FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Avaliação da medida h-index na ordenação de resultados na blogosfera

Tiago Valente da Costa

Relatório final submetido no âmbito da disciplina Dissertação do
Mestrado Integrado em Engenharia Electrotécnica e de Computadores
Major de Telecomunicações

Orientador: Maria Cristina de Carvalho Alves Ribeiro
Co-orientador: Sérgio Nunes

Junho de 2009

U.PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES 2008/2009

A Dissertação intitulada

"AVALIAÇÃO DA MEDIDA "H-INDEX" NA ORDENAÇÃO DE RESULTADOS NA BLOGOSFERA"

foi aprovada em provas realizadas em 22/Julho/2009

o júri

Presidente Professor Doutor Francisco José de Oliveira Restivo
Professor Associado do Departamento de Informática da Faculdade de Engenharia da Universidade do Porto

Professor Doutor José Paulo de Vilhena Geraldes Leal
Professor Auxiliar do Departamento de Ciência de Computadores da Faculdade de Ciências da Universidade do Porto

Professora Doutora Maria Cristina de Carvalho Alves Ribeiro
Professora Auxiliar do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

Mestre Sérgio Sobral Nunes
Assistente do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projecto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extractos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são correctamente citados.

Autor - **TIAGO VALENTE DA COSTA**

Faculdade de Engenharia da Universidade do Porto

Resumo

A constante evolução tecnológica no acesso e uso da Internet contribuiu para que existam cada vez mais pessoas a procurarem nela formas de se exprimirem e partilharem as suas experiências e informação. Os blogues são ferramentas que surgiram dessa necessidade e dessa evolução. O facto de permitirem uma actualização rápida, a partir de textos denominados de entradas, sem que o autor tenha qualquer tipo de conhecimento técnico é a chave principal do sucesso dos blogues. Um bloguista consegue criar um diário, onde pode colocar opiniões, fotografias e conteúdos relativos a assuntos do dia a dia. Com o crescente número de blogues torna-se útil implementar medidas que permitam classificá-los por forma a facilitar a pesquisa e escolha por parte do utilizador. Surge assim a necessidade de procurar e avaliar novas medidas para classificar blogues.

Jorge Hirsch propôs em 1995 a medida h-index, que permite classificar a produção científica de investigadores. Esta medida tem em conta o número de publicações que um cientista produz e o número de vezes que são citadas pelos seus pares. Um cientista tem classificação h se h dos seus artigos tiverem cada um pelo menos h citações. Em 2008, José Mário Branco analisou a aplicação da medida h-index na classificação de blogues. Para este efeito estabeleceu um paralelismo entre as características de ambos os contextos, fazendo corresponder cada blogue a um cientista, e às entradas do blogue os artigos publicados pelo cientista. Foi possível concluir que esta medida é facilmente implementável e mostra resultados positivos na ordenação de blogues.

Os resultados da investigação de José Mário Branco justificaram a realização de uma avaliação mais completa e detalhada. Assim, foram planeadas e implementadas duas experiências com utilizadores e feita a respectiva análise de resultados. A primeira experiência parte de uma ordenação absoluta, e foi pedido ao utilizador a sua preferência perante duas listas de blogues ordenadas por duas de três medidas independentes de interrogações. Na segunda experiência foi pedido aos utilizadores que realizassem interrogações, sendo-lhes pedido a sua preferência perante uma lista de resultados obtidas com ordenações que usavam o h-index como uma das suas componentes.

Com estas experiências foi possível concluir que a medida h-index é promissora na ordenação de resultados na blogosfera. Os resultados desta investigação justificam a continuação da análise da avaliação da medida h-index.

Abstract

The continuous advances made in web-realated technogies, particularly new ways in accessing and using the Internet has played a heavy role in boosting the web's popularity. As with any new trend people tend to seek ways to express and share their experiences and opinions. One way of doing this is through the use of blogs. Anyone without any kind of technical knowledge can create a journal in which he can post and edit reviews, photos and content related to matters of everyday life. With the growth of blog creation, it became useful to identify features in order to classify them, improving the user's search experience. This led to the need of studying and evaluating new features to rank blogs.

In 1995, Jorge Hirsch proposed a measure called h-index, which aims to classify scientists in terms of the quality of their work. This measure takes into account the number of publications authored by a scientist and the number of times they are cited by his peers. A scientist has rank h if h of his papers have been cited at least h times. In 2008 José Mário Branco investigated if the h-index measure could be used in blogs. He established a connection between the characteristics of both contexts, matching each blog to a scientist and each blog entry to the papers published by the scientist. The research showed that it is possible to adapt h-index to blogs and also showed good results in ordering blogs.

The results of José Mário Branco justified the development of a more complete and detailed research. So two experiments with users were planned and implemented in order to analyze the results. The first experience focused on absolute rankings, where the user should pick his preferred list of blogs, up of two lists presented to him and sorted by two of three measures independent of any kind of query. In second experiment the user had to submit queries in order to get a list of results where he had to pick one of them. This list of results was sorted using h-index as one of its components.

With these experiments we have concluded that h-index, as a feature to sort results in the blogosphere, is promising, justifying further and detailed study of the h-index measure.

Agradecimentos

Em primeiro lugar gostaria de agradecer as pessoas responsáveis pela ideia desta dissertação, Prof. Cristina Ribeiro e ao Eng. Sérgio Nunes, por me orientarem, pela paciência e pela oportunidade que me deram.

Queira agradecer a toda a equipa da SAPO responsáveis pela colecção de blogues disponibilizada para este trabalho.

Gostava de agradecer aos meus amigos que me acompanharam e apoiaram durante todo o tempo de realização desta dissertação, nomeadamente: Ana Vieira a correctora de português, João Machado o brincalhão, João Almeida companheiro de grupo durante todo o meu percurso académico, Luís Vigário o grande defesa central, Pedro Correia o madeirense, Sérgio Sá o homem das tecnologias e Tiago Rocha Costa o homem cuja personalidade é inconfundível. Todas estas particularidades, apesar de parecerem brincadeira, foram importantes para me manter motivado e concentrado no meu trabalho.

Um enorme agradecimento à minha família, em especial aos meus pais, Adriano Costa e Ilma Costa, cujo apoio incondicional, enorme respeito e estrondosa paciência sustenta todo o trabalho que fiz e me permitiu chegar onde me encontro hoje.

Por fim, não podia deixar de agradecer às centenas de pessoas que participaram nas experiências realizadas nesta dissertação. Estas pessoas são amigos de verdade porque nunca deixaram de ajudar quando eu mais precisei e sem nunca pedirem nada em troca.

Tiago Valente Costa

"A fool thinks himself to be wise,
but a wise man knows himself to be a fool."

William Shakespeare

Conteúdo

Lista de Figuras

Lista de Tabelas

Abreviaturas e Símbolos

ISP	*Internet Service Provider* - Fornecedor de Serviço de Internet
MSN	*Microsoft Network*
JSP	*JavaServer Pages*
JDBC	*Java Database Connectivity*
API	*Application Programming Interface* - Interface de Programação de Aplicações
SQL	*Structured Query Language*
PDF	*Portable Document Format*
XML	*Extensible Markup Language*
HTML	*Hypertext Markup Language*

Capítulo 1

Introdução

1.1 Contexto

Os blogues são páginas Web que de forma rápida e eficiente permitem organizar e apresentar os seus conteúdos. Neste sentido, dada a facilidade de acesso e a evolução dos meios e mecanismos orientados à Internet, os blogues conquistaram um lugar de destaque na nossa sociedade. Esta evolução deve-se a que o utilizador comum, sem qualquer conhecimento em tecnologias específicas, consegue facilmente criar um blogue e publicar novas entradas com conteúdos criados por si.

Actualmente, a Web é utilizada como a principal fonte de informação em diversas áreas. O utilizador comum, através de motores de pesquisa *on-line*, obtém informação em maior ou menor quantidade, mais ou menos específica, de uma forma rápida e cómoda.

A informação proveniente do meio científico, como artigos, livros e publicações é vasta e nem toda é constituída por informação relevante para uso em investigação. O meio científico, ao longo dos tempos, tem vindo a usar métodos de bibliometria, frequentemente utilizados em bibliotecas e outros fornecedores de acesso a informação, para analisar o impacto do tema, do artigo e do cientista e assim obter uma classificação relativa dos mesmos. Esta classificação é obtida através da análise de textos, citações e conteúdos.

O h-index é uma medida proposta por Hirsch [1], que permite classificar a qualidade e a produtividade dos elementos da comunidade científica ao longo das suas carreiras. Esta medida tem em conta o número de publicações que um cientista produz, bem como o seu impacto na comunidade. A sua como base é o número de vezes que cada um dos artigos é citado pelos seus pares.

Nas palavras de Hirsch:

> "Eu proponho o h-index, definido pelo número de artigos com um número de citações maior ou igual que h, como um índice útil para caracterizar a qualidade científica de um cientista. ... Um cientista tem índice h se h dos seus N artigos têm cada um pelo menos h citações, e os outros $(N-h)$ artigos têm menos que h citações cada um."[1]

José Mário Branco [5] analisou a possibilidade de aplicar a medida h-index em blogues, para este efeito estabeleceu um paralelismo entre o meio dos blogues e o meio científico e propôs aplicar o algoritmo de Hirsch na área da blogosfera. As ferramentas utilizadas e o ambiente experimental utilizado, com base na colecção cedida pela SAPO[1], permitiram elaborar uma caracterização, aplicar três métodos de classificação de importância à colecção, analisar e estabelecer comparações entre os resultados dos métodos utilizados. Os resultados obtidos permitiram concluir que esta adaptação é possível. Contudo, no que diz respeito à utilidade da medida na recuperação de informação, somente foi efectuado um pequeno teste de avaliação do impacto dos índices em pesquisas. Apesar de os resultados serem satisfatórios, não são completos e necessitam de uma investigação mais aprofundada que constitua uma avaliação desta medida de classificação na área da blogosfera.

1.2 Objectivos

A presente dissertação tem como objectivo a avaliação da medida h-index na ordenação dos resultados na blogosfera. Para tal, foi construído um ambiente experimental para testar os resultados obtidos na ordenação de blogues. Propôs-se assim:

- Calcular medidas de importância de blogues, tais como o h-index e a contagem de citações (*inlinks*);
- Planear um ambiente experimental de testes com utilizadores;
- Implementar um motor de pesquisa no qual a medida h-index possa ser usada como uma das componentes de ordenação de resultados;
- Conduzir um conjunto de experiências com utilizadores;
- Estabelecer métodos de avaliação de resultados.

1.3 Estrutura do documento

O presente trabalho está organizado em 6 capítulos. Do Capítulo 2 ao 4, descrevemos o estado de arte e alguns casos de estudos efectuados sobre o presente tema. No Capítulo 5, é descrita a metodologia das experiências realizadas. Os resultados obtidos nas experiências são analisadas no capítulo seguinte. Por fim, no Capítulo 7 apresentamos as conclusões.

[1] *ISP* portuguesa - *www.sapo.pt*

Capítulo 2

A medida h-index

2.1 Definição

O h-index é uma medida de avaliação de qualidade e produtividade de elementos da comunidade científica, proposta em 1995 por Jorge Hirsch [1]. Esta medida tem em conta o número de publicações de um cientista e o impacto que estas produzem na comunidade. O h-index tem como base o número de vezes que cada um dos artigos de cada cientista é citado pelo seus colegas. Ao contrário das outras medidas que somente pesam o número de citações feitas aos artigos de um cientista, o h-index procura avaliar toda a investigação que um cientista produz ao longo da sua carreira. Hirsch teve como base essencial, para a sua proposta do h-index, a área da física, no entanto ela tem sido adaptada para outras áreas científicas.

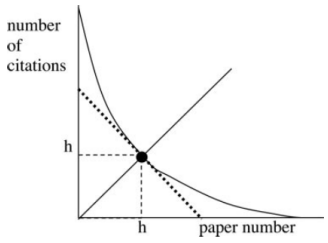

Figura 2.1: Método de cálculo do h-index retirada de, retirada de [1]

Para o cálculo do valor de h-index de um cientista, é tido em conta o número de artigos científicos que este publicou até à data e o número de vezes que estes são citados noutros artigos. A figura 2.1 ilustra o princípio por detrás do cálculo do h-index. Um cientista tem h-index com valor h quando h dos seus artigos são citados pelo menos h vezes e os restantes menos que h vezes. Observemos o exemplo nas tabelas 2.1 e 2.2.

Tabela 2.1: Exemplo de cálculo do valor de h-index

Documento	**No. de Citações**
Amarelo	2
Vermelho	24
Castanho	2
Verde	17
Preto	4
Branco	15
Azul	6
Roxo	11
Rosa	8

Tabela 2.2: Ordenação do exemplo da tabela 2.1 por número de citações

No. do documento	**Documento**	**No. de Citações**
1	Vermelho	24
2	Verde	17
3	Branco	15
4	Roxo	11
5	Rosa	8
6	**Azul**	**6**
7	Preto	4
8	Amarelo	2
9	Castanho	2

Na Tabela 2.1 temos a lista de artigos de um cientista e o número de vezes que esses artigos são citados. Na Tabela 2.2, é feita uma ordenação pelo número de citações desses mesmos artigos. Pode-se verificar que existem 6 artigos que são citados no mínimo 6 vezes e que os restantes são citados menos de 6 vezes, assim, o valor de h do cientista associado é 6.

Segundo Hirsch, o h-index é preferível a outra medidas para a avaliação científica de um indivíduo, porque o cálculo do valor de h utiliza os seguintes critérios:

- **Número total de artigos:**

  Este critério permite medir a produtividade de cada indivíduo, no entanto, não mede a importância ou o impacto dos artigos criados pelo cientista.

- **Número total de citações:**

  O número total de citações permite medir o impacto dos artigos criados pelo cientista. No entanto pode traduzir mal este impacto por vários motivos:

  - o valor pode ser influenciado por um pequeno número de "grandes êxitos";

  - dá um peso indevido nos artigos revistos e é sensível a auto-citações; esta desvantagem é detalhada mais à frente e é denominada de auto-citações.

- **Citações por artigo, razão entre o número total de citações e número total de artigos:**

  Segundo Hirsch, este critério permite comparar cientistas de diferentes idades, no entanto:

  - se um cientista produzir poucos artigos e estes são altamente citados, então a razão é alta e estará a recompensar a baixa produtividade;

  - se um cientista produzir muitos artigos e se o número de citações for menor que o número de artigos produzidos, estará a penalizar a alta produtividade.

- **Número de artigos significativos definido como o número de artigos com y citações:**

  Este critério elimina as desvantagens enunciadas anteriormente e permite dar uma ideia do impacto amplo e sustentado de um cientista, no entanto o valor de y é arbitrário e irá aleatoriamente favorecer ou desfavorecer indivíduos, ou seja, y necessita de ser ajustado para diferentes extensões da carreira.

- **Número de citações para cada q artigos mais citados:**

  O facto de estarmos a analisar o número de citações dos artigos mais citados, permite eliminar a desvantagem indicada no critério acima. Mas este critério não é um número, o que torna difícil de obter e, como consequência, comparar. O valor de q também favorece ou desfavorece indivíduos.

Estes indicadores foram enunciados pelo próprio Hirsch [1]. Wolfgang Glanzel, um ano após a apresentação do h-index, elaborou uma análise às vantagens e desvantagens do h-index [5].

As vantagens enunciadas por Glanzel foram:

- O h-index é um indicador simples e facilmente implementável;
- É combinada produtividade e qualidade do decurso da vida de um cientista;
- O h-index apresenta robustez, visto que um artigo não afecta grandemente os valores de h;
- Pode ser considerado para avaliação qualquer tipo de documento;
- É possível utilizar o h-index em conjunto com outros indicadores.

As desvantagens foram:

- Os cientistas com um pequeno número de documentos produzidos, se bem que de grande qualidade e impacto, obterão um valor de h baixo;
- Mesmo avaliando a produtividade, o h-index possibilita a situação em que um indivíduo não produtivo veja ainda o seu valor de h aumentar, devido às novas citações referentes a trabalhos anteriores;
- Cientistas no início de carreira estarão em desvantagem, numa classificação feita por este método;
- O h-index apresenta resultados positivos na avaliação de cientistas com excelentes desempenhos, mas pode falhar com desempenhos mais medianos.

As análises às vantagens e desvantagens do h-index proporcionaram alternativas a esta medida, nomeadamente o g-index, H(2)index e a-index, o r-index e o ar-index:

- **g-index:** Medida criada por Egghe, L., que alegava a necessidade de aumentar a robustez do algoritmo de Hirsch na parte superior da ordenação, ou seja, acima do número h atribuído. Imaginemos o seguinte caso: temos dois cientistas A e B, que têm o mesmo valor de h e em que o cientista A, nos seus artigos de topo, tem alguns com poucas citações e o cientista B com milhares de citações [5]. Ao contrário do algoritmo de Hirsch que classifica os cientistas como idênticos, o g-index atribui maior classificação ao cientista B, dando maior importância ao impacto dos artigos de maior popularidade no meio científico. Este algoritmo, ao contrário do h-index que assume o número de citações para cada documento, usa o somatório de citações a partir do topo da ordenação σNc e compara com o quadrado da posição desse artigo na ordenação r^2. Quando atinge a última posição da lista onde $\sigma Nc > r^2$, o valor dessa posição é o g a atribuir ao cientista.
- **H(2)index ou a-index:** Criado por Kosmulski, um cientista tem valor $h(2)$ se $h(2)$ dos seus artigos mais citados tiverem cada um pelo menos $h(2)^2$ citações [5].

- **r-index:** Jin et al., criadores desta medida, achavam injusto o a-index para cientistas de topo, porque ao efectuar uma divisão pelo valor de *h* no cálculo desse algoritmo, os indivíduos com um valor *h* alto seriam prejudicados [5]. Assim, para corrigirem esta desvantagem, Jin et al. propuseram que o valor de *h* de um cientista deve ser alterado não pela divisão da média dos seus valores, mas pela sua razão.

- **ar-index:** O ar-index foi uma proposta feita por Jin onde pretendia adaptar o r-index para o contexto temporal. Esta medida, para além de contabilizar a quantidade de citações no núcleo de Hirsch, também considera a idade das publicações. Assim, o valor desta medida pode aumentar ou diminuir ao longo do tempo [5].

2.2 Auto-Citações

Como já foi referido anteriormente, as auto-citações dificultam o cálculo do h-index. Imaginemos que um cientista cita nos seus artigos documentos que criou previamente; se a frequência de citações for elevada, o cientista irá ser beneficiado no seu valor de *h*.

Ao prever esta situação Hirsch refere que as auto-citações devem ser removidas do cálculo do valor do h-index. Esta remoção poderá ser realizada de duas maneiras:

- **Forma absoluta**, em que é removido o número de auto-citações do cálculo na sua totalidade;

- **Forma parcial**, em que são removidas simplesmente as citações presentes nos artigos com número de citações acima do valor de *h* do cientista.

O valor do h-index deverá ser ajustado em função das alterações observadas para obtermos assim um novo valor do h-index. Na última solução mencionada, se um cientista pretender aumentar o seu valor do h-index, terá que procurar citar os seus artigos que estejam abaixo do presente valor do h-index, obtendo assim um aumento gradual do seu valor de *h*.

Hirsch conclui que a melhor opção para o cálculo do h-index é a remoção total destas citações.

Capítulo 3

Blogues, medidas de classificação, motores de pesquisa e avaliação

A evolução da Web tem vindo a disponibilizar serviços de qualidade aos utilizadores. Os motores de pesquisa e as medidas de classificação têm vindo a aumentar ao longo deste percurso evolutivo. E estes serviços têm sido sujeitos a diversas investigações, não só com o intuito de avaliar a sua qualidade, mas também as suas características.

3.1 Blogues

3.1.1 Definição

Um blogue é uma página Web que permite uma actualização rápida a partir de textos denominados entradas. Estes, normalmente, são organizados de forma cronológica inversa, do mais recente para o mais antigo, costumam abordar a temática pela qual o blogue foi criado e podem ser escritos por um número variável de pessoas, de acordo com as regras do respectivo blogue.

O conceito de blogue tem evoluído adaptando-se às necessidades emergentes dos seus utilizadores. Mas o que potencia um blogue é o facto de ser possível a um utilizador comum, sem qualquer conhecimento de linguagens de programação, criar um repositório de informação que considere ser de relevo, sejam textos de opinião, imagens, *links* para notícias, ou outros.

Existe também um elemento singular na blogosfera denominado de *splog*, um blogue de spam. Este possui um conteúdo publicitário que é gerado automaticamente e inunda a blogosfera com a sua informação, geralmente inútil para utilizador. Estes blogues são constantemente actualizados e têm um número muito elevado de ligações. A maneira como são criados, faz com que muitos utilizadores os visitem, melhorando assim a possibilidade de rentabilização com publicidade para venda de produtos.

Em 2001, Nardi et al. [6] publicaram um artigo que analisa as motivações de alguns bloguistas, avaliando vários critérios, nomeadamente razões pelas quais os levaram a criar um blogue e a publicar o mesmo. Para tal, numa pequena amostra de 16 homens e 7 mulheres, com idades entre os 19 e os 60 anos, Nardi et al. fizeram entrevistas: pessoais, telefónicas, email ou mesmo por

mensagens instantâneas. Com estas entrevistas, os autores, puderam salientar cinco motivos que levaram aos utilizadores a criar e publicar os seus blogues:

- Documentação da vida do autor;
- Comentários e opiniões;
- Expressão de emoções pessoais;
- Estruturação de raciocínios e opiniões;
- Criação e manutenção de fóruns de comunidades específicas.

Devido à sua crescente utilização, a blogosfera tem sido alvo de intensa investigação. Aspectos relacionados com os blogues, blogosfera e com a classificação destes, segundo vários critérios, foram já explorados em trabalhos de investigação.

3.1.2 Colecção em uso

Foi estimado em Agosto de 2008 que a blogosfera mundial continha cerca de 133 mil milhares de blogues [2]. O primeiro blogue português foi criado nos anos 90 [2], no entanto, a comunidade portuguesa só se tornou familiar com este conceito em 2006, o mesmo ano em que a SAPO lançou o serviço de blogues[1] dedicado à comunidade portuguesa.

Baseado numa colecção de blogues fornecida pela SAPO, Telmo Couto apresentou uma dissertação [2] que assentava não só na apresentação de múltiplas estatísticas da blogosfera portuguesa, mas também na comparação de resultados obtidos noutros estudos. Esta colecção continha um conjunto de blogues hospedados no serviço de blogues da SAPO e um conjunto de outros blogues provenientes de outros fornecedores.

Um aspecto importante desta colecção é que abrange um vasto período de tempo, contendo blogues e entradas até ao fim de Junho de 2008. Permite a investigação da evolução da actividade da blogosfera portuguesa ao longo do tempo.

Com esta colecção, Couto pôde observar que a blogosfera portuguesa aumentou em número de blogues e entradas criadas. Em Junho de 2008, os blogues portugueses apresentavam uma média de 20 entradas, o que é o dobro do ano anterior. O aumento do número de entradas criadas pode ser observado na figura 3.1.

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

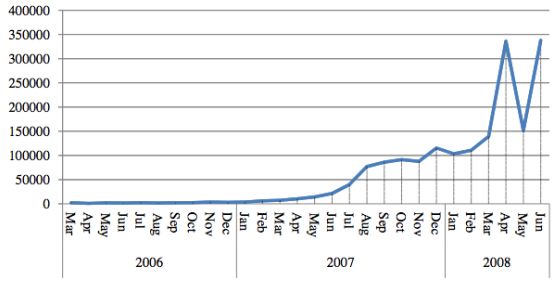

Figura 3.1: Número de entradas criadas ao longo do tempo, retirada de [2]

3.2 Eficácia de motores de pesquisa

Na área da avaliação de motores de pesquisa, Croft et al. [7] fazem três distinções primárias:

- **Eficácia:** mede a capacidade que o motor de pesquisa tem para encontrar a resposta útil para o utilizador;
- **Eficiência:** mede a rapidez com que o motor de pesquisa encontra uma resposta ao utilizador, seja ela útil ou não;
- **Custo:** mede o custo de implementação do motor de pesquisa.

A eficácia e eficiência são afectadas por muitos factores, como a interface usada para disponibilizar os resultados da pesquisa ao utilizador, as técnicas, as interrogações feitas e a relevância da resposta.

Ao longo deste trabalho, serão usados os conceitos de recuperação de informação e de documentos semi-estruturados:

- **Documentos semi-estruturados:** são dados estruturados mas que têm partes significativas em texto integral, o que os torna difíceis de serem interpretados por computadores;
- **Recuperação de informação:** é uma área da computação que lida com o armazenamento de documentos não estruturados ou semi-estruturados e com a recuperação automática da informação associada aos mesmos.

Podemos saber como implementar um certo motor de pesquisa que seja eficiente, para tal é necessário considerar um investimento em processadores, memória, disco, etc. Mas também podemos implementar um motor de pesquisa que seja eficaz, ou mesmo um equilíbrio entre ambos. O custo é usado como critério final de escolha, caso haja dúvidas relativamente aos parâmetros anteriormente citados.

Dois casos extremos para escolha destes factores é pesquisar usando *grep*, ou usando a pesquisa de uma organização como a Biblioteca do Congresso dos EUA [8]. O exemplo que Manning apresenta é fazer uma pesquisa de um texto extenso numa base de dados utilizando *grep*, teremos,

não só uma eficácia e eficiência muito baixa, mas também um custo reduzido. Pesquisar utilizando a Biblioteca do Congresso dos EUA irá produzir resultados com grande eficácia, devido ao esforço manual envolvido, o que o torna dispendioso. Pesquisar directamente, usando um motor de busca eficaz, é o ponto de equilíbrio entre estes dois extremos.

As técnicas de recuperação e indexação num motor de pesquisa têm muitos parâmetros, que podem ser ajustados para optimizar a sua performance, em termos de eficácia e eficiência. Tipicamente os melhores valores para estes parâmetros são determinados através de dados de treino e funções de custo.

3.3 Medidas de classificação

Um dos requisitos básicos na avaliação é a comparação de resultados de diferentes técnicas. Para esta comparação seja razoável é necessário assegurar que as experiências sejam repetidas. As configurações da experiência e os dados a usar têm de ser fixos.

A colecção para teste não pode ser pequena e deve ser constituída por um número de dados que possam reflectir as mudanças feitas ao longo do tempo nos dados e nas comunidades de utilizadores.

Em seguida é descrito um conjunto de avaliações efectuadas a medidas que permitem classificar páginas Web, considerando nesta avaliação os blogues, tema central deste trabalho.

3.3.1 *PageRank*

Em 1998, Page et al. [3] publicaram um artigo que descrevia um método para classificação de páginas Web, denominado *PageRank*. Segundo os autores, nesse ano foi estimado que havia 150 milhões de páginas Web, extremamente diversificadas. No entanto, os motores de pesquisa existentes lidam com utilizadores inexperientes e páginas Web criadas para manipular as funções de classificação dos motores de pesquisa. A Web é constituída por hiperligações e proporciona informação auxiliar no topo dos textos das páginas Web, foi neste ambiente que Page et al. desenvolveram uma medida de classificação para produzir um modelo global de classificações de páginas Web.

A particularidade do *PageRank* é que se baseia em grafos representativos das ligações Web. Segundo Page et al., em 1998 havia 150 milhões de páginas Web, como já foi mencionado, e 1,7 milhares de milhões de hiperligações. Ele definiu para cada página um número de ligações de saídas e de entradas, como é ilustrado na Figura 3.2.

Page et al. mencionam que as páginas Web com mais impacto e de maior qualidade, são aquelas que têm um maior número de ligações, tanto de entrada como de saída. Definiram o seu algoritmo com o objectivo de obter uma aproximação ou medida de importância de uma página. Segundo Page et al.:

> "...uma página tem maior classificação numa ordenação, se a soma dos valores na ordenação de páginas que para ela ligam for alta."

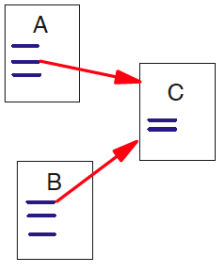

Figura 3.2: A e B têm ligações de saída para C, retirada de [3]

Assim, o cálculo do *PageRank* é obtido pela seguinte fórmula: $R(u) = c\sum_{v \in B_u} \frac{R(v)}{N_v}$ em que u representa uma página Web, u e B_u define um conjunto de páginas que apontam para u. Sendo N_u o número de ligações de u e c um factor de normalização para o valor de *PageRank*, o cálculo desta fórmula obterá a posição numa classificação para uma página a avaliar. É ilustrado na Figura 3.3 uma demonstração do cálculo do valor de *PageRank*, onde podemos verificar que a classificação de uma página Web é igualmente distribuída entre as hiperligações que a constitui para contribuir na classificação das páginas para as quais aponta.

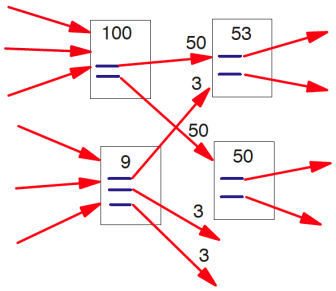

Figura 3.3: Cálculo simplificado do *PageRank*, retirada de [3]

Page et al. descrevem o seu algoritmo como o de um "surfista aleatório", uma entidade abstracta que clica em hiperligação após hiperligação, navegando assim pelo grafo, e eventualmente visitando mais frequentemente páginas mais populares. No seu algoritmo, os autores ainda adicionaram um factor E, com a finalidade de o "surfista" não ficar preso em ciclos infinitos, podendo enveredar por novas hipóteses.

O impacto do estudo feito por Page et al. foi em tudo inovador, causando um tremendo efeito na área de recuperação e classificação de páginas Web.

3.3.2 *BlogRank*

Um dos primeiros documentos de investigação sobre classificação de blogues, surgiu em 2006 por Kritikipoulos et al. [4], e propunha uma versão modificada do método *PageRank* denominada *BlogRank*.

Kritikipoulos et al. defendem que, à parte das relações existentes pelas hiperligações entre blogues, podem ser relacionadas outro tipo de elementos mais implícitos aos blogues, como por

exemplo tópicos comuns, número de participantes e contribuições dos utilizadores. A contribuição dos utilizadores é um exemplo de como as hiperligações não são o mais importante, porque um utilizador pode postar num blogue mas não colocar hiperligações.

Geralmente na blogosfera, as ligações são feitas a nível das entradas e introduzidas por utilizadores. Na Figura 3.4 podemos verificar as relações e hiperligações a nível de entradas, onde as ligações entre entradas são apresentadas por setas, as linhas indicam o autor de cada entrada e ao lado o tópico a que se refere essa entrada. Podemos verificar que existe 11 entradas com somente 3 hiperligações.

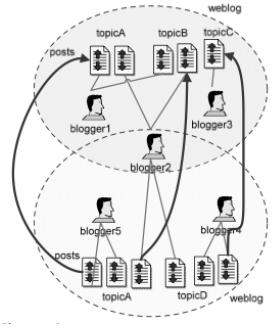

Figura 3.4: Relações e hiperligações a nível de entradas, retirada de [4]

Para solucionar o problema de poucas hiperligações, Kritikipoulos et al. decidiram definir um conjunto de ligações internas entre vários blogues, de forma que partilhassem categorias e autores em comum, podemos verificar esta solução na Figura 3.5.

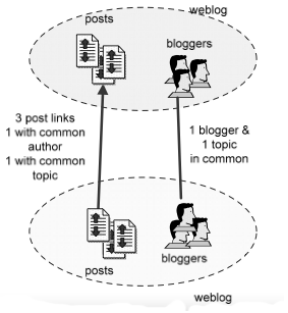

Figura 3.5: Ligações a nível de blogues, retirada de [4]

Ao definir este conjunto de ligações, os cientistas anularam parcialmente o conceito de "surfista aleatório" do *PageRank*. Em substituição consideraram a probabilidade de um utilizador viajar de uma página para a outra, mesmo que não existam ligações entre elas, usando um tema ou autor em comum.

Para fins de teste, Kritikipoulos et al. criaram três métodos para ordenação de blogues:

1. Ignorando as ligações implícitas criadas, correspondente ao *PageRank*;
2. Atribuindo um peso baseado no número de ligações diferentes entre entradas, correspondente à combinação de *BlogRank* e *PageRank*;

3. Atribuindo um peso a ligações implícitas, no caso de *tags* ou autores em comum, correspondente somente ao *BlogRank*.

Os cientistas utilizaram para testes a este método uma colecção amostra fornecida pela Nielsen Buzzmetrics. Kriptikolous et al. observaram que o *BlogRank* era claramente superior aos dois critérios restantes, *PageRank* e *XRank*.

3.3.3 Comparação entre *PageRank* e *Inlinks*

Em 2003, Trystan Upstill estudou qual dos métodos, *PageRank* e *Inlinks*, seria o melhor para classificação de páginas Web [9].

O investigador começou por fazer um pequeno estudo entre os dois métodos. O *PageRank*, como já foi mencionado anteriormente, é uma medida usada pela *Google* para medir a importância e qualidade de páginas Web. Para a visualização do valor deste método, a *Google* dispõe de uma ferramenta denominada de *Google Toolbar* [10]. O valor do *Inlinks* de uma página é o número de ligações existentes de outras páginas para a página em questão. Valorizar estas ligações é assumir que quando existentes para uma página, o autor que as criou está a recomendar a página.

Após esta investigação, Upstill progrediu no sentido de relacionar os dois métodos e analisar os seus resultados. Usando 5370 páginas Web de empresas e 280 páginas de *spam*, a abordagem foi obter o *PageRank* e o *Inlinks* de cada endereço e em seguida analisar as distribuições e relacionamentos entre os dois.

Para a extracção das pontuações de *PageRank* e *Inlinks*, foram usados os motores de pesquisa *Google* e *Fast*. Para relacionar os dois métodos, processou em separado as páginas Web de empresas e nas páginas de *spam*. Upstill encontrou uma relação razoável entre *PageRank* e *Inlinks* para páginas Web de empresas. Em relação às páginas *spam* encontrou uma relação muito semelhante, logo não foi suficiente para identificar qual dos dois métodos era o melhor. Assim, decidiu verificar a tendência relativa a hiperligações de recomendação. Examinou os dados relativos a páginas principais, através de grandes e conhecidas empresas, tal como em relação ao país e tecnologia em uso. Estas experiências reportaram uma relação existente entre *PageRank* e *Inlinks*, nomeadamente a performance e em resultados. Assim sendo, Upstill não vê qualquer tipo de vantagem no uso do *PageRank*, devido aos custos que lhe estão associados, aconselhando assim o uso do *Inlinks*.

3.3.4 Aplicação do h-index em blogues

Em 2008, José Mário Branco [5] propôs a aplicação do h-index para avaliação da importância de um blogue. A actual dissertação é a continuação do trabalho desenvolvido que mostrou resultados muito positivos.

Da mesma forma que nesse trabalho, foi usada uma colecção de blogues maioritariamente portuguesesa (quase na sua totalidade lusófonos), no contexto de uma colaboração entre a Faculdade de Engenharia da Universidade do Porto e a empresa SAPO. Esta colecção é constituída por 54 149 blogues com mais de 3 milhões de entradas e reúne blogues do serviço da SAPO e blogues

de outros serviços que foram encontrados pela SAPO através de *crawling*. Branco começou por analisar a colecção, onde eliminou cerca de 4 mil blogues, por não se encontrarem inseridos no intervalo temporal de Janeiro de 2003 a Dezembro de 2007. Foram considerados 49 940 blogues com 2 993 735 entradas. De seguida, separou esta colecção em três categorias, no que diz respeito a fornecedores de serviço [5]:

- 52% dos blogues são hospedados pela SAPO (25.768 entradas);
- 47% são hospedados pelo Blogspot (23.378 blogues);
- 1% era composto por um conjunto de blogues de diferentes servidores (794 blogues).

Nesta colecção extraiu alguns dados relativos aos hábitos diários de um bloguista português e fez uma análise da sua evolução ao longo do tempo. Após a análise, Branco criou então o paralelismo, já mencionado, para que fosse possível avaliar a medida h-index na classificação de blogues.

No início do seu trabalho, surgiram de imediato duas questões fundamentais, citando-as:

- Deveríamos considerar auto-citações, ou seja, situações em que o bloguista menciona 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 algoritmo na classificação de blogues seria irrelevante, pois só uma pequena parte deste possuiria um valor *h* atribuído).

Da mesma forma que Hirsch aconselha uma remoção das auto-citações [1], Branco também as retirou da sua análise visto que no contexto de blogosfera seria apropriado evitar manipulações nas classificações. No entanto, mostrou resultados relativos ao uso e não uso de auto-citações.

Para que fosse possível avaliar os resultados que iria obter através da ordenação por valores de h-index, Branco decidiu implementar outras duas ordenações, o g-index e a simples contagem de citações (*inlinks*).

Para analisar os resultados, Branco investigou as seguintes situações:

1. Verificou a frequência de ligações entre blogues, ou seja, analisou a quantidade de entradas e blogues que são citados, com e sem auto-citações;
2. Ordenou a colecção de blogues por h-index, g-index e *inlinks*;
3. Utilizou o coeficiente de correlação de ordenações de Kendall, ou coeficiente de *tau* para efectuar uma análise mais profundo nos resultados obtidos;
4. Avaliou as três medidas, com auto-citações, e comparou os seus valores com os obtidos sem auto-citações.

Branco seguiu o exemplo de Kritikipoulos et al. [4] e realizou testes com utilizadores reais, não revelando a nenhum quais os pesos e medidas utilizadas para o cálculo dos resultados que lhes foram apresentados. Tratou-se de uma amostra de 16 pessoas, obtendo no final um conjunto de dados resultante de 284 pesquisas.

Para estes testes, foi construído uma aplicação Web, em JSP, com uma imagem semelhante à de um motor de pesquisa comum. Foi também utilizada a plataforma de recuperação de informação Terrier[2] por forma a criar o índice para a colecção em uso e para futuro uso nas pesquisas. Foi pedido aos utilizadores que fizessem o número de pesquisas que desejassem, e face ao resultado obtido, escolhessem somente uma opção, das que lhe parecesse mais relevante, impossibilitando-os de retroceder na sua escolha.

Em cada pesquisa efectuada, os resultados foram processados segundo dois critérios variáveis, nomeadamente a medida utilizada e o peso da medida na pontuação final do resultado.

A plataforma Terrier ao receber um termo para pesquisa, procura-o no índice que previamente foi criado, e guardado em formato *XML*, e retorna uma lista de resultados com os documentos em que o termo solicitado existe. Este termo não é necessariamente único, podendo ser usado uma combinação de termos, visto o Terrier possibilitar várias formas de pesquisa.

Da lista retornada, encontra-se associada informação relativa à localização do documento aquando da sua indexação, nome, tamanho, extensão e pontuação atribuída. Esta pontuação é configurável e existe um vasta gama de medidas de avaliação disponíveis para uso. Branco optou por utilizar a medida *tf-idf*, que é uma das mais utilizadas.

O valor *tf*, denominado por *term frequency*, corresponde ao número de vezes que o termo surge no documento avaliado, e é normalizado,em função do tamanho do documento, por forma a que seja valorizada a quantidade de vezes que o termo surge no texto, em comparação ao tamanho do próprio. A sua fórmula é:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.1}$$

onde, $n_{i,j}$ é o número de ocorrências do termo i no documento d_j.

O valor *idf*, denominado de *inverse document frequency*, é o valor que permite a discriminação de um termo relativamente à colecção indexada. A fórmula para o cálculo deste valor é:

$$idf_i = \log \frac{|D|}{|d_j : t_i \in d_j|} \tag{3.2}$$

onde, $|D|$ é o número de documentos na colecção e $|d_j : t_i \in d_j|$ o número de documentos onde o termo surge.

Em seguida, a classificação $tf - idf$ foi combinada com a classificação atribuída pelo h-index produzindo uma nova lista ordenada de resultados, por forma a ser avaliada a sua ordenação.

Depois do retorno da lista e feita a escolha do utilizador, a aplicação registava os resultados num ficheiro segundo um formato para esse propósito.

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

A fórmula utilizada pela plataforma para atribuir a medida à pesquisa é dada por:

$$P_t * P_1 + C_m * (1 - P_1) \tag{3.3}$$

P_t representa a pontuação $tf - idf$, o P_1 representa um peso e C_m representa a classificação atribuída a cada um dos blogues segundo a medida utilizada.

Por fim, Branco extraiu os resultados para cada peso e medida, para cada um calculou o valor médio para a posição na ordenação clicada pelos utilizadores. Representou estes resultados num gráfico e concluiu que as medidas usadas para medir a produtividade e importância de cientistas podem de facto ser adaptadas com sucesso nos blogues visto que o h-index apresentava uma curva onde para pesos mais altos apresentava piores resultados, pelo facto de a relevância dos resultados se reduzir no caso da pontuação devolvida pelo Terrier não ser considerada, e para pesos baixos apresentava melhores resultados.

Capítulo 4

Aplicação do h-index à blogosfera

Neste trabalho, pretende-se avaliar a medida h-index na ordenação de resultados na blogosfera. Da mesma forma que Branco [5] teve a necessidade de relacionar as características típicas de um blogue com a produção científica de indivíduos, neste trabalho irá ser usado o mesmo paralelismo que se mostrou ser inovador e com resultados positivos.

Este paralelismo, base do trabalho a desenvolver, associa o blogue, na sua totalidade de entradas ao cientista e as entradas do blogue representam os artigos publicados. Assim sendo, as hiperligações para o blogue em questão seriam o correspondente às citações dos documentos científicos e estimam-se contabilizando o número de vezes em que uma entrada é referenciada em todos os outros blogues da colecção. A Tabela 4.1 resume esta correspondência.

Tabela 4.1: Paralelismo entre blogosfera e a comunidade científica.

Comunidade científica	**Blogosfera**
Cientista	Blogue
Artigos científicos	Entradas no blogue
Citações	Hiperligações

4.1 Experiências

O objectivo das experiências a realizar é avaliar a medida h-index na ordenação de resultados na blogosfera. Este tipo de avaliações têm exigências base que devem de ser cumpridas:

- Reconhecer os cuidados básicos em experiências com utilizadores reais;
- Ter disponível uma colecção de blogues para fins de investigação;
- Definir o número de utilizadores;
- Definir o motor de pesquisa e a interface gráfica a implementar por forma a proporcionar o ambiente experimental apropriado aos utilizadores de teste;

- Definir as medidas de referência a implementar para que seja possível comparar o h-index;
- Definir como irão ser atribuídos os pesos das medidas por forma a estes valores serem comparáveis.

Após cumprir estas exigências é necessário planear a experiência. Neste trabalho a experiência foi dividida em 2 partes. A primeira parte consistiu numa experiência de ordenação estática, que permitia ao utilizador escolher a sua lista ordenada de referência. A segunda parte representava uma pesquisa simples em que o utilizador introduzia uma interrogação para pesquisa e posteriormente seleccionava o resultado que se mostrasse mais relevante para a sua pesquisa.

4.1.1 Exigências base

Uma crítica que é feita aos motores de pesquisa é que, na resposta a uma interrogação, a maior parte das vezes estes retornam sempre os mesmo resultados para qualquer tipo de utilizador [11]. De facto, a maior parte das interrogações são pequenas e ambíguas. No entanto, nem todos os utilizadores têm a mesma necessidade e objectivos com a mesma interrogação. Existe uma solução para resolver este problema, como o mencionado por Dou et al. [11], que defende a pesquisa personalizada. Para a experiência actual, não existe qualquer interesse em personalizar a pesquisa visto que pretendemos avaliar a eficácia da ordenação de blogues da medida h-index e não avaliar métodos para a melhorar. No entanto, esta questão é pertinente, visto que os resultados retornados por esta medida podem ser do agrado de muitos ou poucos utilizadores. A solução passará por definir um número de utilizadores que permita fazer uma avaliação global desta medida, sem ser influenciada por preferências particulares destes.

Pretende-se ultrapassar o número de utilizadores definido por Branco [5] definindo, como objectivo mínimo, atingir 60 utilizadores. A estes utilizadores deve ser proporcionado uma experiência o mais real possível e sem grandes alterações até ao final da mesma. Kohavi et al. [12] mostram que:

- Pequenas modificações na interface gráfica, podem resultar em diferenças significativas nas escolhas por parte dos utilizadores;
- Os utilizadores, quando há variantes nas respostas apresentadas, não devem de ter conhecimento sobre que variante da experiência lhes foi atribuído;
- Deve ser garantido que, até ao final da experiência, cada utilizador deverá manter a sua variante.

4.2 Arquitectura do sistema

A arquitectura do sistema que suporta a experiência pode ser visualizada na figura 4.1, que mostra que o sistema está dividido em 4 partes: *Indexer*, *Retrieval*, *Retrieval Interface* e *System Register*. As duas primeiras partes (*Indexar* e *Retrieval*) são desenvolvidas em *Java*. O *Retrieval Interface* é desenvolvido em *JSP* e *Javascript*. Por último temos o *System Register* que é a componente que regista a informação dos utilizadores.

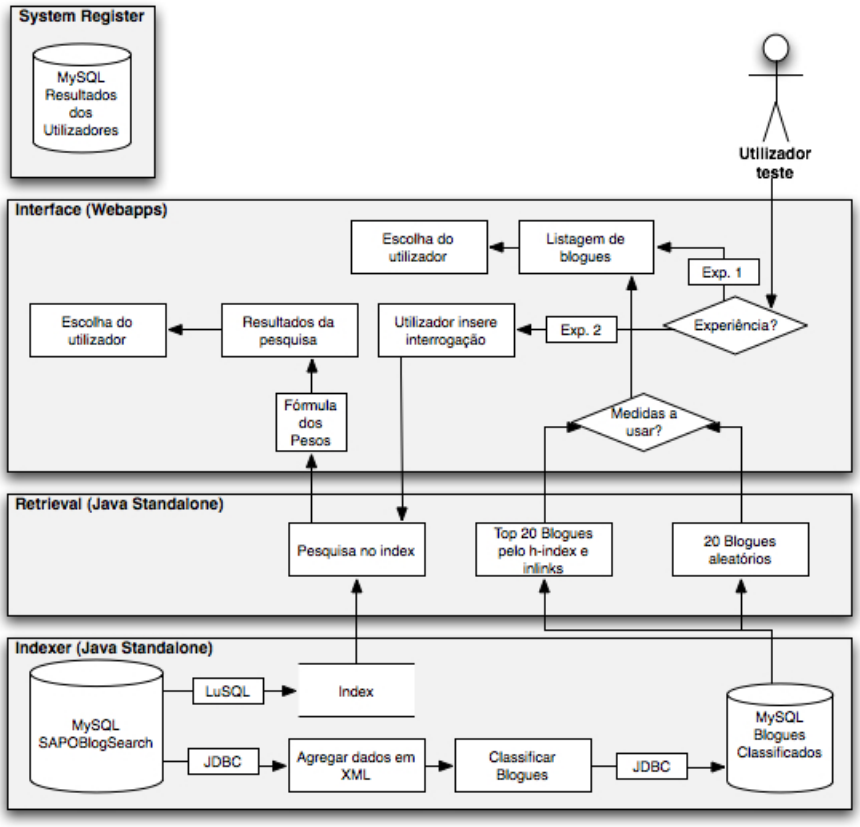

Figura 4.1: Arquitectura do sistema de suporte à experiência

Na figura 4.1 encontram-se esquematizadas as duas experiências realizadas. No entanto, é importante salientar que não foram feitas em simultâneo.

A colecção usada neste estudo foi disponibilizada pela SAPO e tem cerca de 60 000 blogues e 3,5 milhões de entradas [2].

A informação das entradas dos blogues encontra-se armazenada numa base de dados em *MySQL*[1]. A plataforma utilizada para indexação e pesquisa, o *Lucene* que é detalhado mais à frente neste relatório, consegue trabalhar directamente sobre informação proveniente do *MySQL*,

[1] http://www.mysql.com/

visto que tem bibliotecas que permitem agregar os dados existentes na base de dados e transformá-los para ficheiros *Document*, construindo um índice que torna possível ao *Lucene* trabalhar sobre eles.

Na indexação, são guardadas as seguintes informações para cada entrada:

- **Basename:** Nome do blogue;
- **URL:** Endereço de internet da entrada do blogue;
- **Title:** Título da entrada;
- **Postdate:** Data em que a respectiva entrada foi criada;
- **Body:** Conteúdo da entrada;
- **Author:** Autor da entrada.

Não consideramos os dados *description*, *checksum*, *state* e *img* pois não eram relevantes para a análise.

Tendo definido o índice, a questão seguinte são as medidas para efectuar a pesquisa no índice. É necessário definir qual a medida extra a implementar para ser possível compará-la com o h-index. A medida escolhida foi a contagem de citações, o *inlinks*.

Na segunda experiência implementada, é aplicada a metodologia do Teste A/B. O Teste A/B foi utilizado por Kohavi et al. [12], e consiste em atribuir diferentes variantes, ou seja diferentes medidas para o mesmo método, a cada utilizador, a de Controlo (*inlinks*) e Tratamento (h-index). A cada resultado da experiência era avaliado a posição da escolha feita do utilizador.

Segundo Kohavi et al. a implementação de uma experiência numa página Web envolve duas componentes:

- **Algoritmo aleatório:** função que escolhe as variantes e apresenta aos utilizadores;
- **Método de atribuição:** usa o *output* do algoritmo aleatório para determinar a experiência que cada utilizador irá ver no seu *browser*.

São estas duas componentes que permitem dividir os utilizadores para cada uma das duas variantes, Controlo e Tratamento.

O **algoritmo aleatório** tem 3 propriedades:

- É igualmente provável aos utilizadores verem cada variante da experiência, ou seja, assume-se uma separação de 50% - 50%;
- A repetição de atribuições a um único utilizador terá que ser consistente. Ou seja, é atribuído a mesma variante a um utilizador em cada visita à página Web;
- Não deve haver correlação entre experiências. A atribuição de um utilizador a uma variante numa experiência não deve ter qualquer efeito sobre a probabilidade de ser atribuído a uma variante de qualquer outra experiência;

Desta 3 propriedades, a única que não foi aplicada foi a segunda. Optou-se por a um utilizador que repete uma experiência possam lhe ser atribuídas diferentes variantes, podendo assim avaliar a resposta do utilizador para as duas medidas.

O **método de atribuição** activa a página Web da experiência e executa diferentes caminhos de código para diferentes utilizadores. Existe múltiplas formas de implementação deste método [12]. Estas são: Divisão de tráfego, Selecção do lado do servidor e selecção do lado do cliente.

No actual trabalho escolhemos a selecção do lado do servidor, ou seja, a *API* incorporada na página Web dos servidores. Este invoca o algoritmo aleatório e permite uma ramificação lógica, que produz uma experiência diferente ao utilizador em cada variante.

4.3 Implementação

4.3.1 Apache Lucene

O *Apache Lucene* é o motor de pesquisa escolhido. É um motor de pesquisa *open source* e foi desenvolvido no projecto *Apache Jakarta*[2] por Doug Cutting, inteiramente em *Java*. É usado para indexar e pesquisar colecções[13]. A medida utilizada pelo *Lucene* para a indexação de colecções é *tf-idf*, tal como referido em 3.3.4. Esta medida atribui a uma palavra do documento um peso que é proporcional ao número de ocorrências dessa palavra no respectivo documento e inversamente proporcional ao número de documentos na colecção onde a palavra ocorre pelo menos uma vez.

O *Lucene* pode indexar qualquer tipo de informação baseada em texto e permite que esta informação seja posteriormente pesquisada sobre alguns critérios que o utilizador pretenda definir. Apesar de o *Lucene* ter sido desenvolvido para funcionar somente em texto, actualmente existem aplicações extras que permite indexar documentos *Word*, ficheiros *PDF*, *XML* ou mesmo páginas *HTML*, extraindo os respectivos textos. Uma das grandes razões para a popularidade do *Lucene* é a sua simplicidade[14]. Mas, apesar desta simplicidade, o *Lucene* tem no seu interior técnicas de recuperação de informação sofisticadas. Não é necessário um conhecimento aprofundado sobre como funciona o *Lucene*, a fim de começar a usá-lo, visto que a *API* do *Lucene* requer simplesmente que um utilizador aprenda a usar apenas um pequeno subconjunto das classes que o constituem.

Apesar de existirem alternativas, nomeadamente o motor utilizado por Branco [5], o *Terrier*[3], as razões enunciados são os factores que levaram a escolher este motor de pesquisa.

4.3.2 Interface gráfica

Há necessidade de implementar uma interface gráfica para disponibilizar aos utilizadores os resultados do motor de pesquisa. A interface encontra-se desenvolvida em *JSP*, com uma estrutura muito semelhante à de motores de pesquisa conhecidos. A escolha de desenvolver em *JSP* é o facto

[2] http://jakarta.apache.org/
[3] http://ir.dcs.gla.ac.uk/terrier/

da linguagem *Java* ser madura e em constante actualização. Visto que o *Lucene* é desenvolvido em *Java*, foi fácil a integração das componentes.

4.3.3 Medidas de ordenação

Como já foi mencionado, as medidas implementadas para este projecto foram o h-index, medida alvo de avaliação, e o *inlinks*, medida usada para comparação. A informação das entradas dos blogues encontra-se armazenada numa base de dados em *MySQL*. A implementação das medidas directamente sobre a base de dados é um processo pouco eficiente. Assim, a informação foi extraída para o formato *XML* apresentado na Figura 4.2, onde o campo *url* guarda o endereço do blogue, *basename* guarda o nome do blogue e o campo *links* guarda os endereços que são citados no endereço em análise. O procedimento para efectuar esta extracção, pode ser ilustrado na Figura 4.3.

```
<?xml version="1.1" encoding="UTF-8"?>
<entries>
    <entry>
        <url>http://14demarco.blogs.sapo.pt/25436.html</url>
        <basename>14demarco.blogs.sapo.pt</basename>
        <links>[www.new7wonders.com/index.php, www.7maravilhas.sapo.pt/index.html]</links>
    </entry>
</entries>
```

Figura 4.2: Formato *XML* relativo às entradas de cada blogue da colecção.

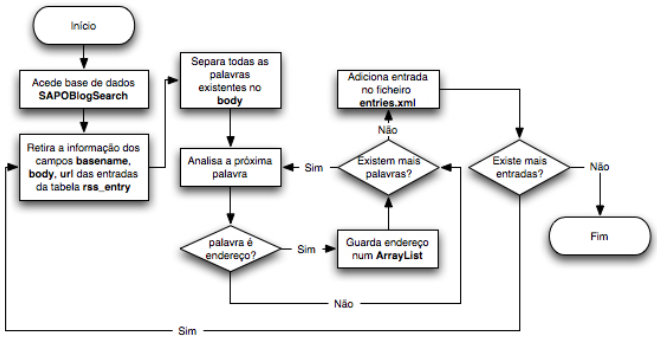

Figura 4.3: Extracção da informação das entradas, existente na base de dados.

Este programa acede a cada entrada existente na colecção, arquivada numa base de dados em *MySQL* (denominada na figura de *SAPOBlogSearch*) e extrai a seguinte informação:

- O endereço da entrada em análise (*url*);
- O nome do blogue da respectiva entrada (*basename*);
- Os endereços que esta entrada cita no seu conteúdo (*links*).

Para obter os endereços que são citados em cada entrada existente na colecção, existe um *parser* que acede ao conteúdo de cada entrada e detecta se existe algum endereço.

Por forma a haver comunicação entre o *Java* e a base de dados *MySQL*, foi usado a API *Java Database Connectivity* (JDBC[4]).

Após correr o programa de extracção das entradas relativas a cada blogue, foi implementado um outro programa, em *Java*, que extrai os endereços que são citados na colecção, conta o número de vezes que estes eram citados e guarda este informação no formato XML 4.4. O algoritmo para este programa é ilustrado na Figura 4.5.

```
<?xml version="1.1" encoding="UTF-8"?>
<links>
    <link>
        <url>http://xl.sapo.pt/mframe_video.html?cid=Xz0003&amp;mid=1&amp;amp;amp;arid=482455&amp;p=index.html</url>
        <basename>http://xl.sapo.pt</basename>
        <citation>1</citation>
    </link>
</links>
```

Figura 4.4: Formato *XML* relativo ao número de vezes que cada endereço é citado.

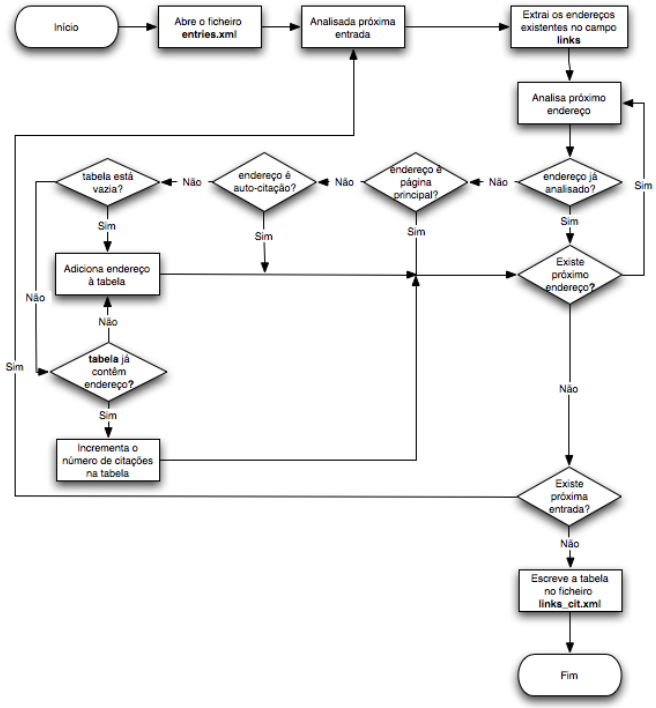

Figura 4.5: Programa criado para extrair a informação relativa a cada endereço que é citado.

É de salientar que este programa verifica se o endereço citado é uma auto-citação ou não. Pela figura 4.5 podemos verificar que o número de citações deste endereço não é incrementado, garantindo que se ignoram as auto-citações no cálculo dos valores de classificação de cada blogue.

Este programa acede a cada entrada, já guardada no ficheiro *XML*, e verifica se a entrada já foi analisada, se é uma página principal e se é auto-citação. Se não se verificar nenhuma destas interrogações, o programa contabiliza essa entrada como uma citação válida. No final, todas as citações que cada entrada tem foram registadas e são guardadas num ficheiro *XML* 4.4.

[4]http://dev.mysql.com/downloads/connector/j/3.0.html

Em seguida foram implementados programas que calculam os valores *inlink* e h-index de cada blogue. Ambos foram implementados em *Java* e os seu algoritmos encontram-se ilustrados nas figuras 4.6 e 4.7.

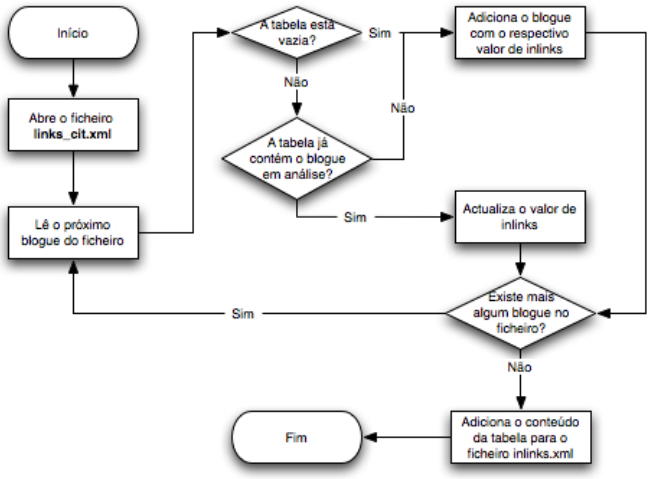

Figura 4.6: Programa criado para calcular o valor *inlink* relativo a cada blogue.

O programa criado para obter a classificação *inlink* de cada blogue começa por aceder ao ficheiro *XML* e a cada entrada que a constitui. Existe um primeiro passo no algoritmo que só ocorre quando a tabela de registo de blogues encontra-se vazia, que adiciona o blogue correspondente à entrada em análise e atribui-lhe o valor de *inlinks* a 1. Nos restantes passos, o programa acede a cada entrada no ficheiro *XML* verifica se o blogue da entrada já se encontra na tabela. Se sim, incrementa o seu valor de *inlinks*, se não adiciona o blogue e atribui-lhe valor 1. No final os valores da tabela são guardados num ficheiro *XML* 4.8.

Para a classificação h-index de cada blogue, o programa 4.7 analisa cada entrada no ficheiro *XML* e adiciona a informação relativa ao endereço da entrada e o número de vezes que ela é citada numa tabela. Em seguida, as entradas nessa tabela são agrupadas pelo nome do blogue que as constitui e ordenadas pelo número de citações. Cada grupo de entradas é analisado e sempre que se verificar que o número de citações é maior ou igual à posição da respectiva entrada, adiciona o blogue que constitui a entrada com o valor da posição somada com 1, que é o seu valor de h-index.

Após cada blogue ser classificado, são guardados os seus valores num ficheiro *XML* 4.9.

Por fim, os valores de *inlinks* e h-index de cada blogue são introduzidos numa tabela na base de dados em *MySQL*, para futuramente ser utilizado pela interface apresentada ao utilizador.

Devido à quantidade de dados a analisar, foi necessário a implementação de programas independentes e, consecutivamente, ficheiros intermédios.

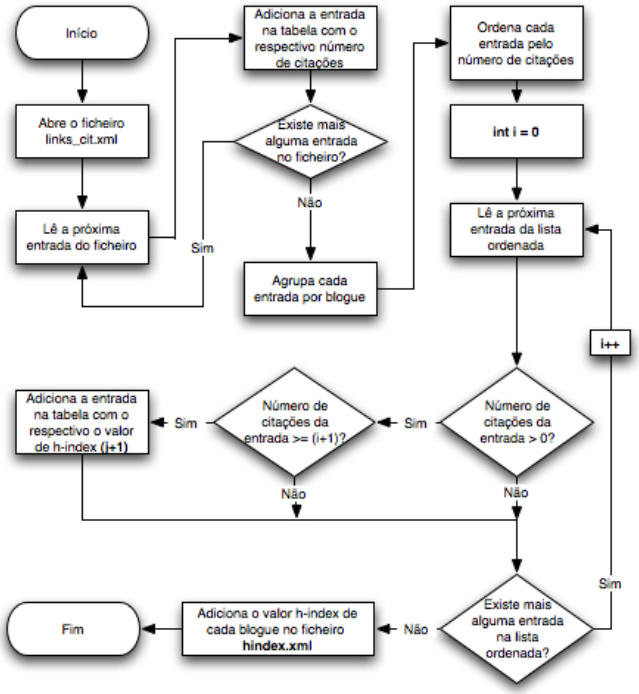

Figura 4.7: Programa criado para calcular o valor h-index relativo a cada blogue.

```
<?xml version="1.1" encoding="UTF-8"?>
<inlink>
    <inlinks>
        <basename>iphonejtag.blogspot.com</basename>
        <value>3</value>
    </inlinks>
</inlink>
```

Figura 4.8: Formato *XML* relativo ao valor *inlink* de cada blogue.

```
<?xml version="1.1" encoding="UTF-8"?>
<hindexs>
    <hindex>
        <basename>iphonejtag.blogspot.com</basename>
        <value>1</value>
    </hindex>
</hindexs>
```

Figura 4.9: Formato *XML* relativo ao valor h-index de cada blogue.

4.4 1ªExperiência - Experiência de ordenação estática

Esta primeira experiência foi implementada para recolher a opinião dos utilizadores perante duas listas de classificação de blogues segundo duas medidas de ordenação. Com esta experiência pretendia-se avaliar a medida h-index pelos utilizadores.

A interface disponibiliza ao utilizador duas listas de classificações de blogues, cada uma com 10 blogues e ordenadas segundo diferentes medidas. De cada vez que um utilizador acede a esta experiência, obtém 2 listas que podem ser uma de 3 combinações:

1. Uma lista ordenada aleatoriamente e outra ordenada pela medida *inlinks*;
2. Uma lista ordenada aleatoriamente e outra ordenada pela medida h-index;
3. Uma lista ordenada pela medida *inlinks* e outra ordenada pela medida h-index.

Em qualquer um dos casos, o utilizador não sabe qual das medidas está a ser usada e as respectivas classificações que lhes eram atribuídas. A interface disponível ao utilizador é a da Figura 4.10.

Figura 4.10: Interface gráfica da experiência 1.

Após o utilizador fazer a sua escolha, esta é guardada numa tabela criada na base de dados *MySQL* com a informação relativa às medidas que foram disponibilizadas e a escolha feita pelo utilizador para posteriormente serem analisados. Um excerto desta tabela é apresentada na Figura 4.11.

feature_left	feature_right	feature_selected	data
random	inlinks	random	Tue Jun 02 16:43:51 WEST 2009
hindex	random	random	Tue Jun 02 16:43:57 WEST 2009
random	hindex	random	Tue Jun 02 16:44:01 WEST 2009
inlinks	hindex	inlinks	Tue Jun 02 16:44:04 WEST 2009
random	hindex	hindex	Tue Jun 02 16:45:11 WEST 2009
inlinks	random	inlinks	Tue Jun 02 17:26:55 WEST 2009
inlinks	hindex	hindex	Tue Jun 02 17:27:00 WEST 2009
inlinks	hindex	hindex	Tue Jun 02 17:27:03 WEST 2009
inlinks	hindex	hindex	Tue Jun 02 17:27:05 WEST 2009
random	hindex	random	Tue Jun 02 17:27:08 WEST 2009

Figura 4.11: Exemplo de como os dados são guardados em SQL.

4.5 2ª Experiência - Experiência de pesquisa simples

Nesta experiência, era pedido ao utilizador que formulasse uma interrogação e escolhesse 1 dos resultados que lhe era devolvido.

A resposta à interrogação efectuada pelo utilizador, não tinha sempre a mesma medida a ordenar a lista, podendo ser *inlinks* ou h-index. Para tal, a aplicação Web, ao ser-lhe inserido a interrogação, atribuía uma medida à pesquisa, calculava a pontuação final de cada resultado e finalmente os resultados eram ordenados e apresentados ao utilizador. Este procedimento pode ser visualizado na Figura 4.12. Para o calculo da pontuação final, era usado a fórmula de pesos, descrito anteriormente, em que P_1 era seleccionado entre o seguinte conjunto: $0;0,2;0,4;0,5;0,6;0,8;1$. Este conjunto de valores foi escolhido tendo como base o trabalho anteriormente realizado por Branco.

Para que fosse possível combinar o peso da medida *tf-idf* e a medida atribuída ao utilizador na experiência, foi necessário normalizar as classificações das medidas *inlinks* e h-index, visto não terem as mesmas ordens de grandeza. O processo mais simples encontrado foi a normalização pela posição em que a entrada se encontrava em ambas as listas ordenadas pela medida *tf-idf* e a medida atribuída. Para tal, na fórmula de pesos em vez de utilizamos o valor atribuído à entrada pelas respectivas medidas, utilizamos as posições. Ou seja, relembrando a equação de pesos:

$$P_t * P_1 + C_m * (1 - P_1) \tag{4.1}$$

Em que o valor P_t representa a posição da entrada quando ordenada pela medida *tf-idf* e o C_m a posição da entrada quando ordenada pela medida atribuída ao utilizador. Este raciocínio é demonstrado na Figura 4.12.

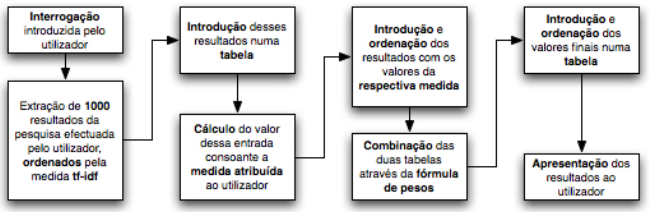

Figura 4.12: Lógica de normalização e ordenação de resultados de pesquisa.

Para melhor compreensão da normalização de resultados, imaginemos que é efectuado uma pesquisa, em que a medida atribuída é o h-index e o peso de *tf-idf* é de 0.2 e do h-index é de 0.8, e obtemos as duas Tabelas 4.2, e 4.3.

Ao aplicar a fórmula dos pesos efectuamos a normalização e obtemos os valores finais de cada entrada. Depois de obter estes valores, as entradas são ordenadas por ordem ascendente, como podemos verificar na Tabele 4.4.

A interface disponível ao utilizador pode ser visualiza na Figura 4.10.

Tabela 4.2: Tabela de resultados ordenados pela medida *tf-idf*.

Posição P_t	**url**
1	http://ablasfemia.blogspot.com/2007/09/no-perder.html
2	http://quartarepublica.blogspot.com/2007/09/parabns.html
3	http://31daarmada.blogs.sapo.pt/1472771.html
4	http://testar.blogs.sapo.pt/429.html

Tabela 4.3: Tabela de resultados ordenados pela medida *h-index*.

Posição C_m	**url**
1	http://quartarepublica.blogspot.com/2007/09/parabns.html
2	http://31daarmada.blogs.sapo.pt/1472771.html
3	http://ablasfemia.blogspot.com/2007/09/no-perder.html
4	http://testar.blogs.sapo.pt/429.html

Tabela 4.4: Tabela de resultados ordenados, após a aplicação da fórmula dos pesos.

Posição	**Equação**	**Resultado**	**url**
1	$2*0.2+1*(1-0.2)$	1.2	http://ablasfemia.blogspot.com/2007/09/no-perder.html
2	$3*0.2+2*(1-0.2)$	2.2	http://31daarmada.blogs.sapo.pt/1472771.html
3	$1*0.2+3*(1-0.2)$	2.6	http://ablasfemia.blogspot.com/2007/09/no-perder.html
4	$4*0.2+4*(1-0.2)$	4	http://testar.blogs.sapo.pt/429.html

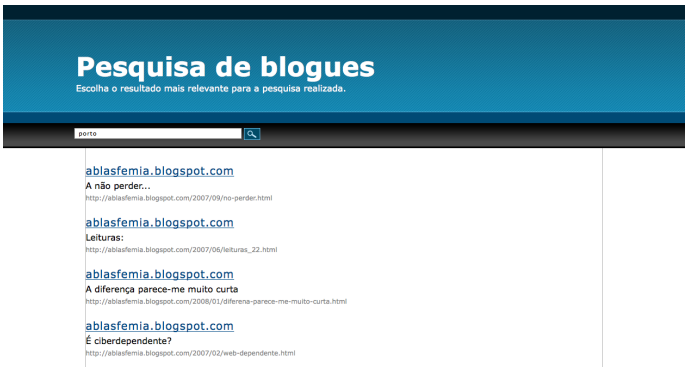

Figura 4.13: Interface gráfica da experiência 2.

Ao seleccionar o resultado considerado relevante pelo utilizador, este é registado numa tabela criada numa base de dados *MySQL* com a informação da interrogação feita, medida utilizada, posição do resultado escolhido, endereço do resultado, peso P_1 atribuído e a data do registo 4.14.

id	query	feature	position	url	weight	date
1	macbook air	inlinks	7	http://oinsurgente.blogspot.com/2006/08/o-iro-e-o-...	0.5	Thu Jun 18 12:07:50 WEST 2009
2	apple	inlinks	6	http://formigabargante.blogspot.com/2007/03/apple-...	0.4	Thu Jun 18 12:08:07 WEST 2009

Figura 4.14: Exemplo de como os dados são guardados em SQL.

Destes dados coleccionados, é calculado o valor médio da posição relativa à escolha do utilizador, consoante a medida utilizada e o peso que lhe é associado. Estes dados vão-nos permitir avaliar o comportamento da medida h-index ao longo da sua combinação com a medida *tf-idf*.

Capítulo 5

Resultados Experimentais

5.1 Experiência 1 - Ordenação estática

Esta experiência consiste na apresentação ao utilizador de duas listagens de blogues onde cada uma é ordenada por uma de três medidas diferentes (*inlinks*, h-index ou ordenação aleatória).

Foi pedido ao utilizador que analisasse cada uma das listas e escolhesse aquela que fosse da sua preferência. Para esta analise, o utilizador tinha a possibilidade de aceder a cada um dos blogues existentes nas listas. No entanto, o utilizador não tinha qualquer influência na criação das mesmas, tornando-o desconhecedor das medidas aplicadas para a ordenação.

Esta experiência contou com a participação de 67 utilizadores anónimos. Os resultados desta experiência podem ser observados na Figura 5.1. As colunas representam as percentagens de escolha dos utilizadores, para dada combinação possível de ordenações.

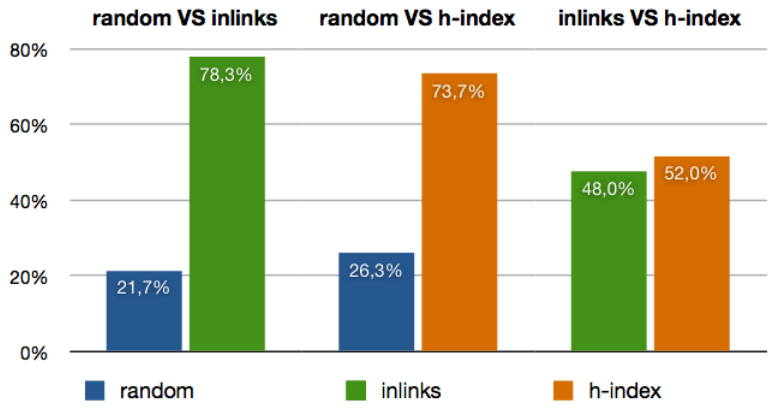

Figura 5.1: Resultados obtidos na experiência 1.

A introdução da lista de ordenação aleatória teve como finalidade validar a metodologia e implementação da experiência. Os resultados obtidos para a lista produzida por esta ordenação foram muito baixos, o que nos leva a concluir que a experiência foi válida.

Todas as vezes que apresentávamos uma listagem ordenada aleatoriamente a um utilizador, os resultados eram diferentes. Os valores baixos que a listagem ordenada aleatoriamente obteve mostram que os utilizadores têm alguma familiaridade com a blogosfera portuguesa.

Quando confrontados com as listas ordenadas pelas medidas *inlinks* e h-index, os utilizadores optaram pela lista ordenada pelo h-index como sendo uma lista representativa dos 10 melhores blogues portugueses. No entanto, esta opinião não é significativa, a diferença entre o h-index e o *inlinks* é de 4%.

Existem muitas variáveis nesta experiência que não são controláveis, como os interesses e o gosto dos utilizadores. É necessário criar uma experiência que possibilite ao utilizador introduzir um tema à sua escolha. Mas, como já foi referido anteriormente, para uma primeira análise, os resultados mostraram-se interessantes, mesmo não podendo controlar estas variáveis o utilizador achou a lista produzida pelo h-index como sendo representativa do top 10 blogues.

5.2 Experiência 2 - Pesquisa Simples

Como já foi referido anteriormente, nesta experiência foi pedido a cada utilizador que efectuasse uma pesquisa livre e, mediante os resultados que lhe eram retornados, escolhesse aquele que se adequava melhor à pesquisa realizada. Ao contrário da experiência 1, o utilizador não tinha a possibilidade de visualizar a página antes de a escolher.

Esta experiência contou com a participação de 153 utilizadores anónimos cujos resultados podem ser observados na Figura 5.2. Estes resultados preliminares requerem ainda validação ao nível da implementação e recolha de dados.

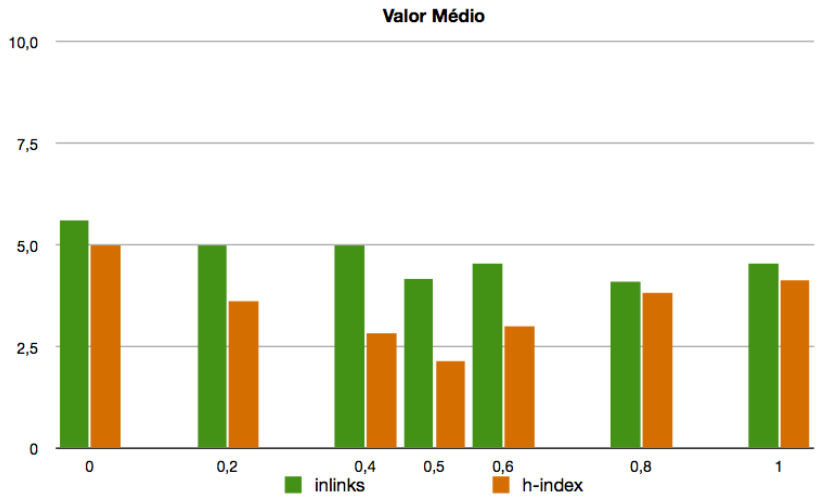

Figura 5.2: Resultados obtidos na experiência 2.

O gráfico da Figura 5.2 apresenta o valor médio das posições dos resultados escolhidos pelo utilizador, mediante a medida utilizada e o peso atribuído, representado no eixo dos x. Relembrando a fórmula de pesos:

$$P_t * P_1 + C_m * (1 - P_1)$$

onde P_t representa a pontuação nativa do *Lucene*, *tf-idf*, P_1 representa o peso a atribuir a cada medida e C_m representa a classificação da medida atribuída para a pesquisa.

Os resultados obtidos não foram o esperado. Esta experiência foi repetida mais do que uma vez, sempre com alterações na metodologia da mesma de modo a corrigir eventuais erros. Analisando a equação da fórmula dos pesos, podemos verificar que, quando era atribuído o peso a P_1 de 0, a lista obtida era ordenada somente pela medida atribuída à pesquisa, *inlinks* ou h-index. Para o peso 1, a lista era ordenada pela medida nativa do *Lucene*, *tf-idf*. Estes resultados são muito próximos entre eles, o que não era esperado, uma vez que a medida *tf-idf* é uma medida robusta, o que deveria de produzir melhores resultados do que uma medida, por assim dizer, "cega"que não tem um grande número de variáveis para a sua formulação.

Estes resultados podem ter sido provocados por má implementação da metodologia. Um dos problemas analisados centrava-se no facto de algumas pesquisa produzirem resultados em que as primeiras posições apresentadas eram referentes ao mesmo blogue. Assim sendo, os resultados tornam-se falseados e de certo modo ambíguos, uma vez que, considerando que numa determinada pesquisa um blogue ocupava as cinco primeiras posições nos resultados de pesquisa e o utilizador escolhesse a quarta posição daria ao blogue uma classificação que não era correcta visto que este também estava a ocupar o primeiro lugar da classificação.

Na experiência 1, verificamos que quando o utilizador foi confrontado com duas listas em que uma era ordenada pela medida *inlinks* e a outra pelo h-index, tinha uma ligeira preferência pelo h-index. Esta preferência reflecte-se no resultado obtido ao analisarmos o gráfico da Figura 5.2 para valores de peso 0, onde verificamos que o *inlinks* teve piores resultados que o h-index. Esta situação é suportada pela análise dos restantes resultados, onde o h-index obteve sempre melhores resultados do que o *inlinks*.

À medida que aumentamos o valor do peso, estamos a aumentar a influência que a medida *tf-idf* têm na classificação. Podemos verificar que o h-index tem uma variação parabólica à medida que o peso aumenta, onde obtém o seu valor mais baixo com o peso 0.5 e o mais alto em 0.

Capítulo 6

Conclusão

O presente trabalho foi conduzido num período de 2 semestres lectivos, tendo sido realizadas pesquisas sobre as áreas associadas ao tema no primeiro e segundo semestre. No segundo semestre foram planeadas e implementadas duas experiências que nos permitiram obter resultados para avaliar a medida h-index na ordenação de resultados na blogosfera. Todavia existe matéria para ser explorada na avaliação desta medida.

6.1 Discussão

Tomando os resultados apresentados no Capítulo 5, podemos concluir que a medida h-index mostra um bom desempenho e oferece ordenações diferentes das que são realizadas pelo *inlinks*. Verificamos assim um potencial desta medida para ser usada na ordenação de resultados na blogosfera.

As experiências foram realizadas em ambientes abertos, isto é, sem qualquer conhecimento do perfil do utilizador, o que torna a experiência mais próxima da realidade. Esta opção permitiu-nos verificar que o h-index obtém resultados semelhantes àqueles obtidos com a medida inlinks. Este resultados tornam o h-index numa medida potencial para uso na ordenação de resultados.

6.2 Trabalho Futuro

Os resultados obtidos mostram a medida h-index como promissora. Contudo, deveriam ser efectuados mais testes, mas em experiências fechadas que permitissem analisar o perfil do utilizador de teste. Este factor é importante, visto que a blogosfera ainda é uma realidade que só alguns utilizadores conhecem e em pequena escala.

Na experiência 1, foram efectuados testes com ordenações de 1 a 10 na classificação e sempre com medidas diferentes em comparação. Seria interessante introduzir lista ordenadas, mas noutro intervalo de classificação, como por exemplo de 10 a 20, e colocá-la em comparação com a ordenação de 1 a 10. Ambas as listas com a mesma medida a ordenar.

A normalização efectuada para a comparação das duas medidas foi efectuada pela ordem que cada entrada mostrava na classificação. Deveria ser efectuada uma normalização pela classificação que cada medida atribuí à entrada, permitindo uma maior analise estatística.

Ao retribuir resultados ordenados pelo h-index ou *inlinks*, o programa pedia ao *Lucene* que devolva os resultados ordenados pelo *tf-idf* e só depois é que era calculado os respectivos valores para a medida em causa. Este processo poderia ser optimizado aquando da criação do índice. Para tal, na criação do índice devia ser agregado um campo com a respectiva classificação de cada entrada da colecção, segundo as medidas a serem usadas.

Por último, deveriam ser eliminado da colecção todos os blogues que actualmente já não existem por forma a tornar esta experiência mais realista e actual.

Anexo A

Anexos

A.1 Anexo 1

Tabela A.1: Top20 blogues ordenadas pela medida h-index

Posição	Nome do blogue	Valor
1	famosas-celebridades.blogspot.com	21
2	doiscliques.blogs.sapo.pt	12
3	7maravilhasdevilavicosa.blogspot.com	10
4	blogueforanada.blogspot.com	9
5	leiriaaminhacidade.blogs.sapo.pt	9
6	abrupto.blogspot.com	8
7	causa-nossa.blogspot.com	8
8	arrastao.weblog.com.pt	8
9	rlx.blogs.sapo.pt	8
10	origemdasespecies.blogspot.com	7
11	fotos_ichliebeth.blogs.sapo.pt	7
12	gloriafacil.blogspot.com	6
13	origemdasespecies.blogs.sapo.pt	6
14	www.gardenal.org	6
15	aba-da-causa.blogspot.com	6
16	ablasfemia.blogspot.com	6
17	daliteratura.blogspot.com	6
18	doportugalprofundo.blogspot.com	5
19	avenida-dos-aliados-porto.blogspot.com	5
20	ansiaonews.blogs.sapo.pt	5

A.2 Anexo 2

Tabela A.2: Top20 blogues ordenadas pela medida inlinks

Posição	**Nome do blogue**	**Valor**
1	ablasfemia.blogspot.com	1902
2	elvirabistrot.blogspot.com	1822
3	famosas-celebridades.blogspot.com	1380
4	causa-nossa.blogspot.com	1277
5	daliteratura.blogspot.com	1138
6	sombarato.blogspot.com	1063
7	abrupto.blogspot.com	1059
8	arrastao.weblog.com.pt	961
9	blogueforanada.blogspot.com	876
10	portugaldospequeninos.blogspot.com	868
11	5dias.net	816
12	31daarmada.blogs.sapo.pt	781
13	corta-fitas.blogspot.com	749
14	origemdasespecies.blogspot.com	664
15	fotos_ichliebeth.blogs.sapo.pt	622
16	doiscliques.blogs.sapo.pt	610
17	incursoes.blogspot.com	571
18	estadocivil.blogspot.com	533
19	gloriafacil.blogspot.com	488
20	blog.uncovering.org	437

Referências

[1] Jorge E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569, 2005.

[2] Telmo Couto. Characterizing the portuguese blogosphere. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, 2009.

[3] Sergey; Motwani Rajeev Page, Lawrence; Brin and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1998. Previous number = SIDL-WP-1999-0120.

[4] Martha Sideri Apostolos Kritikopoulos and Iraklis 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*, page 8, New York, NY, USA, 2006. ACM.

[5] José Mário Castelo Branco. Aplicação do h-index em blogues. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Julho 2008.

[6] Diane J. Schiano, Bonnie A. Nardi, Michelle Gumbrecht, and Luke Swartz. Blogging by the rest of us. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1143–1146, New York, NY, USA, 2004. ACM.

[7] Donald Metzler e Trevor Strohman W. Bruce Croft. *Search Engines: Information Retrieval in Pratice*. Addison Wesley, 2009.

[8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[9] Trystan Upstill. Predicting fame and fortune: Pagerank or indegree. In *In Proceedings of the Australasian Document Computing Symposium, ADCS2003*, pages 31–40, 2003.

[10] Google toolbar, http://www.google.com/tools/firefox/toolbar/ft5/intl/pt-pt/index.html, 14 de janeiro de 2009, 2008.

[11] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 581–590, New York, NY, USA, 2007. ACM.

[12] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967, New York, NY, USA, 2007. ACM.

[13] Thomas Paul. The lucene search engine, www.javaranch.com/journal/2004/04/lucene.html, 20 de maio de 2009.

[14] Otis Gospodnetić Erik Hatcher. *Lucene in Action*. Manning Publications, second edition edition, 2008.