Using the H-index to Estimate Blog Authority

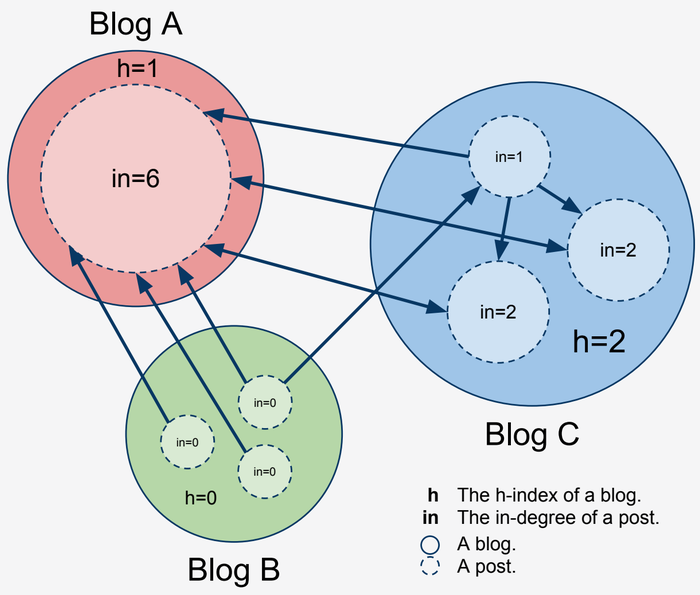
José Devezas[†], Sérgio Nunes[‡§], Cristina Ribeiro[‡§]

[†]Labs SAPO/UP [§]INESC-Porto [‡]DEI, FEUP

Universidade do Porto
Faculdade de Engenharia
FEUP

The H-index and How It Can Be Applied to Blogs

- The h-index was proposed by Hirsch to rank scientists or scholars, based on the productivity and impact of their publications.
- A scientist has index h if h of his or her N papers have at least h citations each and the other $(N - h)$ papers have no more than h citations each.
- A connection between blogs and scientific authorship can easily be established: a blog is comparable to an author or scholar, while its posts are analogous to papers.
- **A blog has index h if h of its N posts have at least h in-links each and the other $(N - h)$ posts have no more than h in-links each.**
- We hypothesise that the h-index might be a strong blog ranking metric that behaves differently than the in-degree — not only it takes into account the number of in-links (quantity), but also the number of blog posts (sustainability).



TREC Blogs08 Collection and Data Preparation

- TREC Blogs08 collection compiles over 1.3 million blogs, with more than 28 millions posts, and a total compressed size of 453 GB.
- We parse each post, extracting the URLs found on the $href$ attribute of the HTML anchors.
- We remove the URLs whose domains don't belong to the blog collection, and create a graph representation of the post network, grouping posts by blog.
- Based on the post graph, we compute the in-degree for each post and then calculate each blog's h-index and in-degree.
- We rank blogs by h-index and by in-degree and compare both metrics values for the top blogs.

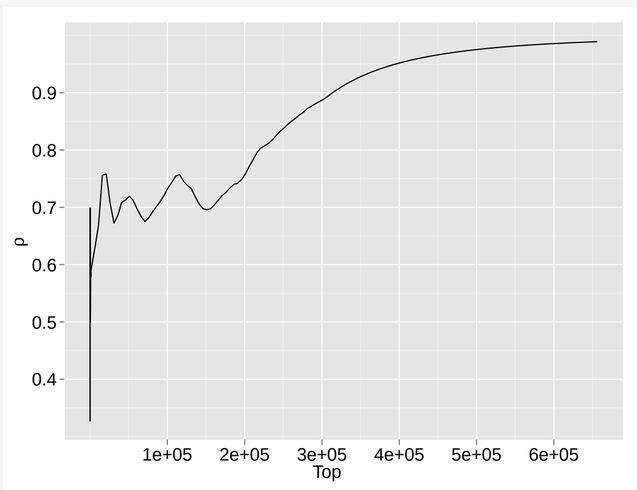
Comparing the H-index with the In-degree as Blog Ranking Metrics

Rank	Domain	H-index	In-degree
1	www.delightfulblogs.com	**700**	2,639,287
2	masalog.com	**603**	604,239
3	taurinerules.blogspot.com	**511**	289,159
4	blogs.nypost.com (TV)	**511**	439,516
5	blogs.nypost.com (Sports)	**473**	335,407

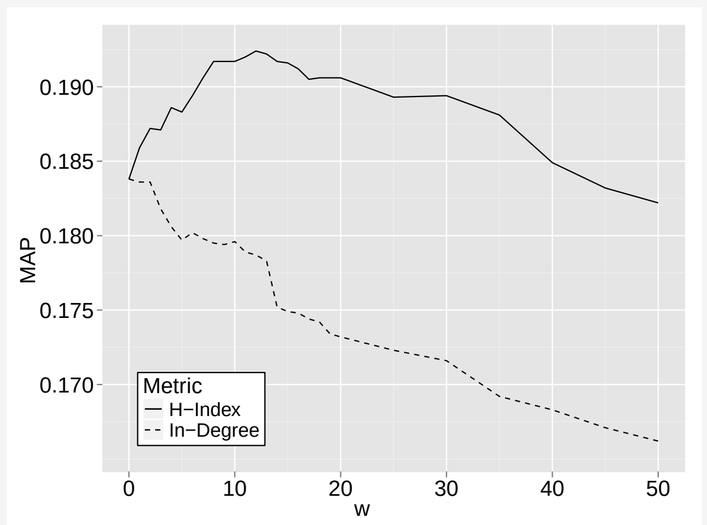
Rank	Domain	In-degree	H-index
1	bodyelectric.blogspot.com	**5,345,032**	2
2	rpc.blogrolling.com	**5,025,547**	301
3	www.delightfulblogs.com	**2,639,287**	700
4	richard-upton.blogspot.com	**1,921,344**	11
5	feeds.feedburner.com	**1,242,343**	24

Top k	ρ	Top k	ρ
25	0.4909091	150	0.6313322
50	0.3272727	175	0.6986296
75	0.4854953	200	0.6438799
100	0.5336898	225	0.6277835
125	0.4990939	250	0.6034763

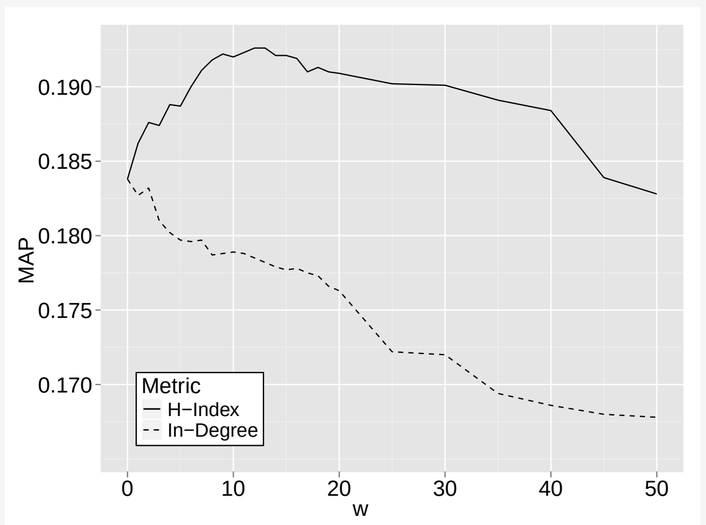
- Even though the in-degree doesn't follow the same behavior as the h-index, blogs with a high h-index also have a high in-degree.
- On the other hand, we can find extremely low h-index values for some of the highest in-degree scores.
- For instance, $bodyelectric.blogspot.com$ is the highest ranked blog according to the in-degree, however it has an h-index of 2, meaning that there are only 2 posts with 2 or more in-links.
- We calculate the Spearman's rank correlation coefficient ρ, for the top k blogs, while increasing k.
- The value of ρ is constantly smaller than 0.76, for cuts below 20,000, even dropping to 0.33 for the top 50 cut. For cuts above 20,000, ρ tends to grow and stabilize around 0.99.
- The 0.76 rank correlation for the highest ranked results led us to further explore the quality of the h-index in the blog distillation task.



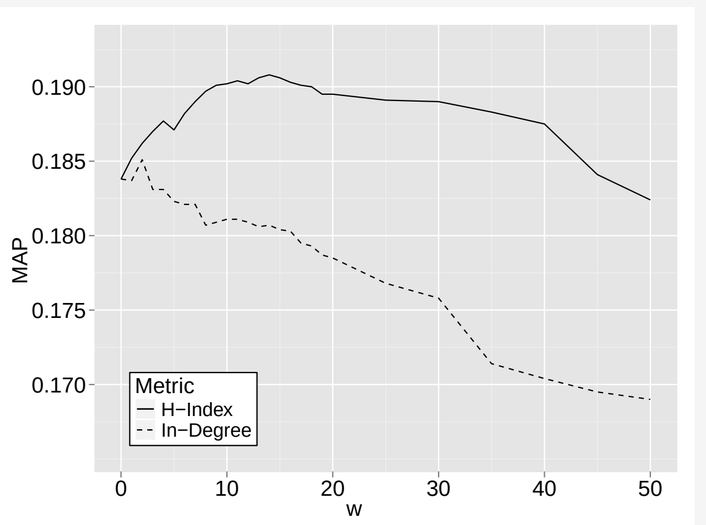
Comparing the H-index with the In-degree as Query-Independent Features in the Blog Distillation Task



(a) Full Blogs08 post graph.



(b) Blogs08 post graph without loops.



(c) Blogs08 post graph without loops and edge multiplicity.

- We begin with the BM25 score as the baseline $(w = 0)$.
- We experiment with two different query-independent features (the h-index and the in-degree) in an attempt to improve blog search over the baseline.
- So, $score(q, b) = BM25(q, b) + w \times log(h\text{-}index(b))$, where q is a query, b a blog and w the weight of the link-based component.
- Using TREC's query relevance assessments, we calculate the mean average precision (MAP) of search results for progressively larger values of w.
- Using the in-degree as a query independent feature actually decreases the quality of the results.
- On the other hand, using the h-index increases the quality of the results.
- We repeat this process in (b) by removing self-citations or loops, and in (c) by removing loops and multiple links between posts.
- Self-citations (or loops) have a slightly negative impact in h-index ranking.
- Edge multiplicity, however, is not an issue and actually results in lower MAP values.