FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Aplicação do h-index em blogues

José Mário Castelo Branco

Relatório de Dissertação

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Maria Cristina de Carvalho Alves Ribeiro

Co-orientador: Sérgio Sobral Nunes

Julho de 2008

Aplicação do h-index em blogues

José Mário Castelo Branco

Relatório de Dissertação
Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo júri:

Presidente: Gabriel de Sousa Torcato David

Arguente: Pavel Pereira Calado

Vogal: Maria Cristina de Carvalho Alves Ribeiro

16 de Julho de 2008

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.

Agradecimentos

Em primeiro lugar gostaria de agradecer à Prof. Cristina Ribeiro e ao Eng. Sérgio Nunes, as pessoas responsáveis pela ideia para esta dissertação, por me orientarem no seu decurso, e por serem impecáveis no seu apoio. Obrigado pela oportunidade, pela ajuda e enorme motivação.

Queria agradecer também aos Eng. João Pedro Gonçalves e Maria João Nogueira, da SAPO, por partilharem dados e conhecimento connosco. Também a Craig Macdonald devo um agradecimento, pelo apoio dado na área de recuperação de informação.

Agradeço de igual modo ao José Pedro Pinto, com quem colaborei na fase inicial do trabalho, e um grande obrigado a Catalin Calistru, pela sua generosidade e prontidão para ajudar.

Um agradecimento especial ao Ivo Navega cuja amizade nunca conseguirei suplantar. Ao António Mota e ao Hugo Valente, pela participação nos testes, ajuda, e pela sua amizade.

Finalmente, agradeço aos meus pais, irmã, e à Carla, cujo apoio incondicional sustenta todo o trabalho que tenha feito e possa alguma vez fazer.

José Mário Castelo Branco

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Na área da recuperação de informação e classificação de blogues é já usual utilizar vários critérios para o cálculo de importância de blogues ou entradas (*posts*). Pretendíamos avaliar o desempenho do h-index como medida de importância de blogues, e a viabilidade da sua inserção como critério no processamento de ordenações de blogues.

1.1 Motivação

Ao longo dos últimos anos, os blogues têm ganho um lugar de destaque como tema de investigação. Foi no âmbito desta área em pleno desenvolvimento que a tese desta dissertação foi proposta. Assim, apresentou-se a ideia de aplicar o algoritmo de J. E. Hirsch, o h-index, como critério para a classificação e ordenação de blogues. Definindo um paralelismo entre as circunstâncias nas quais o índice tem vindo a ser originalmente aplicado, a produção científica de cientistas, e as características típicas de um blogue, adaptámos o índice ao contexto da blogosfera. A ideia colocada em vigor nesta dissertação é inovadora e única até à data, podendo oferecer novos resultados ainda por explorar.

Além do cálculo de uma ordenação através da obtenção do h-index para cada um dos blogues, a utilização deste índice como critério para apresentação de resultados na área de recuperação de informação pode proporcionar-nos a oportunidade de observar e analisar o desempenho de um novo peso de relevância e importância para pesquisas.

1.2 Enunciado da Tese

Pretendemos provar que o h-index poderá ser utilizado com resultados satisfatórios na área de medição de importância de blogues. Foi-nos cedida uma colecção de blo-

gues maioritariamente portugueses (quase na sua totalidade lusófonos), no contexto de uma colaboração entre a Faculdade de Engenharia da Universidade do Porto e a empresa SAPO[1]. Obtivemos assim uma colecção contendo uma parte da blogosfera portuguesa, para análises experimentais. Procurámos formas de aplicar a medida neste novo contexto, para que pudesse vir a ser utilizada como critério de avaliação de importância de blogues. Propusemo-nos realizar testes com utilizadores reais esperando obter resultados relevantes e úteis para o fim pretendido. Além disso, pretende-se que os utilizadores, sem ter conhecimento de que critérios são utilizados para a apresentação de resultados para uma pesquisa, indiquem o h-index como uma mais-valia como critério num motor de pesquisa. Assim, enunciámos a tese da seguinte forma:

"O h-index apresenta-se como uma opção viável para a classificação de blogues, e é passível de ser considerado como uma nova medida, oferecendo valores e consequentes ordenações diferentes das já existentes"

1.3 Objectivos

Tomámos como objectivo fundamental a aplicação do h-index para avaliação da importância de um blogue, bem como a análise dos resultados obtidos. Esperámos obter conclusões relativamente ao desempenho que a utilização do h-index poderá proporcionar, e que estas conclusões sejam também obtidas a partir de testes com utilizadores. Pretendíamos criar situações reais em que um utilizador, ao realizar uma pesquisa por um determinado termo, recebessem os resultados (blogues) ordenados por relevância e pelos seus valores *h*.

Pretendeu-se também fazer uma análise à colecção cedida pela SAPO, uma amostra da blogosfera portuguesa, esperando obter uma análise dos hábitos do bloguista português típico e um número de características genéricas dos dados que nos foram oferecidos.

1.4 Estrutura do Documento

No primeiro capítulo foram analisados trabalhos anteriores relativos ao h-index, com especial atenção à própria medida. Foi descrito detalhadamente o funcionamento do h-index, bem como a sua aceitação na comunidade científica, e análises ao índice, após o seu lançamento. Na segunda parte do capítulo foram descritas variantes propostas para corrigir alguns aspectos do h-index considerados desfavoráveis.

No segundo capítulo, foi feito um apanhado de investigações previamente realizadas na área da classificação de blogues e de análises da blogosfera. Referiram-se casos de estudo quer de blogosferas nacionais, como a portuguesa e a iraniana, quer de comportamentos dos bloguistas, como análises sociológicas da relevância dos blogues para os

[1] http://www.sapo.pt/

media. Posteriormente mencionam-se estudos e artigos na área de classificação de blogues, e métodos utilizados para elaborar ordenações dos mesmos. No terceiro capítulo, é feita uma análise da colecção cedida pelo portal português, e são tiradas conclusões relativas ao comportamento do bloguista português.

No quarto capítulo, é descrita a implementação do h-index para classificação de blogues. Os resultados são analisados e comparados com outros métodos, verificando-se o desempenho de cada um, individualmente. Finalmente, no último capítulo, é descrita a integração do h-index em motores de pesquisa de blogues, utilizando-o como critério para a ordenação de resultados.

Capítulo 2

O h-index de Hirsch

Jorge E. Hirsch propôs em 2005 [Hir05] um índice que proporcionou um novo método de classificar o trabalho de um investigador. Este método, baptizado por Hirsch como h-index, procura avaliar a produtividade e impacto individual de um cientista ao longo da sua carreira. Após a sua apresentação, rapidamente surgiram na comunidade científica análises à eficácia do método aplicado em vários universos, além de propostas de melhoramentos para o mesmo. Neste capítulo descrevemos as principais características do h-index de Hirsch, bem como algumas das variantes posteriormente desenvolvidas.

2.1 O h-index

O método h-index procura exprimir a qualidade de um cientista em função da quantidade de artigos que produz, bem como o interesse que estes despertam na comunidade científica. Assim, ao contrário de outros métodos que pesam somente o número de citações feitas ao artigo de um indivíduo — o que pode de alguma forma medir o seu impacto, mas sendo esta medida facilmente manipulável — o h-index procura avaliar todo o conjunto de trabalho de investigação que um cientista desenvolve.

Hirsch dá como exemplo a atribuição de prémios, nomeadamente o prémio Nobel, que procura distinguir não um trabalho em particular de um cientista, mas sim o seu impacto na comunidade científica ao longo da sua vida profissional. Para o cálculo do valor de h-index de um indivíduo é utilizado o número de artigos científicos que produziu até à data, juntamente com o número de vezes que estes são citados noutros documentos. Nas palavras de Hirsch [Hir05], o valor h-index é obtido da seguinte formulação:

> "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".

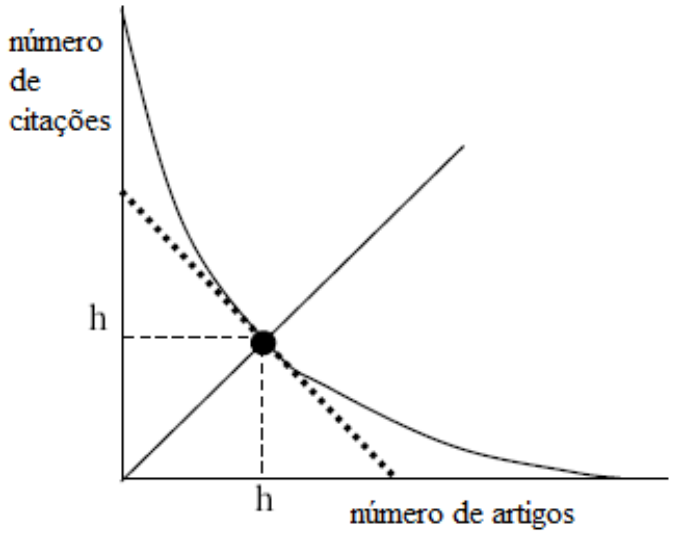

Figura 2.1: Método de cálculo do h-index [Hir05]

A figura 2.1 ilustra o raciocínio por detrás do cálculo do valor *h* de um cientista, enquanto que na tabela 2.1 é demonstrado um exemplo do seu cálculo para um indivíduo. Repare-se no valor a negrito na tabela, representando o valor *h* obtido neste exemplo.

Tabela 2.1: Exemplo de cálculo do valor *h*

Documento	No. de citações
Doc.1	24
Doc.2	17
Doc.3	15
Doc.4	11
Doc.5	8
Doc.6	**6**
Doc.7	4
Doc.8	2
Doc.9	2

A recepção ao h-index foi activa e rapidamente surgiram análises do método. Apenas um ano após a publicação do artigo de Hirsch, haviam já sido publicados pelo menos 30 documentos relatando investigações em redor do índice [BD07]. Um número da revista *Scientometrics* seria dedicado integralmente ao h-index, e o valor é calculado automaticamente na função de relatório de citações no *Web of Science* [Meh07]. Foram propostas muitas variantes para este algoritmo, quer orientando o cálculo deste método para outras áreas, quer procurando corrigir desvantagens que iam sendo apontadas.

Uma questão que rapidamente se coloca é relativa a auto-citações, ou seja, a situação em que um cientista cita no seu mais recente artigo documentos que terá criado previa-

mente. Naturalmente, se considerarmos auto-citações para o cálculo desta classificação, o valor *h* beneficiará cientistas que mencionem frequentemente os seus trabalhos anteriores. A inclusão de auto-citações não influenciará uma classificação de cientistas através do uso do h-index como afectaria, por exemplo, uma obtida através de somente o número de citações. Porém, Hirsch, prevendo esta questão, menciona que será de facto aconselhável remover as auto-citações da contagem para o cálculo de *h* [Hir05]. Esta remoção de auto-citações poderá ser realizada de forma absoluta, removendo o número de auto-citações do cálculo na sua totalidade, ou removendo simplesmente as presentes nos artigos com número de citações acima do valor *h* do cientista. Ajustar-se-ia este valor em função das alterações observadas, obtendo assim um novo valor *h*. A desvantagem apontada pelo autor nesta solução é que desta forma, se um cientista pretendesse aumentar o seu valor *h*, procuraria citar os seus artigos situados abaixo do presente valor *h*, obtendo um aumento gradual. Concluiu assim, que provavelmente a melhor opção para o cálculo de h-index passa pela ausência de auto-citações nas contagens [Hir05].

Um ano após a apresentação do h-index, em 2006, Wolfgang Glanzel publicou uma análise às vantagens e desvantagens do h-index [Gla06], vindo a oferecer uma importante avaliação das características do cálculo do h-index e classificações de cientistas a partir deste valor.

Glanzel enunciou as seguintes vantagens do h-index:

- O h-index é um indicador simples e facilmente implementável;
- É combinada produtividade e qualidade, no decurso da vida profissional de um investigador;
- Apresenta robustez, visto que apenas um artigo não afecta grandemente os valores de h;
- Pode ser considerado para a avaliação segundo este método qualquer tipo de documento;
- É possível ser utilizado em conjunto com outros quaisquer indicadores.

Por outro lado, apresentou as seguintes desvantagens para a medida:

- Cientistas no início de carreira estarão em desvantagem numa classificação através deste método;
- Mesmo avaliando a produtividade, o h-index possibilita a situação em que um indivíduo não producente veja ainda o seu valor *h* aumentar, devido a novas citações referentes a trabalhos anteriores;
- Cientistas com um pequeno número de documentos produzidos, se bem que de grande qualidade e impacto, obterão um valor *h* baixo;

- O h-index apresenta resultados positivos na avaliação de cientistas com excelentes desempenhos, mas pode falhar com desempenhos mais medianos;

A título de conclusão, Glanzel mencionou que podem surgir problemas com a utilização do h-index em algumas situações. Mesmo assim, concluiu que o h-index se revela um método que permitirá obter resultados positivos, especialmente para a avaliação no contexto de pequenas quantidades de documentos[Gla06].

Hirsch propôs que o h-index fosse aplicado a várias outras áreas de investigação, alertando que os valores de avaliação de qualidade por ele definidos muito provavelmente teriam de ser reajustados quando este método fosse usado em novos contextos. Valores h podem, dependendo do contexto em que a medida fôr aplicada, representar níveis diferentes de importância. Em 2006, Cronin e Meho implementaram o h-index para classificação de investigadores na área da Ciência de Informação. Tomaram como objectivo principal a comparação dos resultados obtidos com o cálculo do h-index com os observados através da utilização de uma simples contagem de citações para cada indivíduo. Foram incluídos nesta avaliação 31 investigadores norte-americanos notáveis bem como mais de 100 cientistas de diferentes nacionalidades. Tendo em conta que o valor h valoriza fundamentalmente a produtividade/impacto no decurso de uma carreira, consideraram-se no grupo a analisar investigadores com uma carreira já algo extensa. A contagem de citações e o cálculo do h-index foram feitos de duas formas, considerando e excluindo auto-citações, com vista a avaliar o impacto da inclusão destas na ordenação obtida. Os documentos contados incluíam artigos de opinião, cartas a editores, artigos de conferência e críticas, obtendo-se assim um grupo abrangente da obra escrita dos cientistas a avaliar [CM06] .

Após o cálculo do h-index para o grupo de cientistas-alvo, obtiveram-se valores entre 84 e 1048, se removida a contagem de auto-citações, senão, variaríam entre 79 e 1025. Observa-se que a presença de auto-citações não provoca modificações radicais, mas se estas forem excluídas são visíveis algumas alterações relevantes nos valores h calculados. Assim, Cronin e Meho concluíram que apesar do número de citações obtido por um cientista funcionar como um indicador satisfatório de sucesso profissional, o cálculo do h-index apresenta-se como um factor discriminador valioso se adaptado, como Hirsh refere, às condições e características da área em que é implementado [CM06].

2.2 Alternativas ao h-index

Após a apresentação do h-index à comunidade científica, e no seguimento de análises às vantagens e desvantagens do algoritmo, começaram a surgir propostas de variantes para o índice de Hirsch, com melhoramentos que proporcionavam a correcção de desvantagens existentes, ou melhor adaptação a outros contextos que não os inicialmente pretendidos.

g-index Em 2006 o investigador Egghe, L. publicou um artigo apresentando a sua variante ao h-index, o g-index [Egg06]. Segundo Egghe, era necessário aumentar a robustez do algoritmo de Hirsch na parte superior da ordenação, ou seja, acima do número h atribuído. Apontou o facto de, no caso dos artigos superiores a este nível, ser irrelevante o número de citações que apresentassem, pois o valor de h não seria alterado por muito popular que fossem estes documentos de topo na carreira de um indivíduo. Por outras palavras, dois cientistas A e B podem possuir o mesmo valor h, mesmo tendo o cientista A nos seus artigos de topo alguns com apenas algumas citações (um número superior ao seu h atribuído), e o cientista B artigos com milhares de citações. De acordo com o algoritmo do h-index, estes cientistas são idênticos segundo este critério, o que Egghe aponta como sendo uma característica a corrigir.

O cálculo do g-index permitiria assim uma maior distinção entre o trabalho dos cientistas A e B, do exemplo que referimos, obtendo o segundo um valor g maior do que o primeiro, dando uma importância maior do que no caso do h-index ao impacto dos artigos de maior popularidade no meio científico.

O algoritmo para o cálculo do g-index é, em simplicidade e funcionamento, semelhante ao do h-index. No caso do índice de Egghe, em vez de se considerar o número de citações para cada documento, é usado o somatório de citações a partir do topo da ordenação σNc, e este valor é comparado com o quadrado da posição desse artigo na ordenação, r^2. Na última posição da lista em que o valor de σNc é maior que o r^2 do cientista, o valor dessa posição é o g atribuído. Nas tabelas 2.2 e 2.3 Egghe demonstra a comparação dos valor h e g (a negrito nas tabelas) obtidos por dois cientistas diferentes (ele próprio e Small, H.), em que se observa como o g-index é de facto mais recompensador para cientistas com obras de grande impacto, bem como a exemplificação do cálculo de g e h para cada um deles. Comparando a obra de Egghe com a de Small verificamos que estes possuem valores de h de 13 e 18, respectivamente. Porém, analisando os artigos de topo, é fácil notar que os artigos de Small possuem um número de citações muito superior aos de Egghe, sugerindo que a diferença de valores para os índices atribuídos a cada um deveria de facto ser maior. Através do cálculo do g-index, esta diferença é superior, e aparentemente, mais justa e representativa da realidade, obtendo Egghe um valor de 19, e Small, de 39.

H(2)index e a-index O H(2) index, tal como o g-index, procura proporcionar mais peso a artigos fortemente citados. O seu criador, Kosmulski (2006) [Kos06] definiu o algoritmo:

> "O índice $h(2)$ de um cientista é definido como o número natural mais alto em que os seus $h(2)$ artigos mais citados receberam cada um pelo menos $h(2)^2$ citações."

.

Tabela 2.2: Cálculo dos valores h e g de Egghe [Egg06]

TC	r	$\sum TC$	r^2
47	1	47	1
42	2	89	4
37	3	126	9
36	4	162	16
21	5	183	25
18	6	201	36
17	7	218	49
16	8	234	64
16	9	250	81
16	10	266	100
15	11	281	121
13	12	294	144
13	**13**	307	169
13	14	320	196
13	15	333	225
12	16	345	256
12	17	357	289
12	18	369	324
12	**19**	**381**	**361**
11	20	392	400

Um cientista com $h(2)$ de 10, por exemplo, possuirá no mínimo 10 artigos que tenham sido citados 100 vezes cada um. Tomando o conceito de "núcleo de Hirsh"de Rousseau [JLRE07], que consiste num grupo de publicações de grande impacto na carreira de um cientista, proposto por Jin et al. em 2006 [Jin06] o algoritmo que viria a ser baptizado por Rousseau de a-index. O *a* provém de *average*, pois é calculada a média para o valor h de um cientista. A fórmula de cálculo para este valor é a seguinte:

$$\frac{1}{h}\sum_{j=1}^{h} cit_j \tag{2.1}$$

onde h = h-index e cit= número de citações.

r-index Um ano mais tarde, em 2007, Jin et al. apontaram que o a-index proposto por Jin era injusto para cientistas de topo, pois ao envolver uma divisão pelo valor de h no cálculo desse algoritmo, indivíduos com um valor h elevado são prejudicados [JLRE07]. Assim, apresentaram um novo índice que corrigiria esta desvantagem: o r-index. Neste caso, o valor de h de um cientista é alterado não por uma divisão pela média dos seus valores, mas sim pela sua razão. O cálculo do valor r é calculado segundo a fórmula:

$$\sqrt{\sum_{j=1}^{h} cit_j} \tag{2.2}$$

onde h = h-index e cit= número de citações.

ar-index No mesmo ano, Jin publicou uma nova proposta para índice, o ar-index, uma adaptação do r-index para contexto temporal. Este índice não contabiliza somente a quantidade de citações no núcleo de Hirsch, considera também a idade das publicações aí presentes. Desta forma, ao contrário dos índices anteriores, o valor de *ar* pode não só aumentar ao longo do tempo, como também diminuir. Este índice foi definido por Jin como "a raíz quadrada do somatório do número médio de citações por ano de artigos incluídos no núcleo de Hirsh"[Jin07].

$$\sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}} \tag{2.3}$$

onde h = h-index, cit= número de citações e aj = número de anos desde publicação.

Comparações No ano de 2008 Bornmann et al., investigadores da Universidade de Zurique, publicaram uma análise reportando uma comparação entre o h-index e nove das suas variantes [BMD07]. Estabeleceram como objectivo a resposta à questão:

"Do ponto de vista prático, qual dos índices disponíveis deverá ser escolhido?"

Tomaram como base de teste uma colecção de candidaturas de investigadores doutorados à inscrição na Boehringer Ingelheim Fonds (B.I.F.)[1], uma fundação de apoio à investigação na área da biomedicina. Os candidatos a membro da fundação são avaliados por um quadro de sete investigadores de renome mundial, procurando-se seleccionar investigadores com uma carreira com a excelência pretendida. A amostra usada por Bornmann et al. incluía 414 cientistas candidatos no período de 1990 a 1995, possuindo entre si um total de 1.586 artigos publicados, com um conjunto de 60.882 citações.

Bornmann et al. propuseram também, no seu artigo, um índice variante do a-index. Argumentaram que a média do número de citações não deveria ser utilizada como medidor de tendência, pelo que sugeriram em sua vez o uso do valor da mediana do número de citações obtidos por artigos no núcleo de Hirsch.

Uma das análises mais interessantes efectuadas consistiu no cálculo de regressões logísticas, medição ideal para avaliar a correlação entre as decisões do quadro de avaliadores e as ordenações obtidas pelos índices estudados. Estabeleceu-se a comparação entre

[1] http://www.bifonds.de

os valores de h-index e de m-index com as decisões tomadas no período de 1990 a 1995, através do cálculo de uma medida de associação. Os resultados obtidos aparentaram favorecer o uso do m-index, obtendo este índice valores consideravelmente mais elevados do que o h-index.

Bornmann et al. apontam na conclusão do seu artigo que, apesar de muitas variantes para o h-index terem sido propostas desde o seu surgimento, poucos estudos haviam sido feitos, até à data, que avaliassem a correlação entre estes. Mencionam estudos feitos com este objectivo que concluem que existe entre estes algoritmos uma forte correlação, conclusão que as suas próprias observações parecem apoiar. Estes autores concluiram que apesar do grande número de variantes, estas têm vindo a resultar em poucas ou nenhumas melhorias empíricas. Através de uma análise exploratória Bornmann et al. demonstraram poder assumir a divisão do grupo de índices em dois: o tipo de índices que aponta o núcleo mais produtivo de um cientista, e o número de artigos que este contém; e o tipo que descreve o impacto dos artigos no núcleo da sua bibliografia publicada. Os investigadores apontam ainda que, mesmo apresentando estes dois grupos um foco diferente, se complementam com eficácia, sugerindo que dois índices de diferentes grupos sejam usados em conjunto, para que se obtenham melhores resultados, tal como havia já feito antes Jin et al [JLRE07]. De acordo com os resultados obtidos na análise de regressão, e representando o h-index o grupo de cálculo do núcleo mais produtivo, e o m-index o grupo avaliador do impacto do núcleo produtivo, o segundo grupo apresentou resultados mais semelhantes aos do quadro de avaliação do B.I.F. Segundo Bornmann et al., esta conclusão é bastante explicável:

> "Estes resultados apresentam-se de acordo com a declaração do director do B.I.F, Fröhlich, em que referia que o quadro procurava excelência em performance científica num candidato. A excelência encontra expressão geralmente na qualidade dos melhores artigos de um cientista."

Tabela 2.3: Cálculo dos valores h e g de Small [Egg06]

TC	r	$\sum TC$	r^2
305	1	305	1
239	2	544	4
127	3	671	9
109	4	780	16
86	5	866	25
77	6	946	36
75	7	1023	49
67	8	1098	64
49	9	1165	81
44	10	1214	100
36	11	1258	121
26	12	1294	144
26	13	1320	169
25	14	1346	196
22	15	1371	225
22	16	1393	256
18	17	1415	289
18	**18**	1433	324
15	19	1451	361
12	20	1466	400
10	21	1478	441
9	22	1488	484
8	23	1497	529
8	24	1505	576
7	25	1513	625
6	26	1520	676
5	27	1526	729
5	28	1531	784
5	29	1536	841
3	30	1541	900
3	31	1544	961
2	32	1547	1024
2	33	1549	1089
2	34	1551	1156
1	35	1553	1225
1	36	1554	1296
1	37	1555	1369
1	38	1556	1444
1	**39**	**1557**	**1521**
1	40	1558	1600

Capítulo 3

Classificação de Blogues

Os blogues encontram-se para muitas pessoas entre os mais interessantes elementos produzidos na web. A partir de uma crescente necessidade de individualidade, de auto-expressão e de partilha de experiências, a taxa de criação de blogues tem vindo a aumentar a grande ritmo. Estes elementos permitem que informação seja rápida e facilmente publicada na web. Os blogues tornam possível que um utilizador comum, sem qualquer conhecimento em qualquer tecnologia específica, crie uma espécie de diário, onde poderá colocar fotografias da sua família, opiniões pessoais sobre assuntos do dia-a-dia e muitas mais possibilidades. Sendo a criação de um blogue, e publicação de novas entradas neste, tarefas muito fáceis, esta ferramenta é grandemente utilizada no presente, incentivando e alimentando a vontade do utilizador de se exprimir relativamente ao que faz e pensa. Desenvolvem-se gradualmente blogosferas, conjuntos de blogues e utilizadores que se concentram em redor de um ponto em comum, seja o tema dos blogues, seja a nacionalidade dos seus criadores.

Um elemento singular na blogosfera é também o splog, um blogue de spam, que possui conteúdo de publicidade gerado automaticamente, inundando a blogosfera com informação na sua generalidade inútil para o que o utilizador pretende. Estes blogues são frequentemente actualizados, e contêm um grande número de ligações. Desta forma estes sobem significativamente nas ordenações potenciando que um maior número de utilizadores os visitem, e melhorando assim a possibilidade de venda de produtos. Este género de spam é também conhecido por invadir outros tipos de páginas, como wikis e guestbooks.

Pela sua frequente e abrangente utilização, a área dos blogues é actualmente e tem sido, alvo de intensa investigação. Diversos aspectos relacionados com os blogues, com a blogosfera, e com a classificação destes segundo diversos critérios têm sido explorados. Apresentamos assim, neste capítulo, algum do trabalho já realizado nesta área.

3.1 Caracterizações da Blogosfera

Em 2001, Nardi et al. [NSGS04] publicaram um artigo de análise das motivações dos bloguistas, avaliando as razões pelas quais decidem criar um blogue e publicar neste. Consideraram uma pequena amostra de vinte e três pessoas em redor do campus de Stanford, pelo que claramente esta análise não poderá ser representativa das motivação dos bloguistas a nível global, mas que permitirá uma impressão generalista de algumas das razões pelas quais os bloguistas se dedicam a esta actividade. A amostra incluía 16 homens e 7 mulheres, com idades entre 19 e 60 anos. A maioria dos bloguistas conheceu este universo e começou a participar nesta tendência a partir da leitura de blogues já existentes. A frequência de colocação de novas entradas varia entre vários por dia até menos de um por mês. Segundo Nardi et al., muitos dos bloguistas entrevistados publicam opiniões e factos muito pessoais, dependendo assim a sua produtividade da própria facilidade e disponibilidade com que se exprimirem. Aparentam ter a perfeita noção de que escrevem num local público, tendo cuidado para não ferirem susceptibilidades dos seus conhecidos, não dizendo e revelando tudo o que pensam. Com base nas entrevistas realizadas, Nardi et al. destacam os seguintes motivos apontados por bloguistas:

- Documentação da sua vida, em que os bloguistas vão relatando os acontecimentos do dia-a-dia, como férias ou episódios caricatos que lhes aconteçam.
- Expressão de pontos de vista, em que os bloguistas encaram as suas páginas como o exercício de uma democracia de opinião, comentando assuntos que consideram pertinentes.
- Utilização do blogue como um escape e local de reflexão, em que os utilizadores procuram desabafar e exprimir o que pensam, funcionando este como um diário.
- Estruturação de raciocínios e opiniões. Alguns bloguistas referiram escrever em blogues como forma de exercitar a sua capacidade de expressão e de racionalizar questões e matérias presentes, por exemplo, nos *media*. Afirmam que sentindo-se obrigados a escrever no blogue, acabam por se auto-motivar a pensar e formular as suas opiniões relativamente à mais variada gama de assuntos.
- Utilização dos blogues como um meio de comunicação entre comunidades. No caso de um dos bloguistas entrevistados, o seu blogue permitia-lhe manter-se em contacto, partilhar trabalhos e experiências com a sua comunidade de poesia. Alguns dos entrevistados mencionam o facto dos blogues privilegiarem a comunicação entre grupos de utilizadores com os mesmos interesses.

Em Julho de 2006, Hurst [Hur06] apresentou um documento de análise de uma monitorização de um servidor de *pings*, o Weblogs.com. De cada vez que um blogue ou um

feed eram actualizados, o servidor de *pings* registava esta modificação, permitindo uma análise da actividade da blogosfera envolvida com base nos dados do servidor. Este criava um ficheiro HTML por hora com as actualizações realizadas, permitindo uma organização e leitura simples. Hurst analisou a actividade registada no servidor ao longo de 24 horas, podendo assim tirar conclusões relativamente ao comportamento de bloguistas e criar hipóteses que permitissem funcionar como ponto de partida para outras investigações.

Contabilizando o número de URLs por número de *pings*, Hurst elaborou um gráfico ilustrando esta relação. Esperava já uma curva decrescente, encontrando no entanto uma cauda inferior irregular, bem como um pico no número de URLs com cerca de 24 *pings*. O cientista enunciou que esta distribuição algo irregular se deveria provavelmente à presença de splogs, blogues spam, e retirando uma amostra aleatória de 100 URLs no referido pico, observou que 54% destes referenciavam splogs, confirmando a sua hipótese. Seguidamente, Hurst efectuou uma análise de perfil de utilizadores através da extracção de campos presentes nos perfis dos bloguistas.

Hurst comparou em comportamento e características os utilizadores dos dois domínios mais comuns, o Spaces e o Blogspot. Verificou que: enquanto a maioria das informações de utilizadores no Spaces eram válidas, a maioria das do Blogspot não o eram; e que a idade média dos bloguistas no Spaces é menor do que no Blogspot. Finalmente, analisando as nacionalidades dos utilizadores, Hurst observou certas variações nas características das diferentes blogosferas, como o facto da idade média dos bloguistas chineses ser menor do que a dos norte-americanos, ou as diferentes lideranças em popularidade de domínios, consoante o país em análise.

Analisando o servidor Weblogs, o mesmo autor, através dos registos de *pings* que equivalem a notificações no servidor, avaliou a actividade da blogosfera. Desta forma, Hurst conseguiu obter conclusões relativas à faixa etária de utilizadores, por nacionalidade, os seus hábitos diários como bloguistas, funcionando como um óptimo exemplo de informação passível de ser retirada da análise de uma blogosfera.

Em resposta à necessidade de uma colecção de blogues para investigação, foi criada em 2006 por MacDonald e Ounis a "TREC Blogs 06 Collection"[MO06], após os lançamento de um artigo em que foi feita a caracterização da colecção que obtiveram, e a análise de algumas das mais relevantes características do grupo de blogues. Através do uso de um *crawler*, responsável por extrair o maior número de *feeds* possível, obtiveram uma colecção de 100.000 blogues, com 3 milhões de entradas. Decidiram incluir três tipos diferentes de blogues, de forma a que a *TREC Collection* fosse representativa da blogosfera global. Assim, procuraram blogues populares (70% da colecção), blogues comuns (18%), bem como splogs (12%). A análise feita pelos dois autores é feita principalmente do ponto de vista temporal, tendo verificado que o número de novas entradas diminui consideravelmente no Natal e aos fins-de-semana. O pico inferior na criação de entradas durante o dia revelou-se à volta das 3 horas da manhã, enquanto a hora de maior

actividade localizava-se em redor das 13 horas. MacDonald e Ounis alertaram no entanto [MO06], para o facto dos dados de alguns dos blogues, relativos às suas datas de criação aparentarem não serem de confiança. Alguns blogues apresentavam datas impossíveis para a sua criação, como anteriores a 1998, e outros não continham esta informação de todo, o que poderá ter contribuído para alguma inconsistência em análises temporais da colecção.

Em 2007, Qazvinian et al. apresentaram uma análise da primeira colecção de blogues persas [QRSA07]. Através do uso de um *parser* de HTML, e seguidamente de um conversor para XML, obtiveram uma colecção de 80.000 XML's, incluindo blogues, entradas e os comentários para cada entrada. A colecção compilada continha cerca de 22.000 blogues, 349.000 entradas e cerca de 1.200.000 comentários, proporcionando uma base para a análise. Este estudo coincidiu, de acordo com alguns bloguistas profissionais, com uma queda na actividade do *host* de onde extraíram os dados, o PersianBlog, e esta informação é verificável nos gráficos representativos da actividade distribuída pelo ano. Além disso, foi concluído que a maioria dos comentários eram feitos por utilizadores que eram também bloguistas da mesma blogosfera, e que o seu número se distribuía de forma bastante homogénea pela semana com a excepção das quinta-feiras e sexta-feiras, em que se verificava uma diminuição deste número, bem como nas épocas de Ano Novo e começo de anos académicos. Na época das eleições presidenciais verificaram um aumento significativo na actividade nestes blogues, representando o interesse dos utilizadores neste tema. É de salientar que a inclusão de comentários na colecção de Qazvinian et al. foi uma opção pouco usual em estudos do género, tornando as análises obtidas em redor da produção destes elementos bastante interessante na área de caracterização de colecções.

No que a splogs diz respeito, foram já publicados vários documentos relatando investigações neste campo. Um dos mais interessante foi o apresentado por Kolari et all, em 2006, no artigo "Characterizing the Splogosphere"[KJF06] . Kolari et al. utilizaram a colecção fornecida pelo BlogPulse, um sistema de busca de blogues, que continha a informação de 21 dias em Julho de 2005. Implementando um modelo de aprendizagem de identificação de splogs na colecção, os cientistas obtiveram resultados bastante satisfatórios, com 90% de sucesso na detecção de blogues de spam na colecção.

Relativamente à frequência de palavras, Kolari et al. notaram haver uma grande diferença entre as palavras mais usadas em blogues e em splogues. Palavras como "me", "we"e "my"são raramente utilizadas em splogs, enquanto estes utilizam vocabulário bastante específico de publicidade. Este facto pode proporcionar facilidades na detecção e eliminação de splogues, através da detecção deste vocabulário específico. Segundo Kolari et al.:

> "Modelos de palavras de blogues baseados em características locais criam um tipo de blog interessante, que os separam dos blogues".

Outro aspecto comum na maioria dos splogs é o facto destes classificarem as suas *tags* como sem categoria. Porém, a principal ênfase no documento de Kolari et al. é dado à análise de um servidor de *pings*, verificando o comportamento dos splogs distribuído ao longo do tempo. Os cientistas realçam que à excepção de restrições à frequência de *pings*, mais nenhuma restrição é imposta por estes servidores, tornando-os permeáveis à presença de blogues de spam. Observando a distribuição de *pings* de blogues italianos, Kolari et al. verificaram que a sua distribuição ao longo do dia era bastante realista, com picos de actividade à hora do almoço, e o mínimo cerca das 5 horas da madrugada. Uma distribuição bastante heterogénea como esta indica-nos, segundo os investigadores, que os blogues em línguas que não a inglesa são menos propícios à presença de spam. Ao contrário de no caso dos blogues em italiano, os que são escritos em inglês apresentam uma distribuição muito mais homogénea e com menos picos. A explicação foi obtida através da observação de um gráfico mostrando a distribuição de *pings* de splogs num dia, em que se observava um padrão quase igual para todas as horas. A título de conclusão, Kolari et al. apresentam as duas seguintes conclusões:

- Os splogs constituem 88% de todos os URLs emissores, constituindo no entanto apenas 75% de todos os *pings*. Esta observação pode ser explicada pelo facto de que muitos *pings* de splogs serem enviados com informação de URL de blogues que nada têm em comum com a entidade emissora do *ping*;
- Muitos dos URLs registados são de blogues fictícios. Os *pings* enviados desta forma são chamados de *pings* zombie.

3.2 Classificações de Blogues

No ano de 1998, Page, L. publicou um artigo [Pag98] em que descreveu um método de classificação de páginas web, chamado PageRank, de forma a poder qualificá-las com um valor expressivo de importância. Demonstrou também como aplicar este critério a procuras e a navegações de utilizador. Page começou por apontar as razões fundamentais pelas quais o mundo da web (mas cujas características apontadas se mantêm ainda hoje, dez anos depois) necessitava de um método de avaliação que procurasse medir a qualidade de uma página. As páginas web eram então criadas em grande número sem qualquer controlo de qualidade ou custos de publicação, pelo que facilmente se criava um grupo de falsas páginas só com o intuito de, por exemplo, citarem uma determinada página nuclear com conteúdo comercial, de forma a aumentarem o seu valor numa ordenação por citações. Desta forma, um critério que contasse valores replicáveis de páginas web era facilmente manipuláveis.

Assim, o PageRank foi proposto, com a particularidade de se basear no grafo representativo das ligações na web. De acordo com Page, existiam em 1998 à volta de

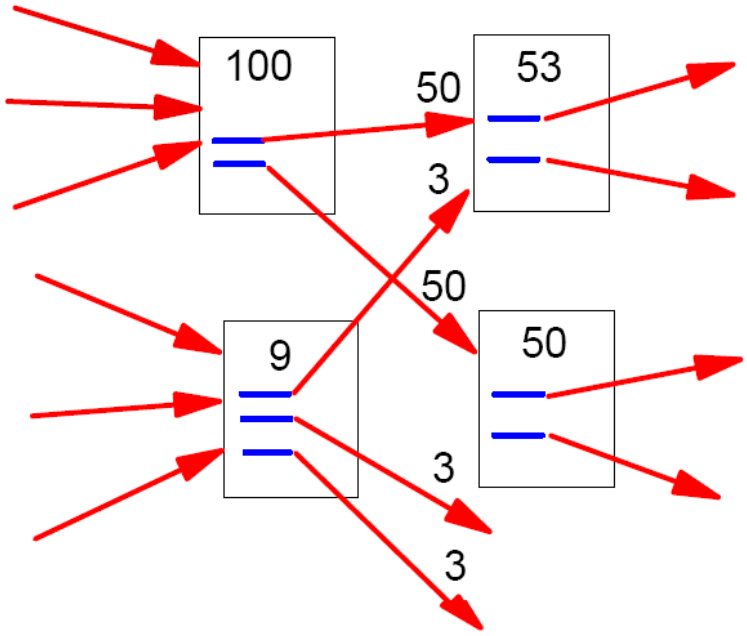

Figura 3.1: Representação do cálculo do PageRank [Pag98]

150 milhões de páginas e 1.7 biliões de hiperligações, e definiu para cada página um número de ligações de saída (ligações dessa página para uma outra) e ligações de entrada (o significado oposto). O cientista mencionou que páginas teoricamente de maior qualidade ou impacto obtinham um número mais elevado de ligações, daí o seu algoritmo ter tomado como objectivo obter uma aproximação ou medida desta importância de uma página. Segundo Page, numa definição simplificada, uma página obtinha um valor alto numa ordenação se a soma dos valores na ordenação de páginas que para ela ligavam era alta. O cálculo do PageRank de uma página é obtida através da seguinte fórmula:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{3.1}$$

em que u representa uma página web. F_u define um conjunto de páginas u e B_u um conjunto de páginas que apontam para u. Sendo $Nu = |Fu|$ o número de ligações de u e c um factor de normalização para o valor de PageRank, o cálculo da fórmula 3.1 obterá a posição numa classificação para uma página a avaliar. No esquema da figura 3.1 é feita uma demonstração do cálculo do valor de Page.

Page descreveu o algoritmo de cálculo do valor por si inventado como o de um "surfista aleatório", uma entidade abstracta que clicasse em hiperligação após hiperligação, navegando assim pelo grafo, e eventualmente visitando mais frequentemente páginas mais populares. Adicionando um factor E, com a finalidade de forçar o "surfista"a sair de ciclos, enveredando por novas hipóteses, Page resolveu facilmente o problema previamente

apontado.

No seu estudo, Page obteve uma colecção representativa da internet de 1998, ao analisar um repositório de 24 milhões de páginas web. O cientista testou o seu método de avaliação em dois motores de pesquisa: um simples motor de pesquisa de títulos de páginas, e no recém-nascido Google, criado por si e por Brin, S. A aplicação num motor de pesquisa por títulos mostrou resultados claramente superiores com o uso do PageRank. Enquanto uma pesquisa simples retornaria meramente as páginas com títulos contendo as palavras procuradas, ao ordenar previamente os resultados pelo seu valor de PageRank, os resultados foram muito mais relevantes e satisfatórios. Ao implementar o método de Page, foi possibilitado um número elevado de inovações. Para aumento de desempenho em pesquisas, criou-se um grupo de páginas web de grande relevância, obtendo-se um grupo de resultados de qualidade para pesquisas que pudessem ser realizadas. Desta forma, é proporcionada a possibilidade de desenvolver o conjunto de páginas que mais provavelmente corresponda ao resultado esperado por um utilizador, quer por relevância, quer por importância destas na web. Outro resultado passível de se obter do PageRank, é a estimativa de tráfego, pelo número de hiperligações e densidade do grafo representativo destas. O impacto do estudo conduzido por Page foi em tudo inovador e audacioso, causando um efeito tremendo na investigação na área de recuperação e classificação de páginas web.

Quase uma década depois, em 2006, Kritikipoulos et al. [KSV06] apresentaram um dos primeiros documentos de investigação sobre classificação de blogues, uma versão modificada do método PageRank. O método apresentado, intitulado BlogRank, introduziu as adaptações necessárias para modelar o PageRank às características particulares de blogues. Os investigadores apresentaram as seguintes razões pelas quais a classificação de blogues apresentava características diferentes do caso de páginas web em geral:

- O número de ligações entre blogues é muito mais reduzido do que no caso de outros tipos de páginas;;
- Informação específica orientada a blogues é insuficientemente investigada.

Os testes a este método foram realizados sobre uma colecção amostra fornecida pela Nielsen Buzzmetrics. Foram feitos ensaios com utilizadores realizando pesquisas. O método de avaliação e ordenação dos resultados variava aleatoriamente, nunca estando o sujeito ciente dos critérios utilizados, sendo assim obtidos resultados imparciais. O método de avaliação, BlogRank, é calculado através da resolução da seguinte fórmula:

$$B(a) = (1-E) + E(FN(U_1 \rightarrow A) * B(U_1) + ...FN(U_n \rightarrow A) * B(U_n)) \qquad (3.2)$$

Em que $B(A)$ é o valor de BlogRank do blog A; $B(Ui)$ o valor de BlogRank do blogue Ui que liga ao blogue A; E um factor entre 0 e 1 com vista a normalizar o cálculo; e

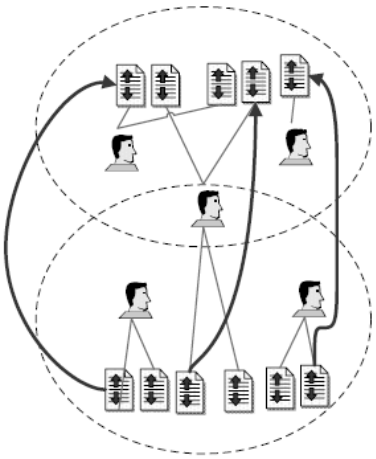

Figura 3.2: Ligações a nível de entradas [KSV06]

$FN(Un->A)$ representa a possibilidade de um utilizador visitante do blogue n visitar o blogue A.

De forma a solucionar o problema da fraca densidade no grafo de ligações da colecção, Kritikipoulos et al. definiram um conjunto de ligações internas entre vários blogues, procurando conjuntos de blogues que partilhassem características em comum, definindo assim ligações a nível de blogue. Blogues que partilhassem categorias e autores foram então agrupados, criando novas ligações e aumentando a densidade do grafo. Enquanto geralmente as ligações na blogosfera são feitas a nível de entradas, em que estas representam as contribuições de utilizadores e hiperligações de entradas para outras (Figura 3.2), os cientistas desenvolveram o conceito de ligações a nível de blogues (Figura 3.3).

Ao criar ligações implícitas ao grafo, Kritikopoulos et al. anularam parcialmente o conceito de "surfista"do PageRank. Neste caso, foi considerada a probabilidade de um utilizador viajar de uma página para a outra, mesmo não existindo ligações entre estas, mas sim um tema ou autor em comum. Para fins de teste, os cientistas definiram três diferentes métodos para ordenação de blogues. O primeiro consistiu em ignorar as ligações implícitas criadas, baseando-se somente nas ligações tradicionais, correspondendo assim ao PageRank. O segundo método consistiu numa extensão do PageRank, atribuindo no entanto um peso baseado no número de ligações diferentes entre entradas, funcionando como uma solução intermédia entre o PageRank e o BlogRank. Finalmente, o terceiro e último método atribuiu peso a ligações implícitas, no caso de tags ou autores em comum, formulando um grafo mais denso, e criando o ambiente para a ordenação segundo o BlogRank. Kritikopoulos et al. realizaram também um número de testes com utilizadores reais, pedindo as estes que executassem pesquisas, e optassem pelos resultados mais relevantes, sem conhecimento dos critérios avaliados.

A equipa de investigadores analisou os resultados, e obervou que o BlogRank obteve

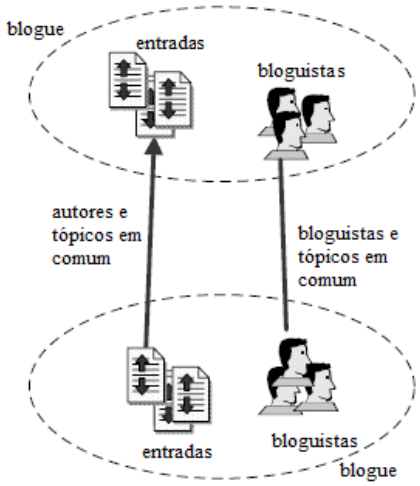

Figura 3.3: Ligações a nível de blogues [KSV06]

um desempenho claramente superior ao dos dois critérios restantes. O Blogrank apresentou um índice de sucesso médio para as pesquisas realizadas de 0,553, enquanto o PageRank e a combinação de ambos obtiveram valores de 0,158 e 0,353, respectivamente. Verificaram também que relativamente ao tempo tomado para indexar uma colecção de 1,5 milhões de blogues, o BlogRank demorou apenas mais 23% do tempo do que o PageRank, equivalente a 2 horas de diferença. Assim, Kritikopoulos et al concluíram que não só o seu método apresentava um desempenho superior ao PageRank, como era possível implementá-lo em motores de pesquisa sem que se exigisse demasiado tempo para indexação e memória, obtendo desta forma resultados francamente positivos para o método que propuseram.

Mishne, G., da Universidade de Amsterdão publicou em 2006 [Mis06] um artigo de análise de critérios que pudessem ter relevo para a apresentação de resultados de uma pesquisa numa colecção de blogues. O cientista avaliou o impacto de três critérios: relevância de tema, expressão de opinião, e qualidade da entrada. Quanto a relevância de tema, procedeu a recuperação de informação clássica, executando uma procura de termos equivalentes ou semelhantes (através de *stemming*). O autor teve também em consideração a componente temporal das entradas de blogues percorridas. Tendo em conta que muitas pesquisas na área de blogues têm como contexto a procura por temas actuais, é frequente o utilizador dar mais importância a artigos recentes do que forçosamente a artigos com mais ocorrência dos termos pelos quais procura. Assim, Mishne achou relevante ordenar os 100 primeiros resultados de uma pesquisa, pela sua data de entrada, com o intuito de fornecer os resultados mais relevantes ordenados por quão recentes as entradas fossem. Outro factor que o cientista implementou como critério para apresentação de resultados foi a expressão de opinião. É frequente o utilizador procurar entradas de blogue que ex-

primam opiniões, e não simplesmente uma notícia (disponível em outros meios). Assim, através do uso de métodos de análise léxica, com que analisou a colecção e procurou por palavras típicas de um discurso onde se exprime uma opinião, conseguiu definir e implementar este critério. Finalmente, a qualidade da entrada do blogue é também definida como medida de relevância para uma pesquisa. De forma a qualificar um blogue pela sua qualidade o cientista utilizou uma aproximação comum, a classificação por número de ligações de entrada, como no caso do PageRank. Mishne deu especial atenção à remoção de splogs, pois estes raramente ou nunca possuem conteúdo de relevo para uma pesquisa. Verificou que a maioria de blogues no domínio *blogspot.com* com um nome excedendo 35 caracteres eram splogs. Já no caso de blogues com domínios como o *livejournal.com* ou *typepad.com* raramente se tratava de splogs. Assunções como estas permitiram ao investigador detectar uma larga percentagem de splogs, criando um filtro para este tipo de blogue. Mishne juntou estes critérios através da associação das classificações segundo os vários critérios, com diferentes pesos. Experimentando diferentes valores aos pesos no cálculo de classificações finais para os resultados de pesquisas, chegou a resultados que considerou positivos. Concluiu que o critério de expressão de opinião contribuiu para uma franca melhoria de resultados, enquanto a avaliação baseada em contagem de citações diminuiu a eficácia. Num todo, Mishne classifica os resultados da sua associação de critérios como substancialmente melhorados, comparativamente a resultados anteriores. O seu artigo constituiu uma base sólida para uma prática que se tornou rapidamente comum, a de associação de vários critérios para a classificação e ordenação de blogues.

Capítulo 4

Caracterização da Colecção

Sendo criadas centenas de blogues diariamente, é extremamente difícil criar uma colecção que englobe e represente a blogosfera completa. Em Portugal, apesar dos primeiros blogues terem sido criados no fim dos anos 90, só agora começa a surgir uma rede de blogues passível de ser intra conectada e que permita a um *spider* eficiente recuperar toda a informação disponível. Nos últimos cinco anos, a SAPO[1] tem vindo a recolher *feeds* de blogues frequentemente, adicionando-os à sua colecção de procura de blogues. Foi esta colecção de *feeds*, contendo um total de 54.149 blogues com mais de 3 milhões de entradas, que nos foi cedida pelo fornecedor para fins de investigação. Apesar de esta colecção ter sido obtida com o intuito fundamental de investigações particulares, como a detecção de tendências e métodos de classificação de blogues, apresentamos neste capítulo uma caracterização simples da colecção. Note-se que lidámos com uma colecção de *feeds*, elaborada por um *crawler* ainda presentemente em funcionamento numa rede social em constante evolução, e que não poderá ser considerada representativa dos 500.000 blogues portugueses que se diz existirem.[2]. Esta caracterização foi realizada em conjunto com Pedro Pinto [Pin08], por partilharmos o interesse de utilizar a mesma colecção como base para experimentações.

4.1 Propriedades da Colecção

Decidimos remover à volta de 4 mil blogues dos 54 mil da colecção em bruto, por estarem fora do intervalo de tempo entre Janeiro de 2003 e Dezembro de 2007. O conjunto final a analisar continha então 49.940 blogues com 2.933.735 entradas distribuídas entre os anos 2003 e 2007. Apesar de estarem representados 50 mil blogues, relembramos que

[1] http://www.sapo.pt/

[2] Maria João Nogueira, http://jonasnuts.blogs.sapo.pt/2008/03/

a colecção usada foi construída recorrendo apenas aos seus *feeds*, podendo assim criar uma falsa impressão de representatividade. Cada blogue foi arquivado sendo as primeiras entradas as apresentadas no primeiro *feed* recuperado.

O critério usado para inserir um blogue na colecção foi simples: seguir todas as ligações de um blogue português, seguido de uma análise linguística nos blogues visados, de forma a confirmar se são ou não escritos em português.

Esta regra poderá explicar a discrepância entre o número de blogues na colecção (à volta de 5 mil) e o número anunciado pela SAPO (200 mil blogues em Março de 2008). Por um lado, uma grande parte do conjunto de blogues deste fornecedor poderá ser brasileira, não sendo assim adicionada à colecção de pesquisa por blogues. Por outro lado, pode também colocar-se a possibilidade de que a rede da blogosfera portuguesa não seja suficientemente densa para que o *spider* seja capaz de executar a pesquisa por *feeds* correctamente.

A colecção fornecida foi separada em três diferentes categorias no que diz respeito a fornecedores de espaços de blogues. Os dois maiores grupos obtidos foram os blogues hospedados pela SAPO, com 52% (25.768 entradas), e os hospedados pelo Blogspot [3], com 47% (ou 23.378 blogues). O restante 1% era composto por um conjunto de blogues de diferentes servidores, contendo um total de 794 blogues diferentes.

4.2 Análise da Colecção

4.2.1 Análise Social

Algumas extrapolações puderam ser feitas relativamente aos hábitos diários de um bloguista português. Os utilizadores portugueses aparentam gostar de escrever durante o dia, com curtas paragens durante o almoço e jantar. Os máximos puderam ser encontrados entre as 12 e as 16 horas enquanto a hora mais frequente para inserir novas entradas revelou-se ser à volta das 23 horas. A figura 4.1 mostra o número de entradas por hora discriminando os três grupos de fornecedores previamente mencionados. Notámos uma diferença de uma hora entre as horas de maior actividade nos blogues da SAPO e os do Blogspot. Mesmo reagindo de forma idêntica, os máximos e mínimos aparecem no SAPO uma hora antes que nos blogues do Blogspot. Esta diferença pode possivelmente ser explicada por diferentes fusos horários ou sistemas operativos diferentes utilizados pelos dois servidores.

A figura 4.2 demonstra os hábitos diários de um utilizador por mês nos últimos cinco anos. Sabendo que a colecção se encontra em constante crescimento, é natural que se observe um crescimento distribuído pelo ano. No entanto, pudémos ver atrasos ou mesmo quedas no crescimento nas férias do Verão e do Natal. O gráfico apresenta um crescimento

[3] http://www.blogger.com/

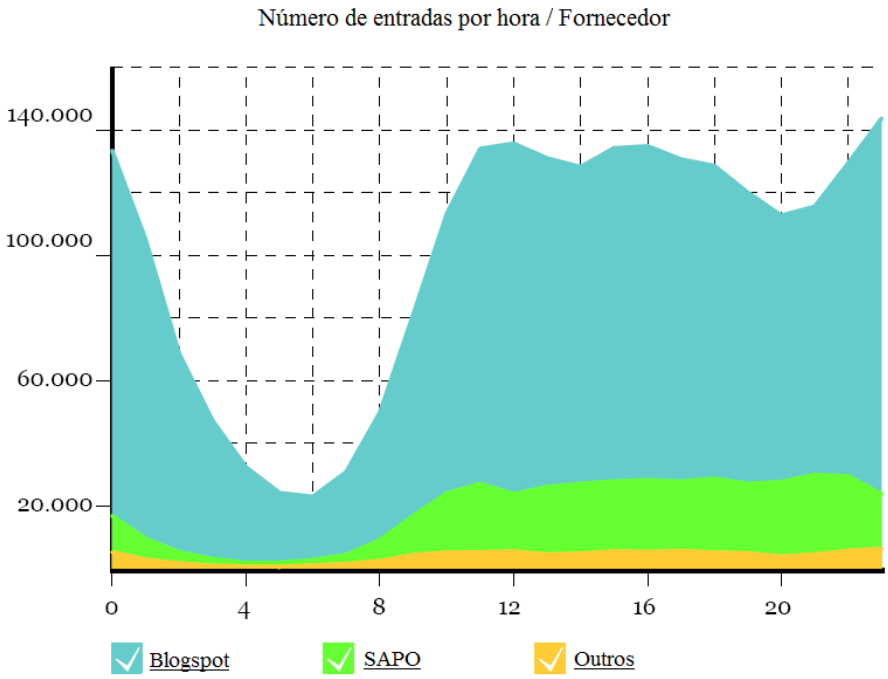

Figura 4.1: Entradas ao longo do dia, por fornecedores

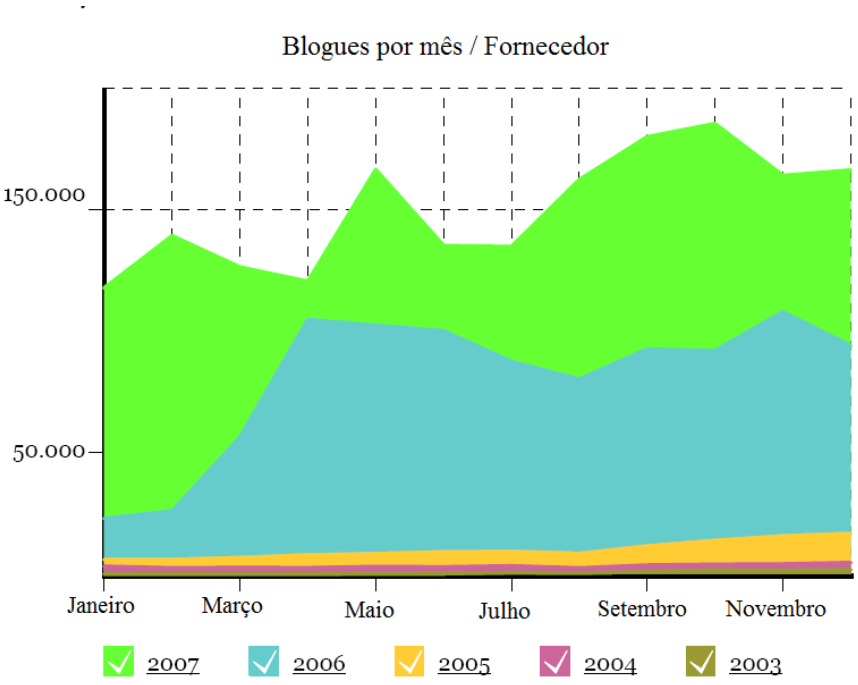

Figura 4.2: Novos blogues ao longo do ano, por fornecedores

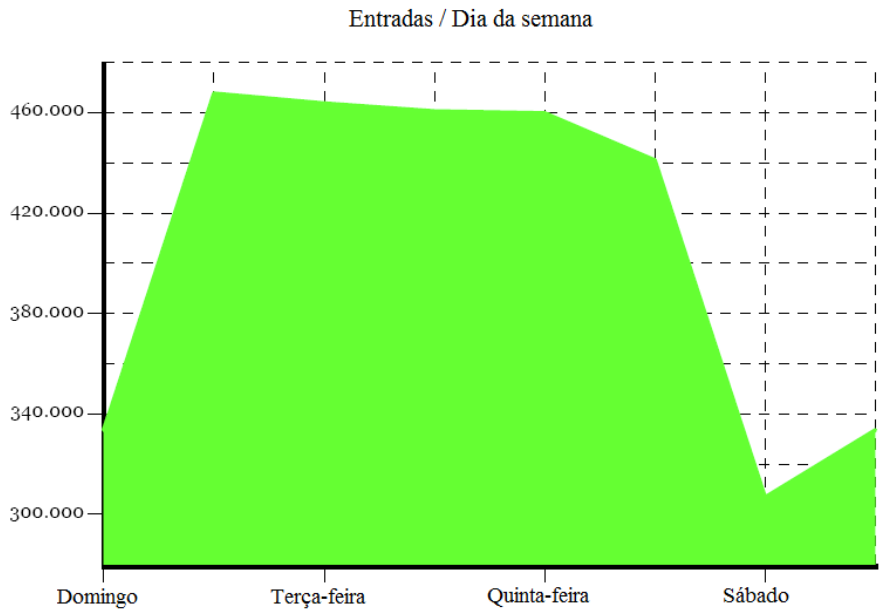

Figura 4.3: Novas entradas ao longo da semana

muito acentuado entre Fevereiro e Abril de 2006, representando-se um aumento de 300% na criação de novas entradas. Esta situação deverá ser estudada no futuro, visto poder estar ligada a algum algoritmo diferente utilizado no *crawl* diário, como a nossa próxima secção mostrará.

Finalmente, a figura 4.3 demonstra os hábitos de *blogging* durante a semana, mostrando um máximo de actividade às segunda-feiras, diminuindo lentamente até às sexta-feiras, vindo depois o fim-de-semana, em que esta actividade sofre uma queda abrupta.

4.2.2 Análise da Evolução

Realizando uma análise do tamanho da colecção ao longo do tempo, detectámos duas mudanças principais no crescimento do tamanho da colecção. Tanto em Abril de 2006 como em Fevereiro de 2007, o crescimento da colecção sofreu abrandamentos, possivelmente devido a alguma mudança no pesquisador de *feeds*. Pode-se observar este comportamento nas figuras 4.5 e 4.4. Aparentemente, a partir do aumento de blogues hospedados pelo SAPO, o *crawler* parou de registar novos blogues criados no Blogspot.

Nas figuras 4.6 e 4.7 é mostrado o número de entradas para cada dia ao longo dos cinco anos de *crawling*. Em Março de 2007 observou-se uma enorme queda no número de entradas em blogues do Blogspot, enquanto o número de entradas no SAPO aumenta o seu crescimento. O mais relevante pico no crescimento de criação de entradas de blogues acontece em Abril de 2006.

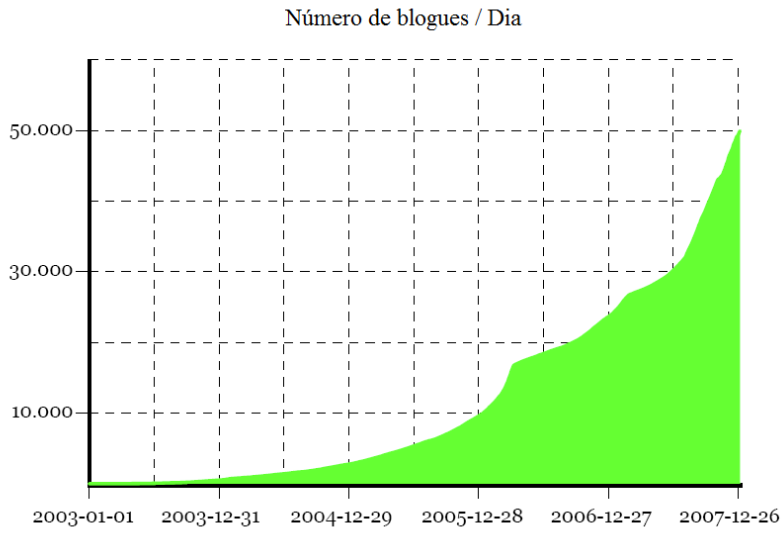

Figura 4.4: Blogues criados ao longo do tempo

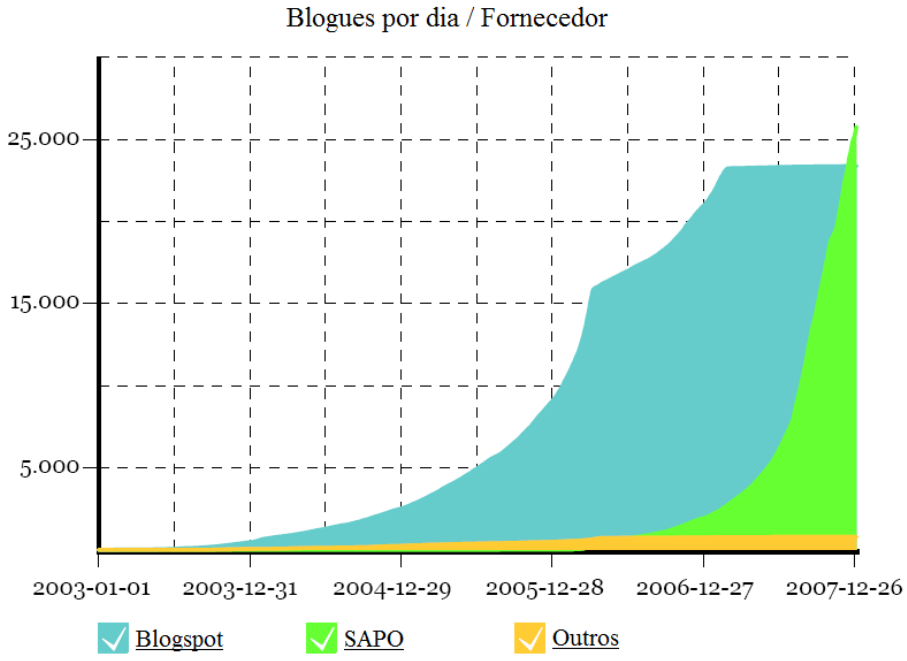

Figura 4.5: Blogues criados, por fornecedor

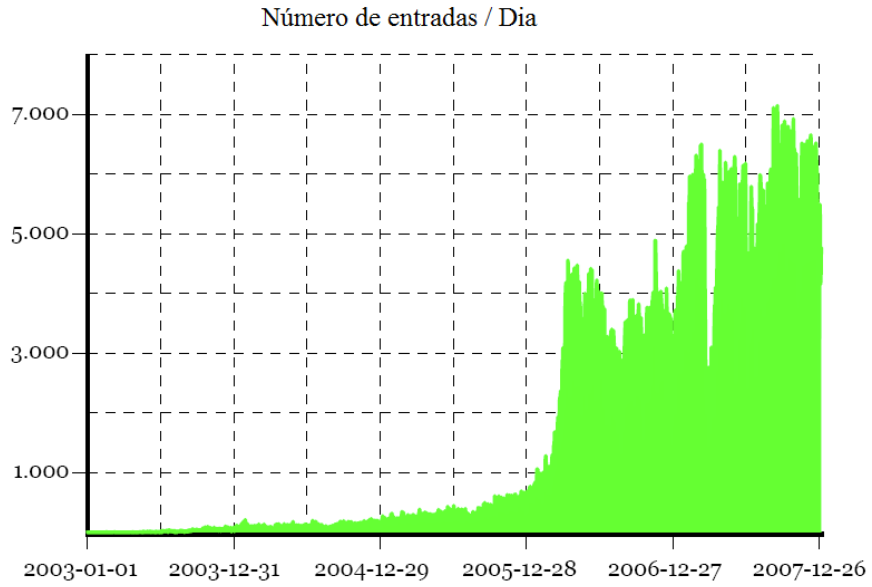

Figura 4.6: Entradas criadas ao longo do tempo

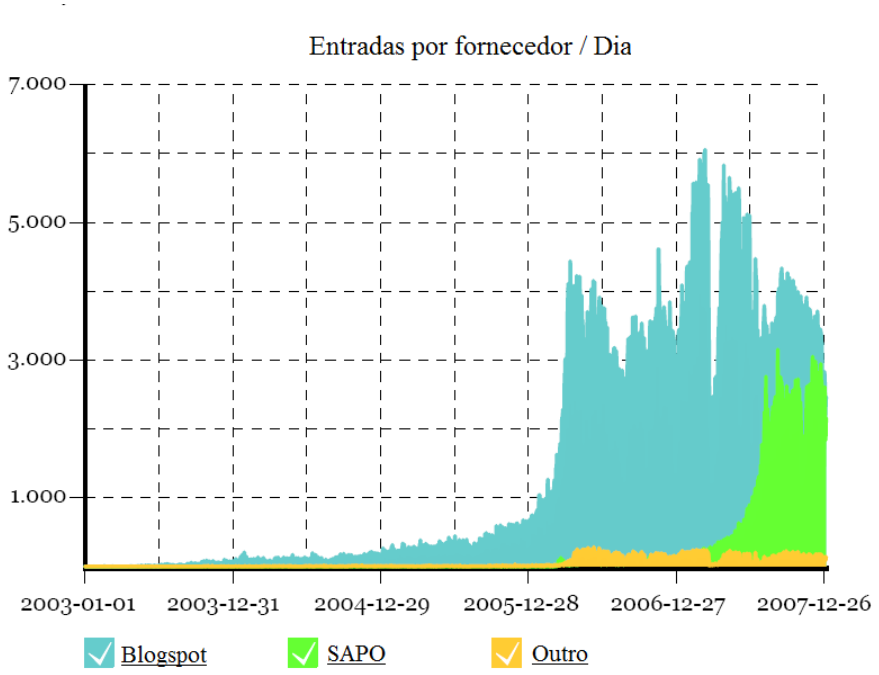

Figura 4.7: Entradas criadas, por fornecedor

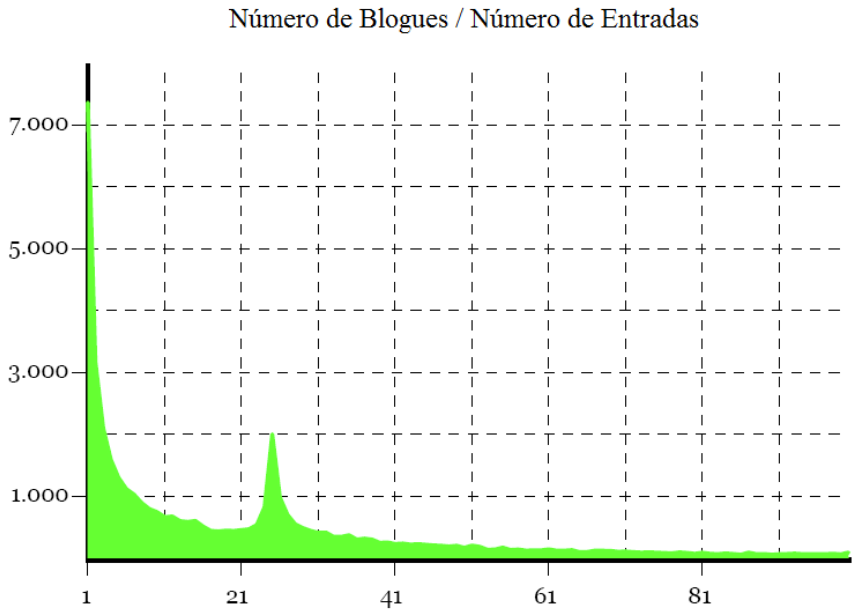

Figura 4.8: Distribuição de número de entradas por blogues

Um dos mais curiosos resultados da nossa análise foi o representado na figura 4.8. Apesar da longa cauda não ser surpreendente nesta análise de número de entradas por número de blogues, o máximo local nas 25 entradas sugeriu-nos que alguma anomalia se verificou naqueles *feeds*. Esta questão deve ainda ser explorada no futuro, visto que a recuperação dos blogues com aquelas 2 mil entradas se revelou inconclusiva. Uma possível explicação seria que o *crawler* passasse só uma vez num blogue, num longo espaço de tempo, e que parasse depois de recuperar mais blogues ao longo do tempo.

No que diz respeito ao conteúdo das entradas, foram feitas algumas análises face aos links e a utilização de palavras, mesmo que a um nível relativamente superficial. Observando as figuras 4.9 e 4.10, poderemos afirmar que os bloguistas portugueses ainda não utilizam muito frequentemente ligações para outros blogues, e que estes utilizam mais ligações para si próprios, seguidos só depois por ligações de saída. No entanto, tanto ligações de saída como as de entrada têm vindo a descrever um crescimento sólido que pode conduzir a uma comunidade mais forte num futuro próximo.

No que diz respeito à utilização de palavras, as entradas são geralmente curtas: metade da colecção é constituída por entradas com um número de palavras entre 0 e 75. No entanto, o número médio de palavras é ligeiramente maior, à volta de 160 palavras por entrada. Esta informação é observável nos gráficos 4.11 e 4.12.

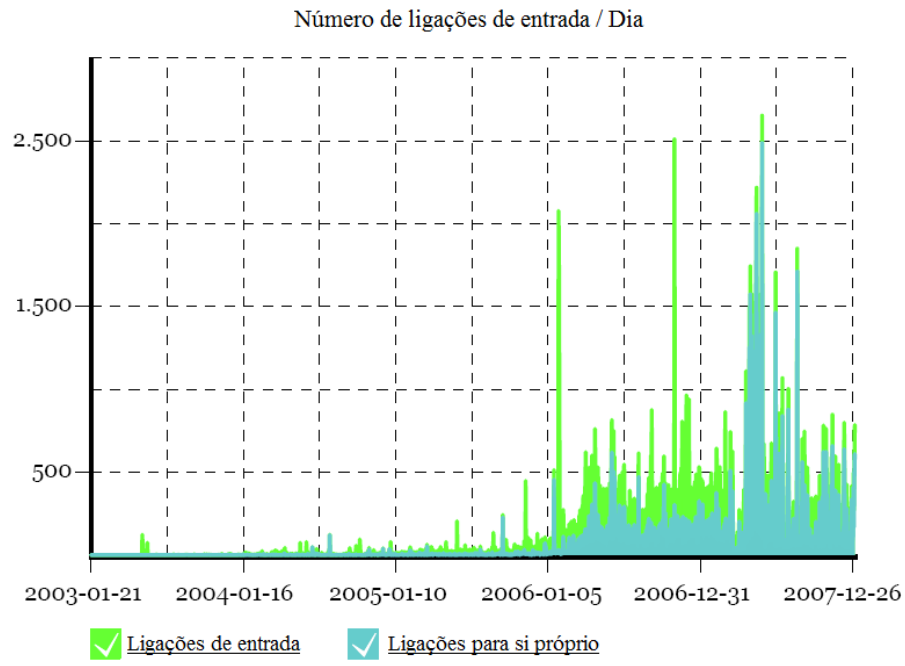

Figura 4.9: Número de ligações de entrada

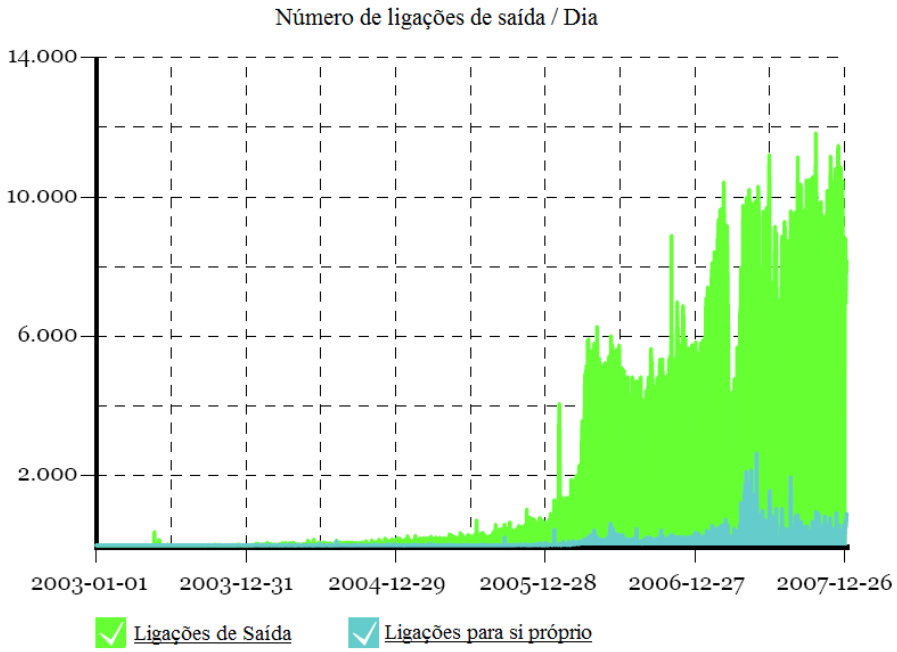

Figura 4.10: Número de ligações de saída

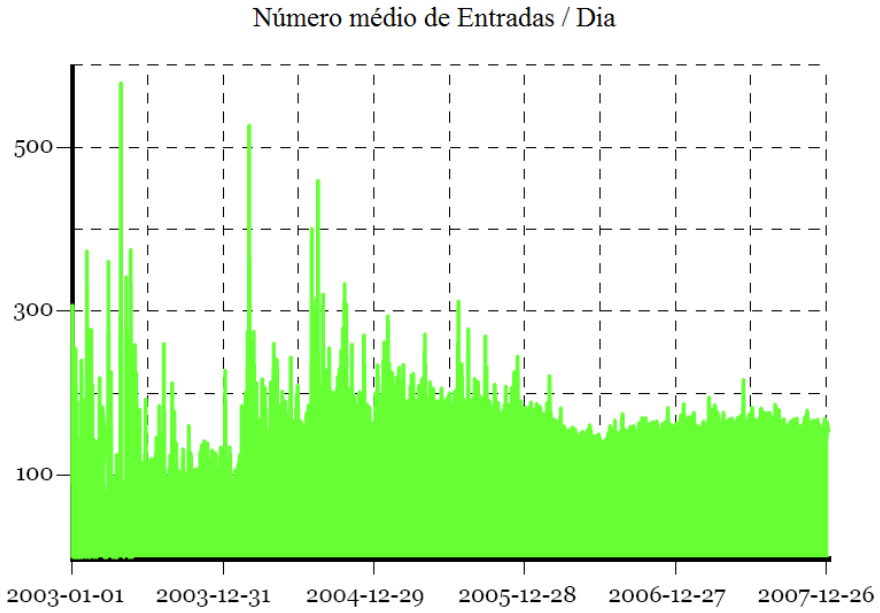

Figura 4.11: Número de entradas criadas por dia

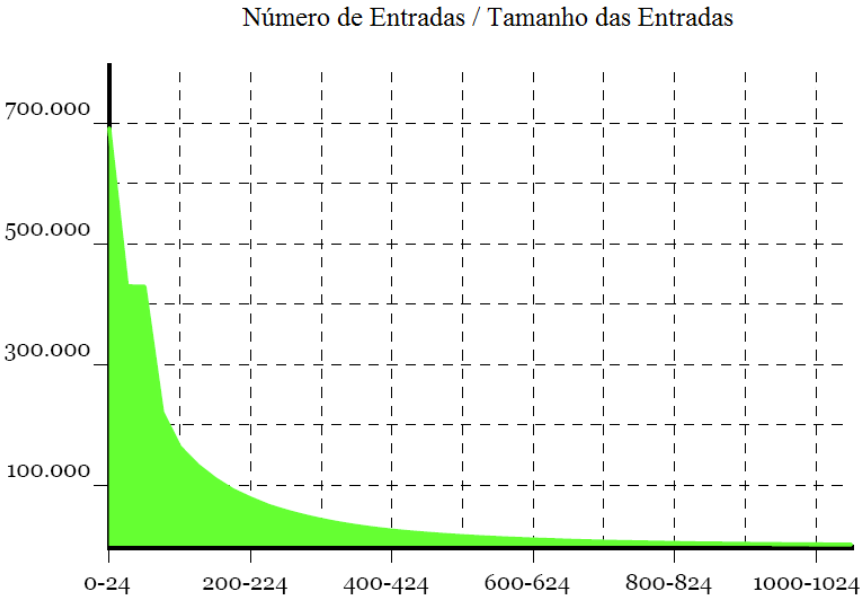

Figura 4.12: Número de entradas criadas, por tamanho

Capítulo 5

Aplicação do h-index em Blogues

Neste capítulo é descrita a implementação do h-index como método de classificação de blogues. Esta implementação permitiu elaborar uma ordenação dos blogues presentes na colecção que nos foi fornecida, de forma a avaliarmos o impacto que o índice pode causar como critério de ordenações. É feita uma análise dos resultados obtidos, fundamentalmente das diferentes ordenações que foram calculadas. Foram obtidos os valores do h-index, g-index e a contagem de links apontados (que serão chamados de citações daqui para a frente, por razões de clareza) para cada blogue, incluindo para cada um destes a presença ou ausência de *selfinlinks* (auto-citações). Desta forma estabelecemos a comparação entre seis ordenações, observando a distribuição destas e aplicando o algoritmo de Kendall [Ken48] para a avaliação da correlação entre os diferentes resultados obtidos.

5.1 Adaptação e Implementação do h-index

Nesta secção descrevemos o processo de adaptação do h-index ao contexto dos blogues, bem como a implementação do índice como valor de medição de importância de blogues.

5.1.1 Modelação

Quando nos foi proposta a ideia de aplicar a medida h-index à classificação de blogues, tomaram-se como objectivos a análise do resultado desta implementação e a observação da existência, ou não, da viabilidade deste valor como critério para avaliação de importância destes sites. O paralelismo estabelecido entre o contexto habitual para a utilização do algoritmo de Hirsch e o contexto da blogosfera foi primeiramente analisado. Tornar-se-ia para nós clara a relação entre os dois ambientes: um blogue, na sua totalidade de

entradas, tomaria o mesmo papel que um cientista, no contexto habitual, tendo como artigos científicos as suas entradas. Hiperligações a apontar para o blogue em questão, tornar-se-iam as nossas citações, do ponto de vista de documentação científica, bastando-nos por isso, contabilizar o número de vezes em que a página principal do blogue ou uma entrada em específico é referenciada noutros blogues da nossa colecção.

No que a ligações (a nossa versão de citações), diz respeito, colocaram-se-nos duas questões fundamentais:

- Deveríamos considerar auto-citações, ou seja, situações em que o bloguista cita uma entrada sua, ou a própria página principal do blogue?

- Existe um número suficiente de citações entre blogues que justifique a aplicação do método em causa? Se este número fosse muito reduzido, provavelmente o impacto do algorimo na qualificação de blogues seria irrelevante, pois só uma pequena parte destes possuiría um valor *h* atribuído.

Hirsch indica que qualquer algoritmo de avaliação baseado em citações deverá evitar a presença de auto-citações [Hir05]. Apesar de no caso do h-index o efeito destas ser muito mais reduzido do que no caso de apenas uma contagem de citações, o cientista sugere que de facto estas sejam removidas, de forma a aumentar a credibilidade dos resultados. No contexto da blogosfera, pareceu-nos ainda mais apropriado que se filtrassem as auto-citações, de forma a evitar manipulações desta classificação, e assim o fizemos. No entanto, apresentaremos mesmo assim na secção seguinte comparações entre resultados obtidos com e sem auto-citações, de forma a constatar o impacto no h-index neste caso.

5.1.2 Extracção de informação

O grupo de dados com que viríamos a trabalhar foi obtida, como já mencionámos, a partir da colecção cedida pela SAPO. De forma a tornar possível a realização dos testes referidos no capítulo seguinte, recorremos a um formato em que pudéssemos conter a informação necessária, nomeadamente o conteúdo de cada entrada, de forma a se poderem realizar buscas pelo conteúdo dos blogues. Para a fácil obtenção da informação em cada uma das entradas, optámos pelo formato XML, que nos pareceu apropriado para recorrer ao motor de pesquisa que viríamos a utilizar. Assim, foi extraída através do uso de Perl a colecção de entradas desejada, do ano de 2003 a 2007, com a informação que nos poderia ser relevante, como o URL da entrada, a data de publicação, o autor e o texto do corpo. Obtivemos desta forma um conjunto de ficheiros XML, um para cada entrada, com o formato mostrado na figura 5.1.

No final da extracção havíamos obtido um total de quase 3 milhões de ficheiros XML, o número correspondente ao número de entradas na colecção.

```
<entry id = "Código único" basename = "URL do blogue" url = "URL da entrada"
<author> Nome do autor </author>
<title>Titulo da entrada</title>
<postdate> Data e hora de publicação </postdate>
<body> Texto principal. </body>
</entry>
```

Figura 5.1: Formato dos ficheiros XML da colecção

Para o cálculo do valor de h-index de um blogue, necessitamos de um dado somente, o grupo de entradas suas que são citadas em outros blogues. De forma a obter estes dados, executamos um *script*, também em Perl, que fazia um *parse* do conteúdo dos entradas da colecção. Quando este *parser* encontrasse uma hiperligação, verificaria se esta apontava para um blogue da própria colecção (ou o próprio blogue a ser analisado) e registaria, se fosse o caso. Fizemos desde logo a distinção entre auto-ligações e ligações de outros blogues, pelo que obtivemos para cada entrada com ligações, o número de ligações de entrada, auto-citações incluídas, bem como o número de ligações somente de outros blogues. Assim, obtivemos um ficheiro contendo o número de ligações apontando para cada entrada, discriminado em dois valores, contendo ou não auto-ligações. Podemos ver a seguir um exemplo do formato do ficheiro final.

donaema.blogspot.com/2006/03/como-num-anuncio-de-tv.html 1 0

Note-se que para o cálculo de classificações baseadas em citações, não necessitamos de mais informação do que esta, pois somente os blogues com alguma citação nos interessaria, obtendo os outros, por omissão, uma posição no fundo da ordenação.

5.1.3 Aplicação do h-index

O algoritmo de cálculo do h-index tem a vantagem já elogiada de ser bastante simples de executar. Tomando o ficheiro com o registo de citações para cada entrada, desenvolvemos em JAVA um *script* que para cada um dos blogues procurasse as respectivas entradas, e as ordenasse pelo número de citações. Após esta ordenação, bastou verificar a posição na ordenação de entradas em que o número correspondente fosse igual ao número de citações para a entrada em questão, obtendo assim o valor de h a atribuir ao blogue. Desta forma, foi-nos possível obter o valor de h para cada um dos blogues com citações, ordenando-os seguidamente pelo seu índice. Viríamos ainda a preencher a listagem com todos os blogues restantes excluídos do ficheiro registo de ligações, pois estes não haviam sido processados para o cálculo do índice, sendo-lhes atribuído um valor h de zero. Para fins de comparação e análise dos resultados com ou sem auto-citações, realizámos o cálculo do valor h considerando o número total de ligações de entrada, bem como a

mesma operação deduzindo ao número total o valor atribuído a auto-citações, de forma a obter os dois resultados diferentes.

5.1.4 Outras implementações

Para estabelecer uma comparação com os resultados que viriam a ser obtidos através da ordenação por valores de h-index, decidimos construir outras duas ordenações: a variante de Egghe, o g-index, e a simples contagem por citações.

A contagem por citações é um dos critérios mais simples e frequentemente utilizados na ordenação de páginas, nomeadamente blogues. Os resultados obtidos são representativos pois, à partida, o número de ligações para um blogue reflectirá a sua popularidade ou impacto na blogosfera. Uma das desvantagens claras deste critério é o facto de poder ser facilmente manipulável, especialmente no caso de considerarem auto-citações, em que um autor pode optar por criar frequentemente ligações para o próprio blogue, causando um grande impacto na posição do seu blogue numa ordenação deste género. A ordenação por ligações de entrada foi rapidamente obtida a partir do ficheiro de registo destas ligações que havíamos extraído previamente, limitando-nos a ordenar o blogue pelo seu número respectivo. Realizámos, tal como no cálculo do h-index, a mesma tarefa tanto no caso de inclusão como de exclusão de auto-citações.

De forma a obtermos uma maior variedade de resultados, e a analisar a distribuição de valores em mais do que uma variante, optámos também por implementar uma ordenação segundo o algoritmo do g-index. Esta opção permitir-nos-ia estabelecer comparações entre os dois índices e analisar as vantagens e desvantagens de cada um dos índices. Relativamente à atribuição do valor g para cada blogue, é mantida em adição à contagem de citações já utilizada para o cálculo do valor h: um somatório do número de citações, bem como o valor do quadrado da posição de entrada na ordenação de entradas por número de citações. Desta forma o valor de g é calculado e atribuído a cada um dos blogues, e é feita a ordenação do conjunto pelos seus valores g.

5.2 Análise dos Resultados

De forma a realizar o cálculo do h-index, verificamos primeiramente a frequência de ligações entre blogues, avaliando assim a quantidade que viria a ser distinguida pelas medidas de ordenação que aplicámos. Observámos que de todas as entradas, 86.008 (3%) dos blogues são citados, e removendo auto-citações, restou-nos um total de 5.154 (0,2%). No que diz respeito a blogues, 39.626 (79%) foram citados, e no caso de remoção de auto-citações, restaram 2.824 blogues (5,7%) citados. Assim concluímos que o grafo de ligações entre blogues apresenta uma característica já antes mencionada [KSV06] existir entre blogues, ser pouco denso em número de ligações entre eles.

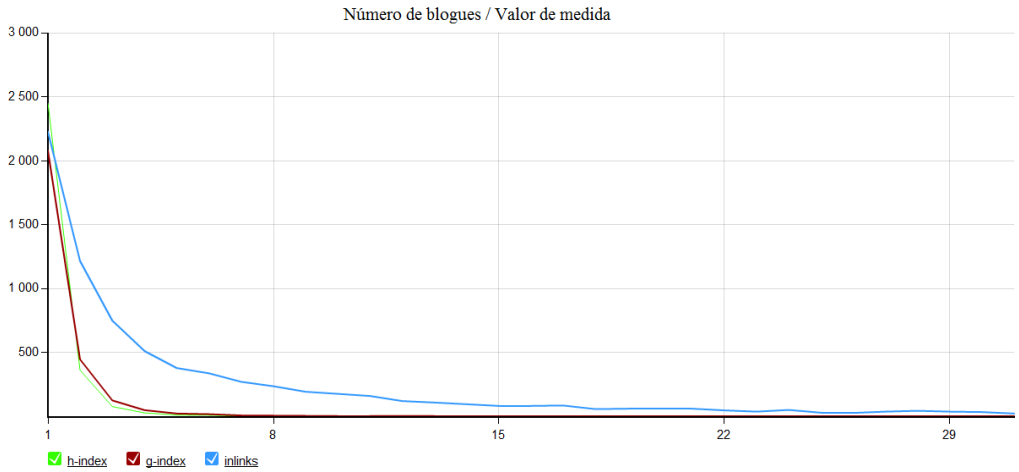

Figura 5.2: Distribuição de blogues segundo diferentes medidas

Depois de proceder ao cálculo para os valores de *h*, *g* e números de ligações de entrada, ordenámos a colecção de blogues por cada um destes três critérios. Nas tabelas 5.1, 5.2 e 5.3 expusémos os primeiros dez das ordenações segundo os três critérios. Os valores máximos obtidos segundo as medidas h-index, g-index, e ligações de entrada, foram respectivamente: 10, 15 e 4087. Os valores mínimos e médios para todas as medidas, tomaram o valor de 0, como esperávamos, dada a reduzida percentagem de blogues com citações. A distribuição dos valores para cada um dos critérios é apresentada no gráfico 5.2. As três medidas desenham um curva similar, com a particularidade do gráfico representando o número de ligações de entrada apresentar uma gama mais alargada de valores.

Um dos objectivos principais desta análise consistiu em comparar os resultados para cada um dos critérios. Assim, foi-nos permitido obter conclusões relativas à diversidade de resultados obtidos pelos dois índices, comparativamente aos obtidos segundo a medida mais tradicional, a da contagem de citações. Na tabela 5.1 verificámos que a ordenação pela medida do h-index apresenta valores bastante diferentes das realizadas com as medidas do g-index e por citações. Pelo maior peso que o g-index dá a blogues de topo, as maiores diferenças no topo da ordenação pertenceram a esta medida. Por outro lado, à medida que fomos alargando o alcance da comparação, o h-index aumentou a sua capacidade de obter resultados diferentes. Esta comparação permite-nos concluir que tanto a medida h-index como o g-index oferecem valores diferentes dos obtidos pelas outras classificações. No gráfico 5.3 demonstramos o grau de semelhança (de elementos contidos) entre diferentes critérios. Aparentemente é o h-index que obtém ordenações mais diferentes dos outros dois métodos, à excepção dos tops 10, 20 e 30.

Para uma análise mais profunda dos resultados obtidos, decidimos utilizar o coeficiente de correlação de ordenações de Kendall [Ken48], ou coeficiente de *tau*. Este coeficiente mede a semelhança entre duas ordenações, classificando-as quanto ao grau de

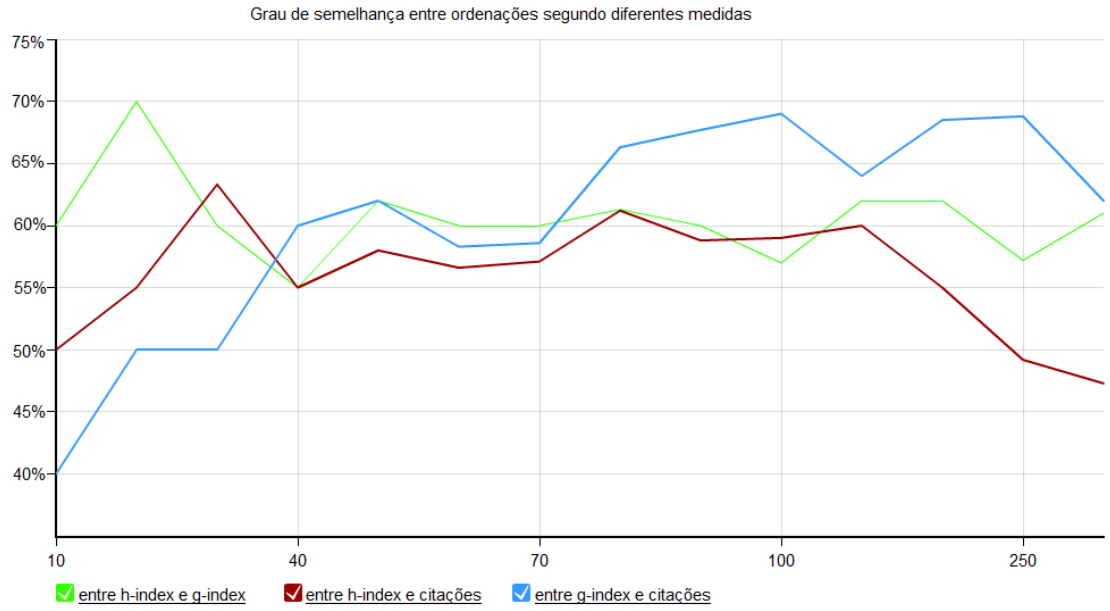

Figura 5.3: Graus de semelhança entre ordenações

parecença entre as duas. Apresenta valores entre 1 e -1, correspondendo o valor de 1 a duas ordenações idênticas, e -1 a ordenações 100% diferentes (ordem ou elementos diferentes). O coeficiente é calculado através da fórmula:

$$\tau = \frac{4P}{n(n-1)} - 1 \tag{5.1}$$

Em que n é o número de elementos das ordenações, e P o somatório do número de elementos classificados após cada elemento comparado entre as duas ordenações.

Na figura 5.4, observa-se a correlação entre as ordenações segundo o h-index e o g-index, nas suas primeiras mil posições. Pode-se verificar uma grande diferença entre as duas ordenações até aos primeiros 700 blogues, começando a partir desse número a tender para valores de maior semelhança. Esta tendência pode ser explicada com o facto de a partir de certo ponto (por volta da posição 3000), passamos a incluir o grupo de blogues que não possuem citações, pelo que terão valores de h e g de zero, ficando assim as ordenações idênticas.

Estabelecendo uma comparação entre as ordenações obtidas pelo h-index e pela medida de citações, obtivemos o gráfico da figura 5.5. Obtivemos também a mesma comparação avaliando a correlação entre o g-index e a medida de citações (Fig. 5.6). Numa comparação entre os dois resultados, verificámos que a ordenação segundo o h-index apresenta uma menor correlação com a ordenação segundo o número de citações. Concluímos assim que o h-index oferece uma grande disparidade de resultados comparativamente com

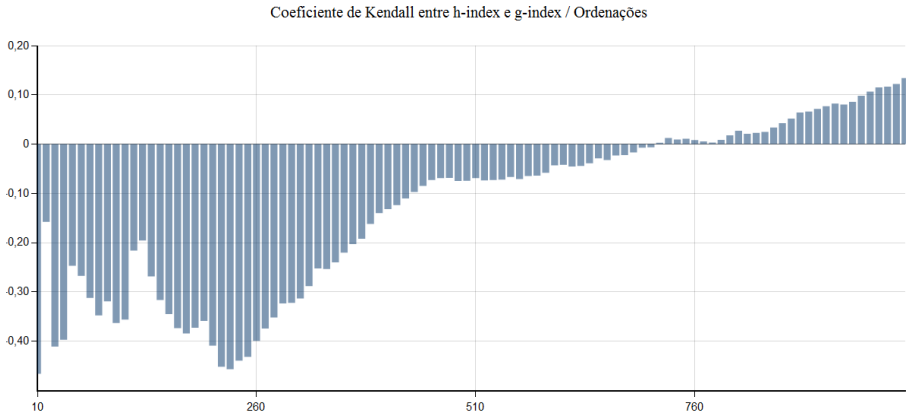

Figura 5.4: Comparação entre h-index e g-index

os resultados obtidos por uma ordenação segundo o g-index. Concluímos também que o h-index apresenta uma ordenação muito menos correlacionada com a de citações do que no caso do g-index.

Com o objectivo de avaliar o impacto da presença de auto-citações na classificação e ordenação de blogues segundo o h-index e g-index, apresentamos os gráficos das Figs. 5.7 e 5.8. Em cada um destes gráficos representamos a avaliação de correlação entre as ordenações obtidas para cada uma das duas medidas e as mesmas mas incluindo auto-citações.

Comparando os gráficos, verificamos que o método g-index é menos sensível à presença de auto-citações, por apresentar um coeficiente mais alto. No entanto, é de notar que nos primeiros dez da ordenação, foi o h-index a obter um coeficiente mais alto. Existia na colecção um pequeno conjunto de blogues com um número de auto-citações anormalmente alto, como é o caso do blogue *antologiadoesquecimento.blogspot.com*, com 36050

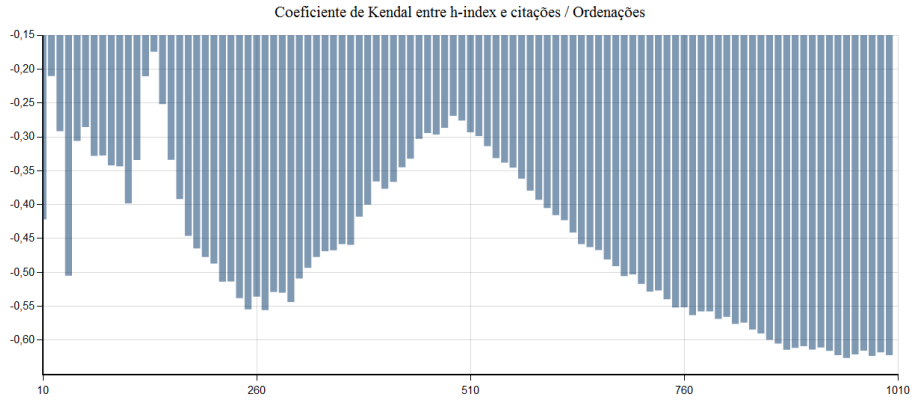

Figura 5.5: Comparação entre h-index e medição de citações

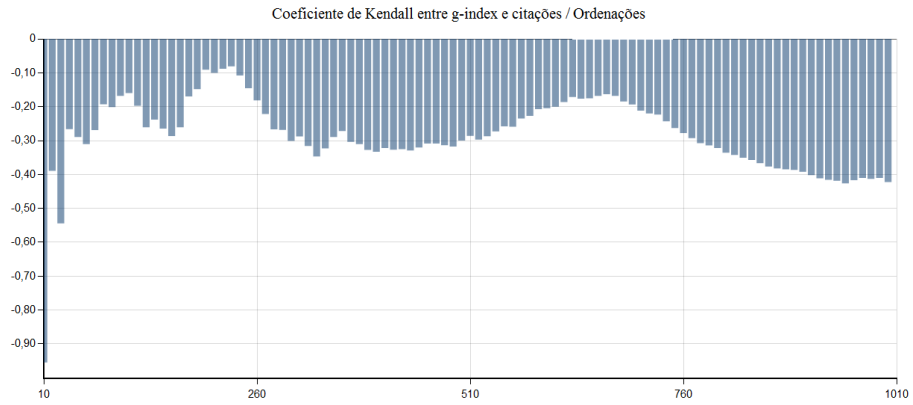

Figura 5.6: Comparação entre g-index e medição de citações

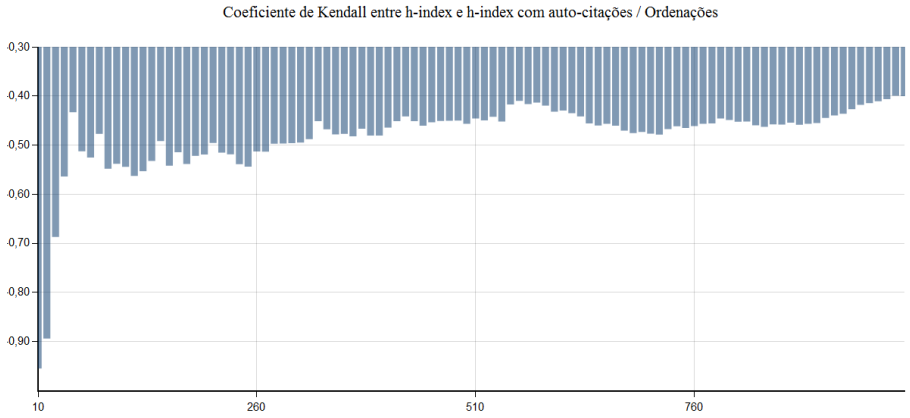

Figura 5.7: Comparação entre h-index e h-index com auto-citações

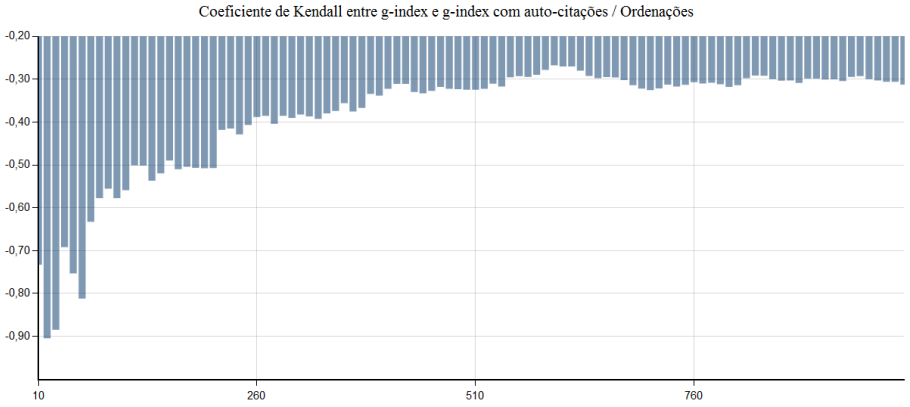

Figura 5.8: Comparação entre g-index e g-index com auto-citações

auto-citações. O facto de o h-index obter uma menor sensibilidade nos primeiros dez blogues, que conterão estes blogues altamente auto-citados, indica que o h-index não será tão vulnerável à presença de auto-citações como os dados anteriormente indicados pareciam indicar.

Tabela 5.1: Top10 de classificação segundo h-index

Pos.	Blogue
1	causa-nossa.blogspot.com
2	ablasfemia.blogspot.com
3	abrupto.blogspot.com
4	blogueforanada.blogspot.com
5	doportugalprofundo.blogspot.com
6	origemdasespecies.blogspot.com
7	geracao-rasca.blogspot.com
8	portugaldospequeninos.blogspot.com
9	aba-da-causa.blogspot.com
10	tugir.blogspot.com

Tabela 5.2: Top10 de classificação segundo g-index

Pos.	Blogue
1	gloriafacil.blogspot.com
2	abrupto.blogspot.com
3	causa-nossa.blogspot.com
4	ablasfemia.blogspot.com
5	doportugalprofundo.blogspot.com
6	geracao-rasca.blogspot.com
7	encarnados.blogspot.com
8	combustoes.blogspot.com
9	blogueforanada.blogspot.com
10	portugalcontemporaneo.blogspot.com

Tabela 5.3: Top10 de classificação segundo contagem de citações

Pos.	Blogue
1	ablasfemia.blogspot.com
2	abrupto.blogspot.com
3	causa-nossa.blogspot.com
4	daliteratura.blogspot.com
5	origemdasespecies.blogspot.com
6	elvirabistrot.blogspot.com
7	portugaldospequeninos.blogspot.com
8	corta-fitas.blogspot.com
9	bloguitica.blogspot.com
10	gloriafacil.blogspot.com

Capítulo 6

Resultados experimentais

No capítulo que se segue é descrita fase de testes às medidas analisadas, com utilizadores reais. Optámos por utilizar a ferramenta Terrier [1], uma plataforma de recuperação de informação tornada disponível pela Universidade de Glasgow para fins de pesquisa na área. Tomámos como objectivo analisar os resultados de decisões de utilizadores face a resultados para as suas pesquisas. Estes resultados foram apresentados segundo critérios variáveis, de forma a avaliar a performance dos diferentes métodos de classificação implementados.

6.1 Modelação para Terrier

A plataforma Terrier (Terabyte Retriever) foi desenvolvida por Ounis et al. em 2006 [OAP+05] para fins de investigação na área de recuperação de informação. O Terrier apresenta características como elevada capacidade de configuração, eficiência, desempenho, e presença de algoritmos actuais e reputados como os melhores para recuperação de informação. A versão *open-source* que utilizámos revelou-se ser flexível, fácil de implementar, e dotada de grande adaptabilidade a diversas possíveis situações em que um utilizador a quisesse utilizar. O Terrier permite ao utilizador construir um índice para uma colecção de grande dimensão, especificando um grande número de configurações existentes, para, da melhor forma, se adaptar a ferramenta às suas necessidades.

Indexação A indexação foi feita sobre a colecção que possuíamos, em formato XML. O Terrier compila os documentos dados como alvo para indexação em objectos do tipo Documento, num objecto do tipo Colecção, e seguidamente a indexação invertida é iniciada. Nesta etapa, especificámos dois requisitos que desejávamos modificar, uma lista

[1] http://ir.dcs.gla.ac.uk/terrier/

de *stopwords*, e indexação em bloco. Construímos a lista de *stopwords* a partir de uma compilação das palavras mais comuns na colecção, incluindo de novo as palavras que considerámos potencialmente de relevo para pesquisas, como por exemplo, "Portugal"ou "Natal". Desta forma palavras mais comuns na língua Portuguesa que achamos irrelevantes para pesquisas (por exemplo, "de", "a", "pois"), não foram consideradas na indexação, poupando tempo e espaço para a formação do índice. A nossa colecção apresentava uma grande dimensão (15 gigabytes), o que resultaria numa indexação muito lenta e custosa em memória. Assim, considerámos válido activar a função de indexação em bloco, que cria estruturas temporárias durante a compilação do índice, originando ficheiros também temporários, de forma a poupar memória para todo o processo. No fim da criação de todos os blocos de índice temporários, o indexador comprime todos os blocos num só índice final. Escolhemos para os blocos do índice uma dimensão de 1000 palavras, e no fim da indexação obtivemos um índice com cerca de 900Mbs de tamanho.

Pesquisa A plataforma Terrier, ao receber um termo de pesquisa, procura-o no índice criado e retorna uma lista de resultados correspondente ao conjunto de documentos em que o termo existe. A pesquisa pode ser realizada com um só termo, ou com uma combinação de vários, visto o Terrier possibilitar várias fórmulas de pesquisa, como por exemplo "+termo1 +termo2", se desejarmos resultados que contenham ambas as palavras.

A informação retornada contém a localização do documento aquando a sua indexação, o seu nome, tamanho, extensão, e pontuação de relevância atribuída. A pontuação de relevância é, no caso do Terrier, configurável, existindo uma larga gama de medidas de avaliação disponíveis para uso. Optámos por utilizar a medida $tf - idf$, uma das mais frequentemente utilizadas, e que cujo funcionamento melhor conhecíamos. O valor tf (term frequency) corresponde ao número de vezes que um termo surge no documento avaliado. Este número é normalizado, em função do tamanho do documento, de forma a que seja valorizado a proporção de vezes que o termo surge no texto comparativamente ao tamanho do próprio. Sendo n_{ij} o número de ocorrências do termo no documento d_j, e o denominador a soma de todas as ocorrências de todos os termos no documento, a normalização é feita segundo a seguinte fórmula:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{6.1}$$

O valor idf (inverse document frequency) mede o poder de discriminação de um termo relativamente à colecção indexada. Sendo $|D|$ o número de documentos na colecção, e $|dj : ti \in d_j|$ o número de documentos onde o termo surge, o valor é calculado através da fórmula:

$$idf_i = log\frac{|D|}{|d_j : t_i \in d_j|} \tag{6.2}$$

Finalmente, multiplicam-se os dois indicadores, obtendo a medida $tf - idf$ por nós utilizada.

$$tfidf_{i,j} = tf_{i,j} * idf_i \tag{6.3}$$

6.2 Resultados

Seguindo o exemplo de Kritikipoulos et al., [KSV06] optámos por testar os resultados de pesquisas com utilizadores reais, não revelando a estes quaisquer dos pesos e medidas utilizados para o cálculo dos resultados que lhe foram apresentados. Testámos uma amostra constituída por 16 pessoas, obtendo no fim um conjunto de dados resultante de 284 pesquisas.

Desenvolvemos uma aplicação web, em JSP, com uma apresentação similar à de um motor de pesquisa comum. Quando executada uma pesquisa, a aplicação forneceu um total de 10 resultados, com ligações directas às páginas respectivas, para que o utilizador pudesse ser directamente direccionado para o URL em que clicasse. Os resultados apresentados correspondem a entradas de blogues cujo texto seja de relevo para a pesquisa efectuada.

Foi pedido aos utilizadores que fizessem o número de pesquisas que desejassem, e que face aos resultados, escolhessem a opção que lhe parecesse mais relevante para o conteúdo que procurava. Não era possível ao utilizador voltar atrás depois de clicar numa das opções, sendo assim o teste de aferência do melhor resultado. Ao ser direccionado para a página correspondente à entrada de blogue em que clicou, o utilizador nunca esteve ciente de qualquer critério que tenhamos usado para o cálculo. Desta forma, garantimos resultados imparciais, característica que, se não existisse, tornaria os resultados inválidos.

Os resultados para cada pesquisa foram processados segundo dois critérios variáveis, seleccionados aleatoriamente:

- Medida utilizada
- Peso da medida na pontuação do resultado

A aplicação registava para cada clique num resultado o método utilizado para a disponibilização dos resultados, o peso com que este havia sido associado à pontuação obtida pelo $tf - idf$, a posição no top10 de resultados em que o utilizador tinha clicado e a(s) palavra(s) procuradas. Assim, para cada pesquisa e selecção de um resultado era adicionado num ficheiro uma entrada com o seguinte formato:

9 3 portugal 0,5

em que o primeiro algarismo corresponde à posição clicada, e o segundo à medida associada, neste caso contagem de ligações. Segue-se a palavra pela qual o utilizador procurou, e o peso dado ao valor da medida associado ao blogue.

As medidas consideradas para o processamento da pesquisa foram as analisadas ao longo desta dissertação, nomeadamente o h-index, o g-index, e a contagem de citações para cada blogue. A aplicação web seleccionou para cada pesquisa uma medida a utilizar, e inseria-a na avaliação das entradas resultantes. A seguir, os resultados foram ordenados pela sua pontuação final, sendo estes finalmente apresentados ao utilizador.

O peso com que a medida foi associada à pontuação atribuída pelo Terrier, foi variável, e também seleccionada aleatoriamente. Utilizámos a seguinte fórmula de pesos:

$$P_t * P_1 + C_m * P_2 \tag{6.4}$$

Em que P_t representa a pontuação $tf - idf$, devolvida pelo Terrier, o P_1 e P_2 representam dois pesos, cuja soma é igual a 1, e C_m representa a classificação atribuída a cada um dos blogues correspondentes às entradas resultado, segundo a medida previamente seleccionada. Os valores para P_1 eram seleccionados entre o seguinte conjunto: $\{0;0,2;0,4;0,5;0,6;0,8;1\}$, sendo atribuído à variável P_2 o restante para obedecer à condição de que a soma dos dois pesos fosse 1. Representámos no gráfico 6.1 os resultados obtidos nos testes realizados. No eixo dos *yy* representamos o valor médio para a posição na ordenação clicada pelos utilizadores, e no eixo dos *xx*, o peso aplicado às medidas. Os resultados revelaram-se promissores, com o h-index e especialmente o g-index a obterem posições de selecção nas ordenações consideravelmente melhores do que no caso da medida de citações. Os resultados foram piores para pesos mais altos, pelo facto de a relevância dos resultados se reduzir drasticamente no caso da pontuação devolvida pelo Terrier não ser considerada.

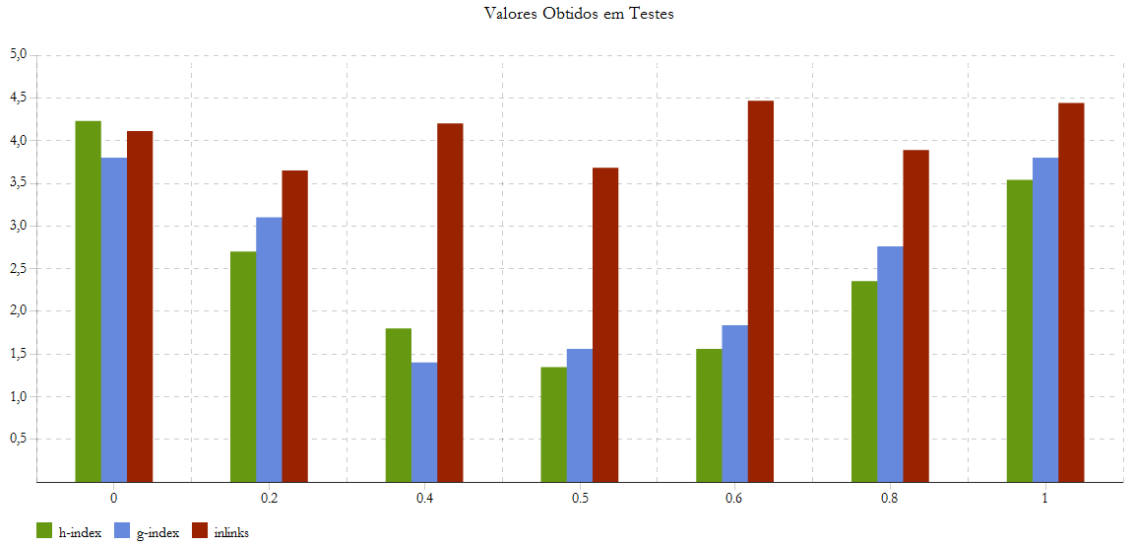

Figura 6.1: Comparação de resultados por peso e método

Capítulo 7

Conclusão

O trabalho que apresentámos neste relatório foi conduzido no período de cinco meses. Durante este período realizámos pesquisa sobre as áreas associadas ao tema da dissertação, elaborámos uma caracterização da colecção que nos foi cedida, aplicámos três métodos de classificação de importância à colecção, analisámos os resultados e estabelecemos comparações entre os resultados dos diferentes métodos. Realizámos um teste ainda simples de avaliação de impacto dos índices em pesquisas e procurámos obter conclusões algo cruas sobre os resultados obtidos. Resta ainda muita matéria passível de ser explorada na adaptação destes métodos de classificação na área dos blogues, que pretendemos continuar a desenvolver, esperando contribuir para esta área de investigação em grande crescimento.

7.1 Discussão

Com base nos resultados apresentados no quinto capítulo deste documento, podemos concluir que as medidas usadas para medir a produtividade e importância de cientistas podem de facto ser adaptadas com sucesso nos blogues. Observámos que a adaptação que realizámos (do h-index no contexto dos blogues) permitiu obter resultados diferentes dos obtidos através de métodos mais clássicos, como a contagem de ligações entre blogues (ao género do PageRank e BlogRank). Com valores diferentes associados aos blogues, e consequentemente a obtenção de ordenações diferentes, concluímos que o h-index e o g-index podem de facto causar um impacto positivo. Tradicionalmente, são procurados vários métodos diferentes para o estabelecimento de um conjunto de critérios combinados, que obtêm uma classificação o mais abrangente e eficaz possível. Através da comparação de ordenações, os dois índices mostraram-se capazes de oferecer novos valores e um novo peso para uma possível inserção, por exemplo, no algoritmo de cálculo

de ordenações de um motor de pesquisas. O h-index revelou resultados especialmente diferentes, em comparação aos resultados obtidos por contagem de citações. Os valores obtidos através do h-index foram mais diferentes do que no caso do g-index, por este valorizar especialmente os blogues de topo. Ambos os métodos aparentam ser robustos face à presença de auto-citações, sendo estes muito menos vulneráveis a este factor do que a classificação por número de ligações. Com base nos testes que realizámos, a tendência aponta para resultados mais positivos da parte do h-index quando utilizado em conjunto com um método de avaliação de relevância para pesquisas. A inserção da medida na apresentação de resultados apresentou um bom desempenho, e a capacidade desta medida de oferecer ordenações diferentes das obtidas com outros métodos é mais um ponto a favor para a sua utilização.

A maior fragilidade destes índices aparenta ser a fraca capacidade de discriminação de valores, visto ser difícil atingir valores altos de h ou g num meio em que existem poucas citações, como acontece no caso dos blogues.

7.2 Trabalho Futuro

Seria de interesse realizar um maior número de testes, com mais utilizadores e resultados, numa plataforma online. Esta tarefa é para nós prioritária, porque a nosso ver, o h-index apresenta já resultados muito favoráveis à perspectiva de ser considerado uma medida válida de importância, podendo estes testes corroborar o que os resultados até à data têm permitido concluir.

Outra possibilidade que consideramos com potencial para explorar seria a implementação de uma variante do h-index com componente temporal. Tal como Mishne [Mis06] refere, a actualidade dos resultados de uma pesquisa pode ser fundamental para um utilizador, pelo que implementar uma medida que tivesse em conta este aspecto poderia proporcionar valores interessantes.

Finalmente, gostaríamos também de ter elaborado mais a caracterização da colecção que possuíamos. A partir do trabalho que realizamos a esse nível, poderá ser realizado posteriormente um outro que analise mais cuidadosamente situações como a quebra na contagem de blogues no Blogspot e das suas entradas a partir de Março de 2007. Seria interessante responder às questões que se formularam em redor do máximo local de 25 entradas para alguns blogues. Também não temos conhecimento, por exemplo, de quantos blogues estão ainda activos, nem se os resultados estatísticos, e gráficos deles derivados, são de facto resultado de um comportamento de publicação ou se são alterados pela actividade da *spider*.

Referências

[BD07] L. Bornmann and H. Daniel. Convergent validation of peer review decisions using the h index extent of and reasons for type i and type ii errors. *Journal of Informetrics*, 1(3):204–213, 2007.

[BMD07] L. Bornmann, R. Mutz, and H. Daniel. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5):1381–1385, 2007.

[CM06] B. Cronin and L. Meho. Using the h-index to rank influential information scientists. *Journal of the American Association for Information Science and Technology*, 57(9):1275–1278, 2006.

[Egg06] L. Egghe. An improvement of the h-index: the g-index. *ISSI Newsletter*, 2(1):8–9, 2006.

[Gla06] W. Glanzel. On the opportunities and limitations of the h-index. *Science Focus*, 1:10–11, 2006.

[Hir05] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, November 2005.

[Hur06] M. Hurst. 24 hours in the blogosphere. *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[Jin06] B. Jin. h-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1):8–9, 2006.

[Jin07] B. Jin. The ar-index: complementing the h-index. *ISSI Newsletter*, 3(1):6+, 2007.

[JLRE07] B. Jin, L. Liang, R. Rousseau, and L. Egghe. The r- and ar-indices: complementing the h-index. *Chinese Science Bulletin*, 52(6):855–863, 2007.

[Ken48] M. Kendall. Rank correlation methods. *Charles Griffin & Company Limited*, 1948.

[KJF06] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*, May 2006.

[Kos06] M. Kosmulski. A new hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3):4–6, 2006.

[KSV06] A. Kritikopoulos, M. Sideri, and I. Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *AAA-IDEA '06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, volume 3169, pages 229–240, New York, NY, USA, 2006. ACM.

[Meh07] L. Meho. The rise and rise of citation analysis. *Physics World*, 20(1):32–36, 2007.

[Mis06] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. *2006 TREC Blog Track*, 2006.

[MO06] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. *DCS Technical Report Series*, 2006.

[NSGS04] B. Nardy, D. Schiano, M. Gumbrecht, and L. Swartz. Im blogging this: A close look at why people blog. *In submission to CACM*, 2004.

[OAP+05] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier information retrieval platform. *Proceedings of the 27th European Conference on IR Research*, ECIR 2005, 2005.

[Pag98] L. Page. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Library Working Paper*, pages 1999–0120, 1998.

[Pin08] J. Pinto. Detection methodologies for blog trends. Master's thesis, Faculdade de Engenharia da Universidade do Porto, July 2008.

[QRSA07] V. Qazvinian, A. Rassoulian, M. Shafiei, and J. Adibi. A large-scale study on persian weblogs. *The proceedings of 12th international joint conference on Artificial Intelligence, workshop of TextLink2007*, 2007.