Introduction to IR and IE

Vasudeva Varma

IIIT Hyderabad

Approaches to IR

- Two types of retrieval
  - By metadata (subject headings, keywords, etc.)
  - By content
- Metadata as manually assigned information
  - Human agreement is not good
  - Expensive for most data
- Metadata assigned automatically
  - Quality is reasonable, but not high for many applications
- Metadata in general
  - Requires *a priori* prediction of headings, keywords, …
- Most successful IR approaches are content-based

basic approach to IR

- Successful content-based approaches are statistical
  - Rather than actual "understanding" of text
  - Text understanding effectiveness is very poor
    - Exception: works better in some restricted domains
- IR statistics used in different ways
  - Past/concurrent queries and relevance judgments
    - Collaborative systems
  - Document and query similarities

relevant items are similar

- Much of IR depends upon idea that
  similar → relevant to same queries

- Usually measure query-document similarity
  - Can consider document-document similarity

- "Similar" can be measured in many ways
  - String matching
  - Same vocabulary
  - Probability arise from same model
  - Same meaning
  - …

“bag of words”

- An effective and popular approach
- Compares words without regard to order

- Consider reordering words in a headline
  - Stocks fall on inflation fears
  - inflation stocks fall on fears
  - fall inflation stocks on fears
  - fall fears inflation stocks on
  - fall fears inflation on stocks

- Q: How far can we push “bag of words”?

IR engines: State of the Art

- Wide variation in retrieval results
  - User topic
  - Retrieval system
- Different approaches work for different systems.
- No way to determine which approach will work for a particular query.

Solution:

- **Deeper analysis of the content and Query**

Motivation for Deeper Analysis

- Texts are one of the major sources of information and knowledge.

However, they are not transparent.

They have to be systematically integrated with the other sources like data bases, numerical data, etc.

NLP/IR/IE for better analysis

IA for better presentation

IR vs. IE vs. IA

- To search and retrieve documents in response to queries for information

Vs.

- To extract information that fits pre-defined database schemas or templates, specifying the output formats

Vs.

- To make the required information accessible to the user in their choice of language, mode, level of detail and format