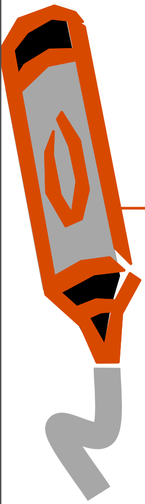
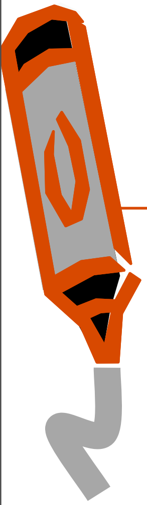
boolean retrieval

some slides courtesy James Allan@umass



what is a retrieval model ?

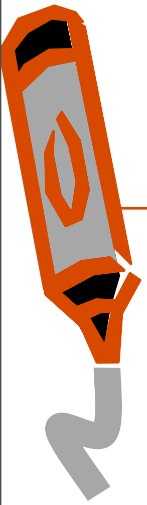
- Model is an idealization or abstraction of an actual process
- Mathematical models are used to study the properties of the process, draw conclusions, make predictions
- Conclusions derived from a model depend on whether the model is a good approximation of the actual situation
- Statistical models represent repetitive processes, make predictions about frequencies of interesting events
- Retrieval models can describe the computational process
  - e.g. how documents are ranked
  - Note that how documents or indexes are *stored* is implementation
- Retrieval models can attempt to describe the human process
  - e.g. the information need, interaction
  - Few do so meaningfully
- Retrieval models have an explicit or implicit definition of relevance



retrieval models

today

-boolean
-vector space
-latent semantic indexing
-statistical language
-inference network



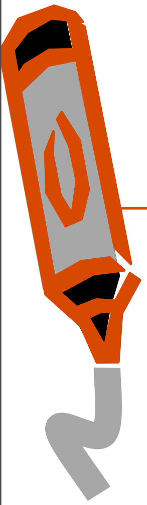
exact vs. best match

- Exact-match
  - query specifies precise retrieval criteria
  - every document either matches or fails to match query
  - result is a set of documents
    - Unordered in pure exact match

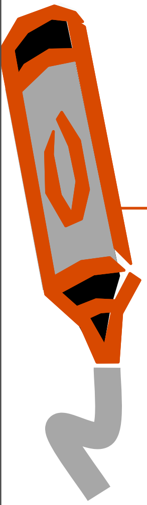
- Best-match
  - Query describes good or "best" matching document
  - Every document matches query to some degree
  - Result is *ranked list* of documents

- Popular approaches often provide some of each
  - E.g., some type of ranking of result set (best of both worlds)
  - E.g., best-match query language that incorporates exact-match operators

IR vs Databases



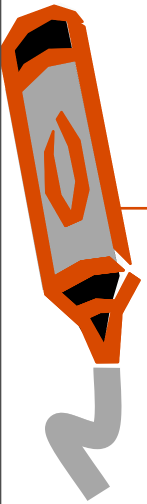
	Databases	IR
Data	**Structured**	**Unstructured**
Fields	**Clear semantics** (SSN, age)	**No fields** (other than text)
Queries	**Defined** (relational algebra, SQL)	**Free text** ("natural language"), Boolean
Recoverability	**Critical** (concurrency control, recovery, atomic operations)	**Downplayed**, though still an issue
Matching	**Exact** (results are *always* "correct")	**Imprecise** (need to measure effectiveness)



exact match retrieval

- Advantages of exact match
  - Can be very efficiently implemented
  - Predictable, easy to explain
  - Structured queries for pinpointing precise documents
  - Work well when you know exactly (or roughly) what the collection contains and what you're looking for

- Disadvantages of exact match
  - Query formulation difficult for most users
  - Difficulty increases with collection size
  - Indexing vocabulary same as query vocabulary
  - Acceptable precision generally means unacceptable recall
  - Ranking models consistently shown to be better
  - • Hard to compare best- and exact-match in principled way



best match retrieval

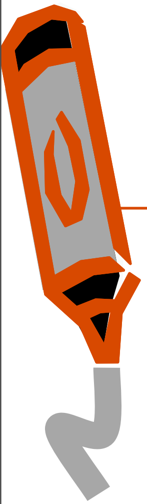
- Retrieving documents that satisfy a Boolean expression constitutes the Boolean exact match retrieval model

- Best-match or ranking models are now more common

- Advantages:
  - Significantly more effective than exact match
  - Uncertainty is a better model than certainty
  - Easier to use (supports full text queries)
  - Similar efficiency (based on inverted file implementations)

- Disadvantages:
  - More difficult to convey an appropriate cognitive model (“control”)
  - Full text does not mean natural language understanding (no “magic”)
  - Efficiency is always less than exact match (cannot reject documents early)

- Boolean or structured queries can be part of a best-match retrieval model

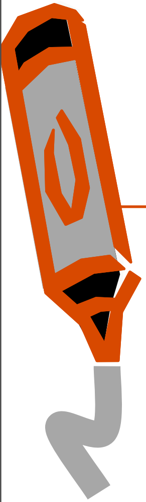


boolean retrieval

- Boolean model is most common exact-match model
  - queries are logic expressions with document features as operands
  - In pure Boolean model, retrieved documents are not ranked
- Most implementations provide some sort of ranking
  - query formulation difficult for novice users
- Boolean queries
  - Used by Boolean model
  - and in other models (Boolean query ≠ Boolean model)
- “Pure” Boolean operators: AND, OR, AND-NOT
- Most systems have proximity operators
- Most systems support simple regular expressions as search terms to match spelling variants

boolean query languages

- Many users prefer Boolean
  - Especially professional searchers
  - Many WESTLAW, DIALOG searches still use Boolean
  - “Control”
  - Understandability
- For some queries or collections, Boolean often works better (e.g., using AND on the Web)
- Boolean and free text find different documents
- Need retrieval models that support both
  - “Extended Boolean” vector space
  - Probabilistic inference network
- Need interfaces that provide good cognitive models for ranking

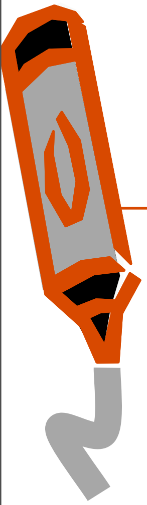


example

	nuclear	nonproliferation	treaty	Iran
D1	0	0	0	0
D2	1	0	0	1
D3	0	0	1	0
D4	0	0	1	1
D5	1	1	0	0
D6	0	0	1	1
D7	1	0	1	0
D8	0	1	1	1

query :

(nuclear **AND** treaty) **OR** ((**NOT** treaty) **AND** (nonproliferation **OR** Iran))



example

	nuclear	nonproliferation	treaty	Iran
D1	0	0	0	0
D2	1	0	0	1
D3	0	0	1	0
D4	0	0	1	1
D5	1	1	0	0
D6	0	0	1	1
D7	1	0	1	0
D8	0	1	1	1

query :
(nuclear **AND** treaty) **OR** ((**NOT** treaty) **AND** (nonproliferation **OR** Iran))

retrieved docs : D7,D5,D2