Are Popular Documents More Likely To Be Relevant? A Dive into the ACLIA IR4QA Pools

Tetsuya Sakai
and
Noriko Kando

EVIA 2008, December 16, 2008@NII, Tokyo
What is ACLIA IR4QA?

- ACLIA=Advanced Cross-lingual Information Access Task Cluster
- IR4QA=Information Retrieval for Question Answering Task

The IR4QA test collections:
- About 100 topics (CS, CT, JA and English)
- 545,162 CS (Simplified Chinese) docs
- 1,150,954 CT (Traditional Chinese) docs
- 419,759 JA (Japanese) docs
- Graded relevance assessments collected through pooling

See IR4QA Overview paper for more details
Pooling for relevance assessments

Target Documents
- CS: Simplified Chinese
- CT: Traditional Chinese
- JA: Japanese

Run depth = 1000

System 1
- Run 1
  - Pool depth >= 30

System N
- Run N
  - Pool depth >= 30

Topic A

Relevance assessments
- L2-relevant
- L1-relevant
- L0: judged nonrelevant

L2: relevant
L1: partially relevant
L0: judged nonrelevant
Different pool depths for different topics

- Assess depth-30 pool
  - Assess depth-50 pool (minus depth-30 pool)
    - Assess depth-70 pool (minus depth-50 pool)
      - Assess depth-90 pool (minus depth-70 pool)
        - Assess depth-100 pool (minus depth-90 pool)

Relevance assessments coordinated independently by Donghong Ji (CS), Chuan-Jie Lin (CT) and Noriko Kando (JA)

See IR4QA Overview Tables 29-31 for details
Sorting the pooled documents for assessors

- Traditional approach: Docs sorted by IDs
- IR4QA approach: Sort docs in depth-X pool by:
  - #runs containing the doc at or above rank X (primary sort key)
  - Sum of ranks of the doc within these runs (secondary sort key)

Present ``popular'' documents first!

X=30 in this study
Assumptions behind the sort

1. Popular docs are more likely to be relevant than others.
2. If relevant docs are concentrated near the top of the list to be assessed, this is easier for the assessors to judge more efficiently and consistently.

Objective of this very short talk: Show that Assumption 1 is valid for the IR4QA test collections!
L0 (Judged nonrelevant)

L1 (partially relevant)

L2 (relevant)

L1+L2

L0 increases (and eventually decreases due to different pool sizes across topics)

L1 does not necessarily follow this pattern

L1+L2 is top-heavy and decreases almost monotonically;

Similar pattern for L2

Counts summed across topics
L0 increases (and eventually decreases due to different pool sizes across topics)

L1+L2 is top-heavy and decreases almost monotonically;
Similar pattern for L2

L1 does not necessarily follow this pattern
L0 increases (and eventually decreases due to different pool sizes across topics)

L1+L2 is top-heavy and decreases almost monotonically;
Similar pattern for L2

L1 does not necessarily follow this pattern
Conclusions

Assumption 1: “Popular docs are more likely to be relevant than others” is correct at least for the IR4QA collections!

Moreover, we observed that “Popular docs are more likely to be highly relevant than others.”

So our sorting strategy may be reasonable.

More on ACLIA IR4QA in the afternoon of NTCIR-7 Day 3 (18th)!