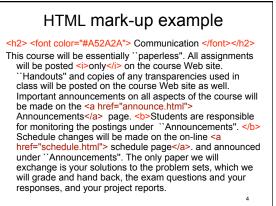


## Ranking

· What intuitive criteria?

## Enhanced document model

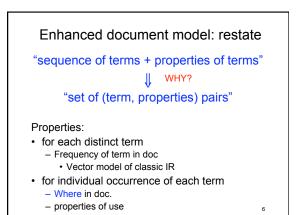
- · First model: set of terms - term in/not in document
- · Next: bag of terms
  - know frequency of terms in document
- Now: sequence of terms + additional properties of terms
  - sequence gives you where term in doc • derive relative position of multiple query terms
  - Special use? (e.g. in title,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



### vields

#### Communication

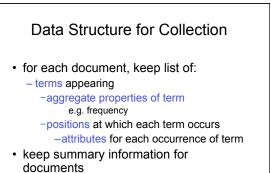
This course will be essentially ``paperless". All assignments will be posted only on the course Web site (see Schedule and Readings). "Handouts" and copies of any transparencies used in class will be posted on the course Web site as well. Important announcements on all aspects of the course will be made on the <u>Announcements</u> page. Students are responsible for monitoring the postings under "Announcements". Schedule changes will be made on the on-line <u>schedule page</u>. 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.



## 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

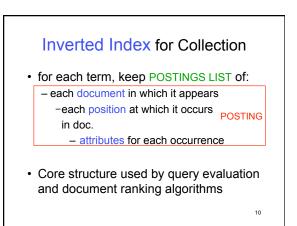
   info beyond documents and query ?

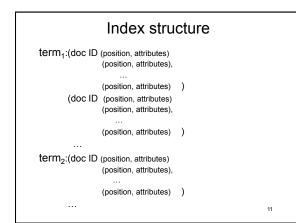


# Data Structure for Collection: Invert

- 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





Model	Document	Query	Satisfy
Boolean	set of terms	Boolean expression over terms	evaluate boolean expression
Vector dictionary of <i>t</i> terms	<i>t</i> -dimensional vector	<i>t</i> -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