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

Tetsuya Sakai† Noriko kando*
†NewsWatch, Inc., tetsuyasakai@acm.org
*National Institute of Informatics, Noriko.Kando@nii.ac.jp

Abstract

The ACLIA IR4QA Task at NTCIR-7 is an ad hoc document retrieval task involving three document languages. Although IR4QA used pooling for collecting relevance assessments, it was unique in that the pooled documents were sorted before presenting them to the assessors, based on the assumption that "popular" documents are more likely to be relevant than others. We show that this assumption is indeed valid for the IR4QA test collections.
Keywords: *test collection, pooling, relevance assessment.*

1 Introduction

The ACLIA (Advanced Cross-lingual Information Access) IR4QA (Information Retrieval for Question Answering) Task at NTCIR-7 is an "ad hoc" document retrieval task involving three document languages: Simplifield Chinese (CS), Traditional Chinese (CT) and Japanese (JA). Following previous work on building large-scale test collections, the IR4QA relevance assessment data were collected through *pooling*. However, IR4QA was unique in that the documents in the depth-X pools were *sorted*, first by the number of runs containing the document at or above rank X (the larger the better), and then by the sum of ranks of that document within those runs (the smaller the better) [1]. Thus, documents retrieved by many teams, especially those retrieved early in the ranked lists, were presented to the relevance assessors with the highest priority.

The assumptions behind the IR4QA relevance assessment strategy were:

Assumption 1 "Popular" documents (i.e., those retrieved at high ranks by many systems) are more likely to be relevant than others;

Assumption 2 If there are more relevant documents near the top of the list of documents to be judged than near the bottom, then this makes it easier for the assessors to make judgments more efficiently and consistently than when relevant documents are randomly spread across the list.

This paper shows that **Assumption 1** is indeed valid for the IR4QA test collections.

2 A Dive into the Pools

We used the depth-30 pools of the CS, CT an JA test collections of IR4QA since, at the time of this writing, the relevance assessments have not yet been done beyond this depth for some topics [1]. The size (i.e., the number of documents) of the depth-30 pools range from 85 to 466 for CS, 133 to 393 for CT, and 127 to 347 for JA. As mentioned earlier, each pool file is a sorted list of document IDs, where the primary sort key is the number of runs containing the document at or above rank 30, and the secondary sort key is the sum of ranks of that document within those runs. The IR4QA relevance levels are: $L2$ (relevant), $L1$ (partially relevant) and $L0$ (judged nonrelevant).

For each topic for each test collection, we created bins of document ranks in the pool, where the first bin corresponds to ranks 1-10, the second bin corresponds to ranks 11-20, and so on. Then, for each bin, we counted the number of $L2$-relevant, $L1$-relevant and $L0$ documents. Finally, the counts were summed across topics. Figure 1 shows the results.

By looking at the blue bars (representing $L2$ and $L1$ documents), it can be observed that our results support **Assumption 1**. That is, popular documents are more likely to be relevant than others. Moreover, by looking at the red bars (representing $L2$ documents only), we can further claim that popular documents are more likely to be *highly* relevant than others. Whereas, the yellow bars (representing $L1$ documents only) do not necessarily follow this pattern: For example, the CS graph shows that there are more $L1$-relevant documents in the "41-50" bin than in the "1-10" bin.

Acknowledgements

We thank the ACLIA IR4QA organisers and the participants for making this research possible.

References

[1] Sakai, T. *et al.*: Overview of the NTCIR-7 ACLIA IR4QA Task, *Proceedings of NTCIR-7*, to appear, 2008.

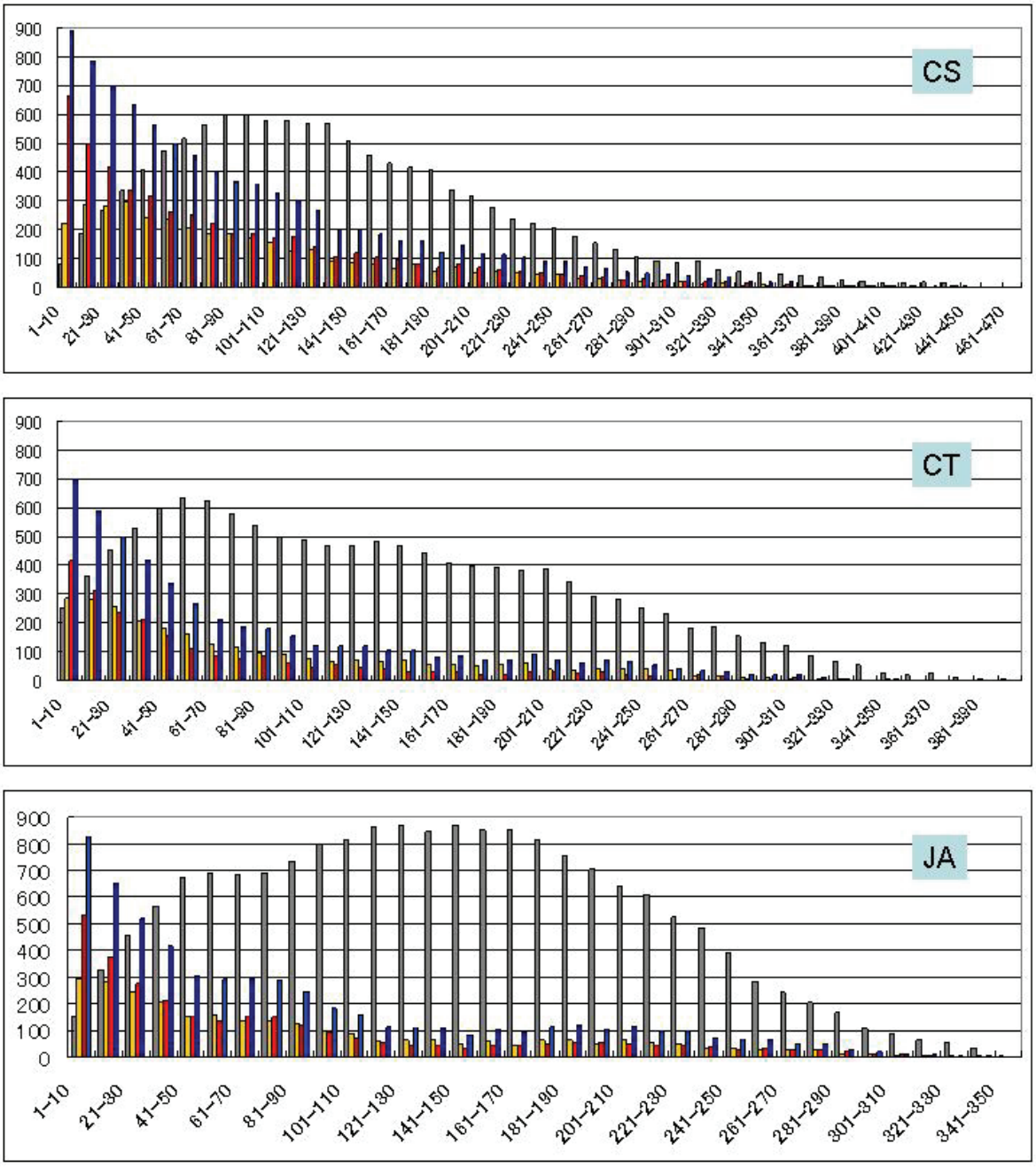

Figure 1. Judged nonrelevant ($L0$) documents (gray), $L1$-relevant documents (yellow), $L2$-relevant documents (red), and sum of $L1$- and $L2$-relevant documents (blue). The horizontal axis represents bins of document ranks in the sorted pool, and the vertical axis represents the document counts summed across topics. For example, "1-10" denotes ranks 1-10 in the sorted pool.