Web Searching: The Challenge

• Query:

“Decline in rainfall and impact on farms near the Sahara”

Web Pages include:

• Web Page 1:

Rainfall measurements in the Sahara continue to show a steady decline starting from the first measurements in 1961. In 1996 only 12mm of rain were recorded in upper Sudan, and 1mm in Southern Algiers…”
Web Pages include:

• Web Page 2:

Dan Marino states that professional football risks loosing the number one position in heart of fans across this land. Declines in TV audience ratings are cited...

Web Pages include:

• Web Page 3:

Alarming reductions in precipitation in desert regions are blamed for desert encroachment of previously fertile farmland in Northern Africa. Scientists measured both yearly precipitation and groundwater levels.
Challenges:

• How to retrieve 3 instead of 1 and not 2?
• How to rank 3 as best?

• How do we evaluate a search engine?
• Later in Webware: how do we build a “good” search engine?

Evaluation of Search Engines (Web and Text collections)

• Two main measures:
  – Recall
    • proportion of relevant material actually retrieved
  – Precision
    • proportion of retrieved material actually relevant

• Also:
  – Form of Presentation
  – Effort Required/Ease of Use (e.g., Boolean queries)
  – Time and Space Efficiency
Relevance

• In what ways can a document be relevant to a query?
  – Answer precise question precisely.
    • What is the distance from the Sun to Jupiter?
      – 750,000 kms (470,000 miles)
  – Partially answer question
    • Which theaters are showing “The Hurricane”? 
      – Solomon Pond, White City
  – Suggest a source for more information.
  – Give background information.
  – Remind the user of other knowledge.

These issues have been studied in Information Retrieval

• Precision

\[
\frac{\text{# relevant retrieved}}{\text{# retrieved}}
\]

• Recall

\[
\frac{\text{# relevant retrieved}}{\text{# relevant in collection}}
\]