

This course will be essentially ``paperless''. All assignments will be posted *only* on the course Web site (see Schedule and Readings). ``Handouts'' and copies of any transparencies used in class will be posted on the course Web site as well. Important announcements on all aspects of the course will be made on the Announcements page. **Students are responsible for monitoring the postings under ``Announcements''.** Schedule changes will be made on the on-line schedule page. and announced under ``Announcements''. The only paper we will exchange is your solutions to the problem sets, which we will grade and hand back, the exam questions and your responses, and your project reports.

5

Enhanced document model: restate

“sequence of terms + properties of terms”

⇓ WHY?

“set of (term, properties) pairs”

Properties:

- for each distinct term
  - Frequency of term in doc
    - Vector model of classic IR
- for individual occurrence of each term
  - Where in doc.
  - properties of use

6

Model

- Document: set of (term,properties) pairs
- Query: sequence of terms
  - Can make more complicated
- Satisfying: AND model
  - relax if no document contains all?
- Ranking: wide open function
  - info beyond documents and query ?

Data Structure for Collection

- for each document, keep list of:
  - terms appearing
    - aggregate properties of term
      e.g. frequency
    - positions at which each term occurs
      - attributes for each occurrence of term
- keep summary information for documents

Data Structure for Collection: Invert

- for each term, keep list of:
  - documents in which it appears
    - positions at which it occurs in each doc.
      - attributes for each occurrence
- keep summary information for documents
- keep summary information for terms

Inverted Index for Collection

- for each term, keep POSTINGS LIST of:
  - each document in which it appears
    - each position at which it occurs in doc.
      - attributes for each occurrence

  POSTING
- Core structure used by query evaluation and document ranking algorithms

Index structure

```
term_1:(doc ID (position, attributes)
               (position, attributes),
                 …
               (position, attributes)   )
       (doc ID  (position, attributes)
               (position, attributes),
                 …
               (position, attributes)   )
     …
term_2:(doc ID (position, attributes)
               (position, attributes),
                 …
               (position, attributes)   )
     …
```

Models have seen

Model	Document	Query	Satisfy
Boolean	set of terms	Boolean expression over terms	evaluate boolean expression
Vector dictionary of *t* terms	*t*-dimensional vector	*t*-dimensional vector	vector measure of similarity Doc.s ranked by score
Extended	set of pairs (term, properties)	sequence of terms	Boolean AND Doc.s ranked; flexible scoring algorithm