Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > stat > arXiv:1107.2691

Search or Article-id (Help | Advanced search) All papers Go!

Statistics > Computation

On the Weakenesses of Correlation Measures used for Search Engines' Results (Unsupervised Comparison of Search Engine Rankings)

Paolo D'Alberto, Ali Dasdan

(Submitted on 13 Jul 2011)

The correlation of the result lists provided by search engines is fundamental and it has deep and multidisciplinary ramifications. Here, we present automatic and unsupervised methods to assess whether or not search engines provide results that are comparable or correlated. We have two main contributions: First, we provide evidence that for more than 80% of the input queries - independently of their frequency - the two major search engines share only three or fewer URLs in their search results, leading to an increasing divergence. In this scenario (divergence), we show that even the most robust measures based on comparing lists is useless to apply; that is, the small contribution by too few common items will infer no confidence. Second, to overcome this problem, we propose the fist content-based measures - i.e., direct comparison of the contents from search results; these measures are based on the Jaccard ratio and distribution similarity measures (CDF measures). We show that they are orthogonal to each other (i.e., Jaccard and distribution) and extend the discriminative power w.r.t. list based measures. Our approach stems from the real need of comparing search-engine results, it is automatic from the query selection to the final evaluation and it apply to any geographical markets, thus designed to scale and to use as first filtering of query selection (necessary) for supervised methods.

Comments: 16 pages, 19 figures
Subjects: **Computation (stat.CO)**; Information Retrieval (cs.IR)
Cite as: **arXiv:1107.2691 [stat.CO]**
(or **arXiv:1107.2691v1 [stat.CO]** for this version)

Submission history

From: Paolo D'Alberto [view email]
[v1] Wed, 13 Jul 2011 22:35:07 GMT (1541kb,D)

Download:

- PDF
- Other formats

Current browse context:
stat.CO
< prev | next >
new | recent | 1107

Change to browse by:
cs
cs.IR
stat

References & Citations
- NASA ADS

Bookmark(what is this?)

Which authors of this paper are endorsers?

Link back to: arXiv, form interface, contact.