Using the H-index to Estimate Blog Authority

José Devezas[†], Sérgio Nunes[‡§], Cristina Ribeiro[‡§]
[†] Labs SAPO/UP
[§] INESC-Porto
[‡] DEI, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
{jld,ssn,mcr}@fe.up.pt

Abstract

Link analysis is a technique frequently used in the ranking of web sites. On the web, we often encounter content that is organized by entries, sorted from recent to old, and generally follows the structure of a blog. In this paper we explore and evaluate the usage of a bibliometrics measure, called h-index, for the task of blog ranking, in an information retrieval context. We base our experiments on the TREC Blogs08 collection, which comprises over 28 million posts. The results obtained indicate that the h-index is a robust metric that allows for an improved relevance discrimination between blogs, when compared to the in-degree. Additionally, tests performed using distinct versions of the post graph, indicate that this metric might tolerate a certain level of link clutter.

Introduction

Over the past few years, blogs have grown in popularity among the masses, and recently branched into new services focused on short message content sharing. These services, microblogging platforms like Twitter, Tumblr or Posterous, diverted some of the attention blogs were previously getting from the general public. As a consequence, resilient bloggers invest in producing high quality content that surpasses the simplicity of sharing a short thought or the last holiday's photos — take for instance Mashable or the Gawker blog network. As content creators spread among these services, blogs become professionalized. Moreover, blogs are a good choice for generating structured, dated content, so it is not uncommon for web sites to use a blogging platform, like Wordpress, as their content management backend.

Link analysis is a technique frequently used in the ranking of web pages or sites. Metrics like in-degree or PageRank (Brin and Page 1998) are frequently used to impose some relevance order in the web graph. In the context of blogs, however, we can take advantage of distinctive features that are not generally present in other types of web pages. The date and the grouping of posts by blog are some of these unique characteristics that give us the chance to explore different ranking methods.

In this paper, we present a few experiments based on a large blog collection, the TREC Blogs08 collection (Macdonald, Ounis, and Soboroff 2010, Section 2), containing over 28 million posts. Branco (2008) has previously explored the application of the h-index in the blog ranking task. We conduct here a comprehensive and controlled assessment of this approach. During our participation in TREC 2010 Blog Track (Devezas, Nunes, and Ribeiro 2010), we combined the h-index with query-dependent ranking functions, applying it to a much larger collection than the collection used by Branco (26 times more blogs and 10 times more posts), looking for an optimal weight for the h-index component while examining the gain introduced in the retrieval system.

We expand this work and evaluate the results of the experiments using the query relevance assessments provided by TREC and comparing the in-degree and h-index of the top ranked blogs according to both metrics. Our main concern is to determine the h-index's robustness and quality as a blog ranking metric, in real case scenarios where, for instance, the existence of link clutter should be taken into consideration.

The H-index

Hirsch proposed the h-index (Hirsch 2005) as a measure for the scientific output of a researcher. In bibliometrics, the h-index is used to rank a scientist or scholar, based on the productivity and impact of his/her publications. The central idea of the metric is that the number of citations alone or the number of published papers alone aren't directly correlated to the author's importance. Thus, the h-index depends both on work volume and number of citations. If an author published just a few papers, that turned out to be highly cited, and then ceased publishing, the h-index would be bound by the low number of publications. Similarly, if an author published a considerable amount of papers, but didn't get many citations, the h-index would be limited by the low number of citations. As proposed by Hirsch:

> 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.

Experience with the h-index outside the bibliographic realm (Branco 2008) has shown that this metric might cause a positive impact on the blog ranking task. We aim to further investigate this hypothesis, by exploring larger collections, and by emphasizing some of the advantages of using

Post	**1**	**2**	**3**	4	5
In-degree	**16**	**16**	**6**	3	1

Table 1: Example of a post in-degree list.

the h-index in real blog networks.

Blog Ranking

There are several methods for blog ranking, ranging from the simplicity of the in-degree to the specificity of the iRank (Adar, Zhang, and Adamic 2004), a metric that takes into consideration the importance of blogs in the propagation of information.

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. In resemblance to bibliometry, a blog's relevance should be measured by taking into account both the number of in-links and the number of published entries, thus taking advantage of the grouping of posts by blog, together with their individual number of in-links, in order to more efficiently rank blogs. Hence, we hypothesize that the h-index might be a strong ranking metric while requiring a relatively small processing power to compute. From the definition of h-index:

> 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.

Calculating the H-index of a Blog

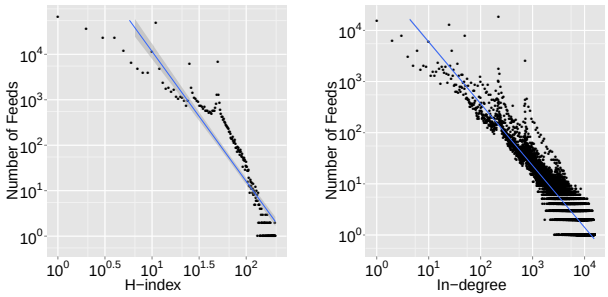
We determine the h-index of a blog based on the in-degree of its posts. Algorithm 1 illustrates the simple steps taken for the calculation of the h-index. Given a list with the number of in-links for each post, we sort it and reverse it. Then, we increment the value of h until the in-degree is larger than h or we reach the end of the list.

Table 1 shows an example of a blog with 5 posts, each with a given number of in-links. According to the h-index definition, the value of h is 3, as there are 3 posts with at least 3 in-links each and the other 2 posts have no more than 3 in-links. To clarify, if we assumed $h = 2$, we would find that one of the other 3 posts has more than 2 in-links, which violates the definition. On the other hand, if we to assumed $h = 4$, we wouldn't find 4 posts with at least 4 in-links, since the post at rank 4 has 3 in-links.

Algorithm 1 Blog's h-index calculation.

Input: List L of post in-degrees for a blog.
Output: H-index of the blog.

```
L ← sort(L)
L ← reverse(L)
h ← 0
while L_h > h do
    h ← h + 1
end while
```



(a) H-index distribution. (b) In-degree distribution.

Figure 1: The TREC Blogs08 Collection.

Experimentation

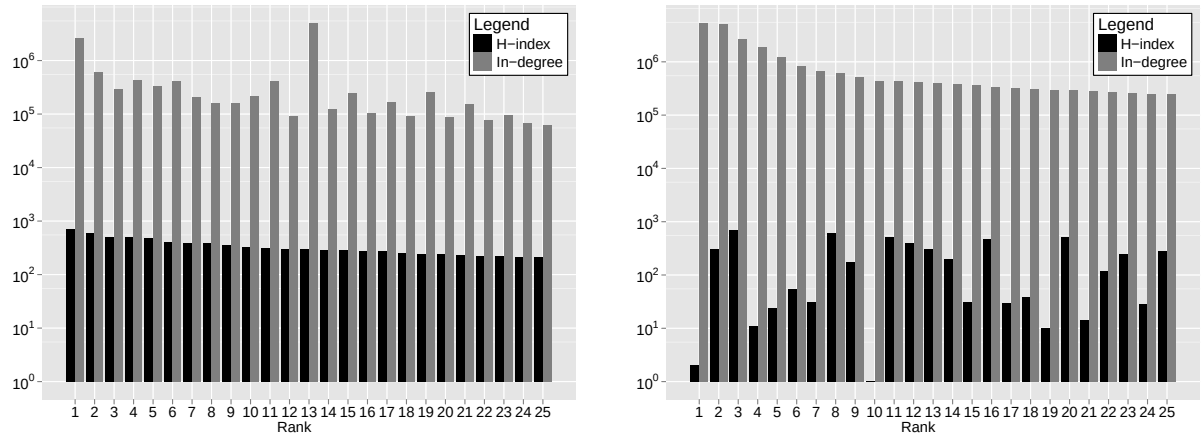
We conduct some experiments based on the TREC Blogs08 collection, using the in-degree as baseline to assess the quality of the h-index on the blog ranking task. We analyze both distributions, comparing the in-degree and the h-index, for the top ranked blogs according to both metrics. We then explore the influence of link clutter while combining these link-based metrics, as a query-independent feature, with the BM25 score, in an information retrieval context.

TREC Blogs08 Collection

The TREC Blogs08 collection is a compilation of over 1.3 million blogs, comprising more than 28 million posts, with a total compressed size of 453 GB. The files are formatted in a way that can easily be indexed by the Terrier platform (Ounis et al. 2006). By using this platform and by taking advantage of the query relevance assessments that TREC provides, we are able to explore the influence of the h-index in the blog distillation task. We want to know whether we can improve the blog retrieval process by combining the h-index, a query-independent feature, with query-dependent scores, such as the BM25 score. We carry on this line of work, by exploring the application of the h-index in the same task, now by calculating it based on three different versions of the post graph. The first version of the graph corresponds to the full graph we used for TREC, which includes post self-citations and link multiplicity. The second graph is identical to the full graph without self-citations or loops. And finally, for the third graph, we apply another layer of cleaning, by removing both loops and link or edge multiplicity.

Data Extraction

In order to compute the h-index, we first need a post in-degree list for each blog. To calculate these values, we follow some simple steps. First, we parse each post, extracting every URL found on the *href* attribute of the HTML anchors in the entries. Next, we remove the URLs whose domains don't belong to the blog collection, and reverse the representation of the graph. This results in a text file containing a list of posts associated with their respective in-links. Sorting this file allows us to easily group the posts by blog, count the number of in-links of each post, and calculate the blog's h-index.



(a) Top 25 blogs according to the h-index.

(b) Top 25 blogs according to the in-degree.

Figure 2: Comparison between the h-index and the in-degree as blog ranking metrics, in the full Blogs08 post graph.

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

Table 2: Top 5 blogs ranked by h-index according to the full Blogs08 post graph blog.

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

Table 3: Top 5 blogs ranked by in-degree according to the full Blogs08 post graph.

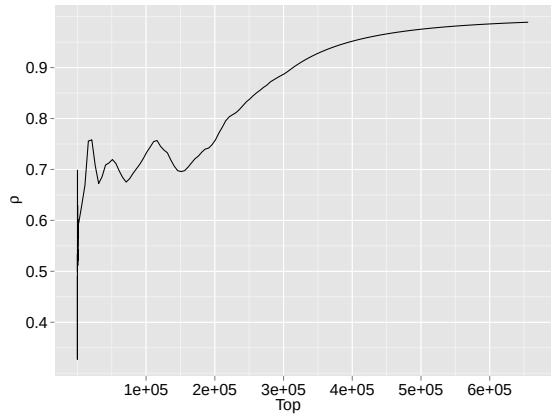


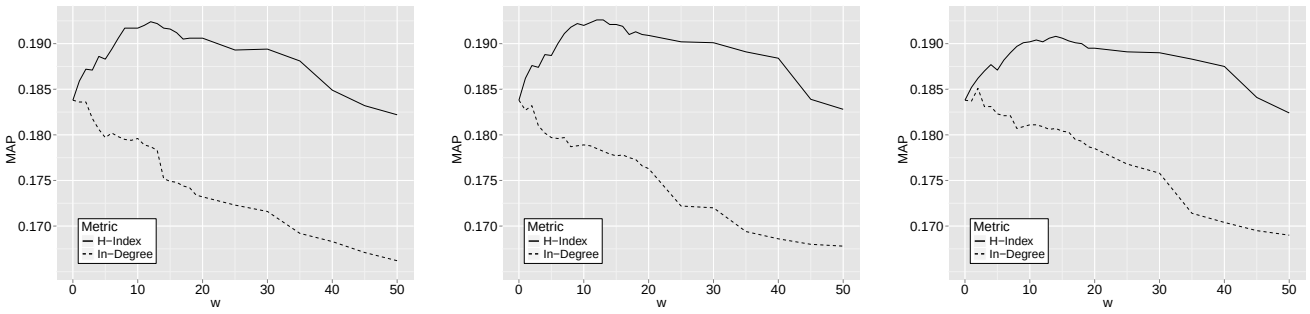
Figure 3: Spearman's correlation coefficient.

Data Analysis

Figure 1 depicts the in-degree and h-index distributions for the TREC Blogs08 collection. Values for the in-degree are highly diverse, ranging from 0 to 5,345,032, with 15,309 distinct values. On the other hand, there are only 205 distinct h-index values, ranging from 0 to 700 — the h-index is bound by the number of posts in a blog. Figure 2 depicts the values of both the in-degree and the h-index, for the top 25 blogs, according to (a) the h-index and (b) the in-degree. Looking at Figure 2(a), we verify that, 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 — Table 2 illustrates some of these values. On the other hand, Figure 2(b) depicts a different scenario, where we can find extremely low h-index values for some of the highest in-degree scores. By looking at Table 3, we can see that the h-index for the highest ranked blog according to the in-degree is 2, meaning that there are only 2 posts in $bodyelectric.blogspot.com$ with 2 or more in-links from other blogs in the collection. This illustrates one of the strengths of the h-index when compared with the in-degree, in the blog ranking task, the ability to account for the number of posts in a blog.

We rank blogs according to their in-degree and h-index, and calculate Spearman's rank correlation coefficient (Bar-Ilan 2005) for the $top\ k$ blogs, where k ranges from 25 to 1000, in steps of 25, and then from 1000 to 659,452 (the total number of blogs with at least one in-link or one out-link), in steps of 5000. Figure 3 depicts the evolution of Spearman's ρ for a progression of cuts of the in-degree and h-index rank lists. The value of ρ is constantly smaller than 0.76, for cuts bellow 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. We use the TREC Blogs08 collection, that we had previously indexed using the Terrier platform, and



(a) Full Blogs08 post graph.

(b) Blogs08 post graph without loops.

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

Figure 4: MAP for the combined BM25 score and link-based metric, in the TREC Blogs08 Collection.

base our explorative work on the query topics and relevant assessments provided for TREC 2010. We determine the values for the link-based metrics — the query-independent features of our system — based on the three different versions of the post graph previously described. We then include the h-index and the in-degree in the final score, by adding $w \times log(metric)$ to the BM25 score, where $metric$ is either the h-index or in-degree value. Equation 1 shows how to calculate the score for a query q and a blog b based on the h-index metric:

$$score(q, b) = BM25(q, b) + w * log(h\text{-}index(b)) \quad (1)$$

Figure 4 depicts the evolution of the mean average precision (MAP) across the one hundred TREC topics, as we increase w, the weight of the in-degree or h-index in the final score. As we remove link clutter, we notice that the MAP values for the results based on the in-degree metric tend to improve in the absence of graph loops and edge multiplicity. On the other hand, MAP values for the results based on the h-index are very similar for (a) and (b), indicating that the removal of graph loops has only a light influence on the results. By removing edge multiplicity (c), the values of MAP actually decrease, indicating that edge multiplicity is relevant to the computation of the h-index for the blog ranking task.

Conclusions

We have ranked blogs according to two distinct metrics, the in-degree and the h-index, while studying their behavior when affected by different levels of link clutter. We applied these metrics to an information retrieval context by using them as query-independent features, and evaluated their quality based on the relevance assessments for the query results we obtained. Using the h-index metric we were able to consistently improve the quality of the results over the BM25 baseline, for three different versions of the post graph. The boundaries imposed by the h-index, regarding the number of posts and in-links, allow this metric to establish a more balanced measurement of blog relevance than the in-degree metric, specially improving the ranking of the top blogs.

Acknowledgements

We thank Filipe Coelho for all the fruitful discussions. This work was partially supported by a research grant from Labs SAPO/UP.

References

Adar, E.; Zhang, L.; and Adamic, L. 2004. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*.

Bar-Ilan, J. 2005. Comparing rankings of search results on the Web. *Information Processing & Management* 41(6):1511–1519.

Branco, J. 2008. *Aplicação do h-index em blogues*. Master's thesis, Faculty of Engineering, University of Porto.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*.

Devezas, J. L.; Nunes, S.; and Ribeiro, C. 2010. FEUP at TREC 2010 Blog Track: Using h-index for blog ranking. In *The Nineteenth Text REtrieval Conference Proceedings (TREC 2010)*.

Hirsch, J. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA* 102(46):16569–16572.

Macdonald, C.; Ounis, I.; and Soboroff, I. 2010. Overview of the TREC 2009 Blog track. In *The Eighteenth Text REtrieval Conference Proceedings (TREC 2009)*.

Ounis, I.; Amati, G.; Plachouras, V.; He, B.; and Macdonald, C. 2006. Terrier: A high performance and scalable information retrieval platform. *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.