

Resumo

O presente relatório pretende dar a conhecer o trabalho realizado no âmbito da adaptação da medida de importância h-index ao contexto dos blogues. O h-index foi proposto em 1995 por Jorge E. Hirsch, como um valor de classificação de qualidade e produtividade de elementos da comunidade científica. A medida tem em conta não só o número de publicações que um cientista produz, como também o impacto destas no resto da comunidade. O impacto da produção de um cientista é avaliado com base no número de vezes que os seus artigos são citados pelos seus colegas. Nas palavras de Hirsch, o valor h atribuído a um cientista significa que:

“Um cientista tem index h se h dos seus N artigos têm cada um pelo menos h citações, e os outros $(N - h)$ artigos não têm mais que h citações cada um”.

A adaptação do índice a blogues criou a necessidade de que estabelecêssemos um paralelismo entre as características de ambos os contextos, tarefa que acabou por ser de simples execução. Optámos por que cada blogue correspondesse a um cientista, e as suas entradas à documentação publicada pelo indivíduo.

É apresentada uma análise da colecção em que realizámos os testes, descrevendo alguns dos comportamentos que verificámos nos bloguistas portugueses. Observámos a distribuição dos blogues na nossa colecção por fornecedor, bem como a frequência de novo conteúdo ou blogues ao longo do tempo. Estas análises permitiram-nos verificar o crescimento rápido em que a blogosfera portuguesa se encontra, devido às características apelativas dos blogues, e obter alguns dados curiosos relativos ao comportamento típico de um bloguista português.

Implementámos o h-index, bem como uma das variantes existentes do método, o g-index. Para fins de comparação, extraímos também a ordenação de blogues segundo o número de ligações de entrada de cada blogue. Obtivemos resultados que nos permitiram concluir que os dois índices obtêm classificações bastante diferentes da obtida segundo o número de citações. Dados os resultados satisfatórios na comparação do h-index com uma medida mais tradicional, tornou-se apelativo testar a medida na ordenação de resultados de pesquisas em blogues. Decidimos utilizar a ferramenta open-source Terrier, um motor de pesquisa altamente configurável e flexível. Testámos pesquisas com o h-index, o g-index e a contagem de citações como critérios. Com quase trezentos testes, o desempenho do h-index como critério para ordenação de resultados de pesquisa apresenta resultados positivos.

Concluímos que o h-index e o g-index, permitem o cálculo de valores para uma ordenação de blogues bastante diferente dos métodos já existentes, pelo que podemos afirmar que estas medidas possuem valor considerável para serem utilizadas ou pelo menos testadas como critério de avaliação na área de recuperação de informação.

Abstract

This report presents the work done in the adaptation of the h-index to the blogs context. The h-index was proposed in 1995 by Jorge E. Hirsch, as a measure for the quality and productivity of a scientist. It considers not only the number of produced articles but also their impact in the rest of the scientific community. The impact of a scientist's work is evaluated based on the number of times his work is cited by someone else. Hirsch states that:

“A scientist has index h if h of his/her N papers have at least h citations, and the other $(N - h)$ papers have no more than h citations each”.

Adapting the index to blogs required us to establish a parallelism between the characteristics of both contexts, a task proved easy. We defined that each blog would correspond to a scientist, and its posts to the papers published by the individual.

We present an analysis of the dataset that we worked with, describing some of the behaviors we observed on the portuguese bloggers. We describe the distribution of the blogs in our dataset by provider, and the frequency of new posts or blogs published through the years. This analysis allowed us to observe the rapid growth of the portuguese blogosphere, and to obtain some interesting data related to the typical behavior of the portuguese bloggers.

The h-index was implemented, as well as one of its variants, the g-index. For the purpose of comparison, we extracted the ranking of blogs according the number of inlinks for each blog. We obtained results allowing us to conclude that the two indexes obtain very different rankings from the ones resulting from the use of more traditional methods. Given these good results, we have decided to use an open-source tool, Terrier, a highly customizable and flexible search engine. We conducted tests with the h-index, the g-index and the inlinks counting as features, which allowed us already to get to some conclusions. With almost three hundred tests, the h-index has obtained positive results, with a good performance if compared with the other methods.

We conclude that the h-index and its variant, the g-index, obtain very different values for the ranking of blogs, allowing us to state that these measures are worth being used or at least tested as criteria for evaluation and ranking in the area of information retrieval.