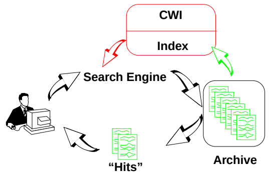
Dissemination of Collection Wide Information in a Distributed Information Retrieval System

Charles L. Viles and James C. French

Presented at the 18th Conference on Research and Development in Information Retrieval
Seattle, WA, July 9-13, 1995

Problem Context



Collection Wide Information (CWI) derived from the document corpus is used to enhance the effectiveness of user queries.

What is Collection Wide Information?

Collection Wide Information (CWI) is statistics and data structures built from the entire document collection.

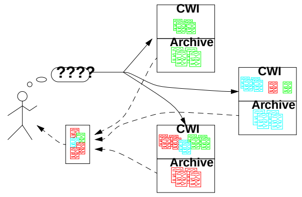
A Sample:

$$weight_{\text{Doc i, term k}} = freq_{ik} \times idf_k$$

where *inverse document frequency (idf)* is

$$idf_k = \log\left(\frac{\text{Collection Size}}{\text{Docs containing term k}}\right)$$

Incomplete CWI: Distributed Scenario



In a distributed system, each site's "view" of CWI may differ.

Approach

Basic Tenets:

- Distributed search, merge results.
- Each site's "view" of CWI may differ from the true CWI.
- Archives can communicate with each other.
- Communication level should be just enough to maintain retrieval effectiveness.
- Distribution of "content" may affect required communication level.

Parameterizing Communication

Dissemination Level (*d*)

- A site builds its view of CWI from:
  - its own documents and
  - a fraction, *d*, of the documents at other sites.

Interpretation

- *d = 0*, use only local information
- *d = 1*, use all information from all sites

Parameterizing Content Allocation

Content Allocation (*a*)

- The distribution of content in the system may affect retrieval when sites have imperfect knowledge.
- Model:
  - Determine content-similar documents
  - Assign content-similar documents to the same site with an *affinity* probability *a*.
  - Assign to a random site with probability *1-a*.

Interpretation

- *a = 0*, content-uniform system
- *a = 1*, content-skewed system

Methodology

Data:

- Four document collections (two large, two small)
- MED and CACM (1000-3000 documents)
- AP88 and WSJ (80,000 documents each)

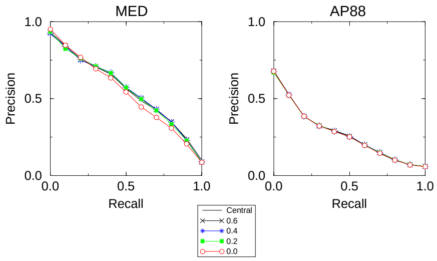
Parameters:

- Number of sites = 20
- For each collection, vary two parameters, d and a
- "Configuration" is (d, a, collection)

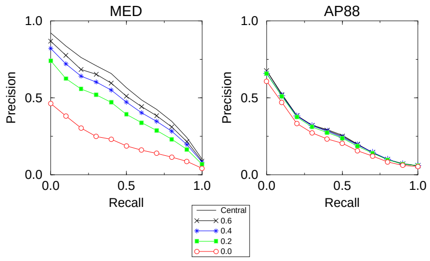
Evaluation:

- Multiple runs at each configuration
- Use standard IR evaluation measures
- Compare against an "omniscient" Central archive

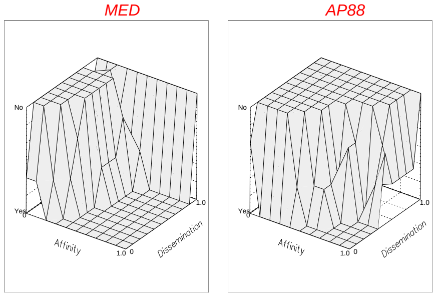
Content Uniform Results (a = 0, d = *)



Content Skewed Results (a = 1, d = *)



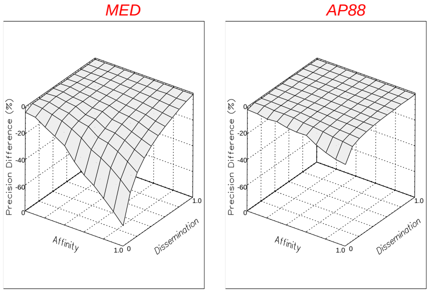
Is There a Statistical Difference?



MED

AP88

Is There a Practical Difference?



MED

AP88

Implications and Future Work

Implications

- Degree of communication is tied to content allocation
- Content-skewed systems must have inter-site communication for best search quality.
- But, communication can be "lazy" or delayed.

Future Work

- Are operational distributed IR systems content-skewed?
- How does CWI drift over time?