Traps and Pitfalls of Order-Based Correlation Indices for Topic-Biased PageRank

Paolo Boldi, Roberto Posenato, Massimo Santini and Sebastiano Vigna

2006
Traps and Pitfalls of Order-Based Correlation Indices for Topic-Biased PageRank

Paolo Boldi\textsuperscript{1}, Roberto Posenato\textsuperscript{2}, Massimo Santini\textsuperscript{1}, and Sebastiano Vigna\textsuperscript{1}

\textsuperscript{1} Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, via Comelico 39/41, I-20135 Milano, Italy
\textsuperscript{2} Dipartimento di Informatica, Università degli Studi di Verona, strada le grazie 15, I-37134 Verona, Italy

Several studies have been recently published about the correlation between rank vectors for the same web graph obtained with different ranking techniques, or computed with bias towards different topics. The correlation is usually defined using an order-based measure, such as Kendall’s $\tau$ or Spearman’s $\rho$, computed on the given ranks. In this work we focus on Kendall’s $\tau_b$ – a variant of Kendall’s $\tau$ that takes ties into account.

In an attempt to duplicate the correlation data published by Haveliwala \cite{Haveliwala} about rankings biased towards different topics (where the correlation was computed using a measure similar to Kendall’s $\tau$), we met significant difficulties due to the number of different ways in which PageRank can be defined and computed, and to the lack of public data over which to replicate the experiments.

The first contribution of our work is a publicly available snapshot of the .uk domain, together with topic bias data derived from the ODP hierarchy. We believe such a public, well-defined data set is essential to continue research on topic-biased ranking.

Biasing w.r.t. a topic, however, requires additional care. Real-world snapshots contain a significant percentage of dangling nodes (nodes without outlinks). In the well-known surfer metaphor for PageRank, the way in which the surfer chooses the next node when she is at a dangling node is an issue resolved in different ways by different authors. We distinguish clearly between strongly preferential PageRank, in which the distribution of random jumps and of jumps out of dangling nodes is identical, and corresponds to a topic or personalisation bias, and weakly preferential PageRank, in which the random jump and the dangling node jump distributions are not identical, and, in principle, uncorrelated. We extend the closed formula given by Del Corso, Gulli and Romani \cite{DelCorso} for strongly preferential PageRank to a general formula that applies also to weakly preferential PageRank. Using this formula, any biased, weakly preferential PageRank vector whose distributions are a linear combination of a set of base vectors can be computed using the pseudorank vectors associated to the base vectors. The computation of a pseudorank vector requires the same amount of computational effort as computing PageRank, but once pseudoranks have been computed it is immediate to compute and compare several different biased ranks.

More precisely, denoting with $P$ the natural random-walk matrix of a web graph (i.e., $p_{ij}$ is one over the outdegree of $i$ if there is an arc from $i$ to $j$, zero

\textsuperscript{*} This work is partially supported by the EC Project DELIS.
otherwise), with \( v \) the distribution of random jumps, with \( u \) the distribution of jumps out of dangling nodes, with \( \alpha \) the damping factor (the probability of following a link) and with \( d \) the characteristic vector of dangling nodes, a simple computation shows that PageRank is defined by the solution \( z \) of the linear system

\[
(I - \alpha P^T - \alpha u d^T) z = (1 - \alpha)v.
\]

When \( u = v \), PageRank is strongly preferential. Now, let \( u^* = (I - \alpha P^T)^{-1} u \) and \( v^* = (I - \alpha P^T)^{-1} v \), the pseudoranks associated to \( u \) and \( v \). Using the Sherman–Morrison formula, we show that

\[
z = (1 - \alpha) v^* + (1 - \alpha) \frac{u^* d^T v^*}{1 + d^T u^*}.
\]

Note that by properly ordering multiplications, no matrix computation is necessary to compute the formula above. When \( u = v \), the formula reduces to the one provided in [2]: the pseudorank is just a multiple of the actual rank. The condition number in the computation of pseudoranks is the same as or better than that for PageRank, and there is no increase of computation time.

The last issue we tackle, however, is probably the most interesting one. Correlation measures such as Kendall’s \( \tau \) are based on the number of exchanges appearing in the rank list (e.g., positions \( i \) and \( j \) such that \( i < j \) but the rank of \( i \) is larger than the rank of \( j \)) and on the number of ties, and so on. The point that appears to have been completely missed in the literature (including that previously contributed by the authors [3]) is that the computation of ranks is almost always the result of interrupting a limiting process (e.g., the power method). The interruption is usually based on a threshold satisfied by the \( \ell_1 \) or \( \ell_2 \) measure.

As a result, a number of correct digits appearing in the ranks is unpredictable, as it just depends on the computational process. The abovementioned norms guarantee on average a certain number of significant digits, but unless the much more demanding \( \ell_{\infty} \) measure is used, almost no guarantee can be provided for single ranks.

In the case several very close values appear in the ranking list, the effect of such an unpredictable precision turns out to be catastrophic, in particular with certain computational methods (such as Gauss-Seidel). Namely, the value returned by Kendall’s \( \tau_0 \) is a function of the significant digits considered in its computation.

To prove the impact of this observation experimentally, we present data obtained by working out the strongly preferential PageRank computation in a standard fashion, using the Gauss-Seidel method, for a certain bias vector. We stopped the computation at different stages, having every time a known (lower bound on the) number of correct digits in the computed ranks, and then we computed Kendall’s \( \tau_0 \) using only a limited number of digits in the ranks. To limit the number of significant digits we used, we turned each floating point-number into its bitwise IEEE 754 representation, and manipulated it directly so to delete all digits beyond a certain threshold. This procedure, applied with threshold \( k \),
has the effect of batching all values in the interval \([j2^{-k} . (j + 1)2^{-k}]\), into the value \(j2^{-k}\). The net effect is that several ranks that appeared to be exchanged because of unpredictable noise in the last digits are now considered as equal. (We remark that due to the size of the data we use, these computations require thousands of hours of CPU time).

The resulting graphs (an example is presented in Figure 1) are quite surprising: even the \(\tau_0\) of a certain rank vector computed against the same vector, but with a different precision can go down as low as 0.2. Of course, as far as the computation of \(\tau_0\) uses no more digits than those that are guaranteed to be correct, the correlation is 1, but it rapidly drops as soon as more digits are considered; in particular, computing \(\tau_0\) blindly (i.e., without any form of batching) can bring essentially to random results.

![Graph](image)

**Fig. 1.** Values of Kendall’s \(\tau_0\); when rank values are batched using more bits than the number of significant bits guaranteed in PageRank computation, the value of \(\tau_0\) drops significantly. These data are determined from the .uk web graph, using the “adult” preference vector (from ODP), \(\alpha = .85\) and the Gauss-Seidel method.

More evidence is needed to corroborate the data we present, and we plan to provide it in the full paper, which will be based on the larger, 100M-pages snapshots we are gathering for the EU DELIS project. But already the preliminary data we present show that order-based correlation indices must be managed with great care, and have probably given rise to biased results in the past.

**References**