3
Caracterização do Repositório Experimental

Devido ao alto custo para reproduzir o ambiente computacional das atuais máquinas de busca, a simulação tem sido empregada na literatura para avaliar as políticas de revisitação (Cho03a, Wol02, Eck08). Para que a simulação tenha validade, o simulador deve reproduzir aspectos relevantes do ambiente, como as modificações das páginas, a capacidade do canal de comunicação, e a concentração de páginas por servidor.

De modo a tornar a simulação mais próxima da realidade, a parametrização do simulador deve ser baseada nas características de um repositório contendo páginas *Web* "representativas" daquelas encontradas nos repositórios de máquinas de busca. Para este propósito construímos o repositório WEBBASE, cujo processo de construção e principais características são apresentados a seguir. As características do repositório WEBBASE são utilizadas nas simulações ao longo da tese. Um repositório com artigos da Wikipedia (Wik07) é utilizado nos experimentos do Capítulo 6.

A caracterização do repositório WEBBASE leva em conta os aspectos abaixo:

Concentração de páginas por servidor. Servidores com mais páginas têm maior chance de serem afetados pela restrição de *politeness*. Se grande parte das páginas estão hospedadas em poucos servidores, então espera-se que a restrição de *politeness* tenha grande impacto no *freshness* do repositório.

Teste da Suposição 1.4. A hipótese de que as páginas na Internet se modificam segundo um processo de Poisson é utilizada nesta tese e em outros trabalhos na literatura. Portanto, é importante verificar a validade desta hipótese no repositório WEBBASE.

Distribuição das taxas de modificação. Quando as páginas são modificadas segundo um processo de Poisson, as taxas de modificação são suficientes para caracterizar as modificações das páginas.

3.1 Construção e Monitoramento

O repositório WEBBASE é composto por páginas coletadas a partir de uma amostra aleatória de URLs dentre aquelas disponibilizadas no repositório *General Crawl* do Projeto *Stanford WebBase* (Web07). A seleção de páginas para o repositório *General Crawl* não esteve restrita a nenhum tema específico, e portanto este repositório é composto por páginas típicas da Internet. O processo de construção do repositório WEBBASE seguiu os seguintes passos:

1. Seleção de uma amostra aleatória com 5.462 URLs encontradas no repositório *General Crawl* em novembro de 2007.

2. *Download* de até 100.000 páginas partindo de cada URL na amostra, utilizando o *crawler* WIRE (Wir07). Servidores com até 100.000 páginas foram totalmente coletados. Este processo forneceu 14.513.924 páginas para o repositório WEBBASE, hospedadas em 5.462 servidores.

Uma vez construído o repositório WEBBASE, um conjunto de páginas deste repositório foi monitorado para identificar os padrões de modificações das páginas. Este monitoramento seguiu os seguintes passos:

1. Seleção de uma amostra com 363.466 páginas do repositório WEBBASE para monitoramento diário. Para evitar sobrecarga de servidores, o número de páginas monitoradas de um mesmo servidor foi limitado em 100. Assim, para cada servidor com mais de 100 páginas, 100 páginas foram escolhidas aleatoriamente para compor a amostra. Servidores com até 100 páginas tiveram todas as páginas inseridas na amostra.

2. Monitoramento diário das páginas da amostra durante 275 dias (20/04/2008 até 22/01/2009). Uma página foi considerada modificada em um determinado dia quando apresentava alguma diferença com relação à sua cópia do dia anterior que não fosse os caracteres espaço, tabulação ou nova linha.

O monitoramento periódico das páginas é necessário para entender como as páginas se modificam, visto que o histórico de modificações da maioria das páginas na Internet não está disponível. Um aspecto crítico deste monitoramento é a frequência com que vamos revisitar cada página para verificar se ela sofreu modificação. Se o monitoramento é feito em intervalos longos nossas estimativas tornam-se imprecisas, visto que (i) a modificação pode ter ocorrido em qualquer ponto do intervalo e (ii) mais de uma modificação pode ter ocorrido no intervalo. Por outro lado, como a frequência

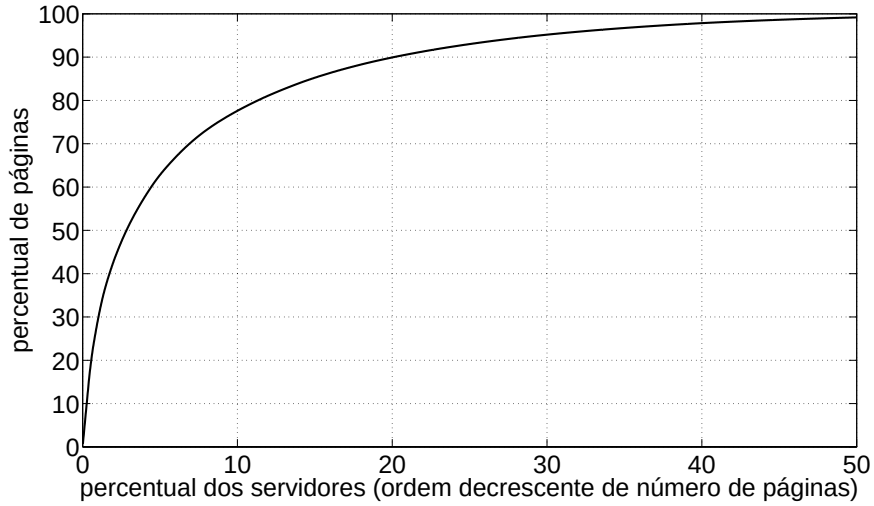

Figura 3.1: Concentração das páginas nos servidores com mais páginas (repositório WEBBASE).

de requisições a um dado servidor *Web* é limitada pela restrição de *politeness*, intervalos curtos de monitoramento restringem bastante a quantidade de páginas de um mesmo servidor que podem ser monitoradas.

Como consequência disso, o monitoramento do repositório WEBBASE possui as seguintes desvantagens: (i) revisitando as páginas uma vez por dia, foi possível monitorar apenas 363.466 páginas (2,5% do repositório); (ii) para 10% das páginas, o servidor *Web* retornou um código de erro em todos os pontos de monitoramento; (iii) 32% das páginas não sofreram modificação durante o período de monitoramento; e (iv) 10% das páginas estavam modificadas em todos os pontos de monitoramento. Um período maior de monitoramento melhora o item (iii), e uma maior frequência de monitoramento melhora o item (iv). Entretanto, frequências altas de monitoramento e longos períodos de monitoramento produzem uma grande quantidade de requisições aos servidores Web, que podem por esta razão bloquear o *crawler*, ignorando requisições futuras.

Um repositório com artigos da Wikipedia não possui estas desvantagens, visto que os instantes de modificação dos artigos estão disponíveis no *Web site* da Wikipedia (Wik07). Desta forma a simulação pode utilizar os instantes reais de modificação, ao invés de gerá-los através de um processo de modificação estimado. Por outro lado, este repositório não pode ser considerado representativo das páginas na Internet. De fato, as modificações dos artigos na Wikipedia não podem ser modeladas por um processo de Poisson (Alm07). A construção, caracterização e experimentos com um repositório de artigos da Wikipedia são apresentados no Capítulo 6.

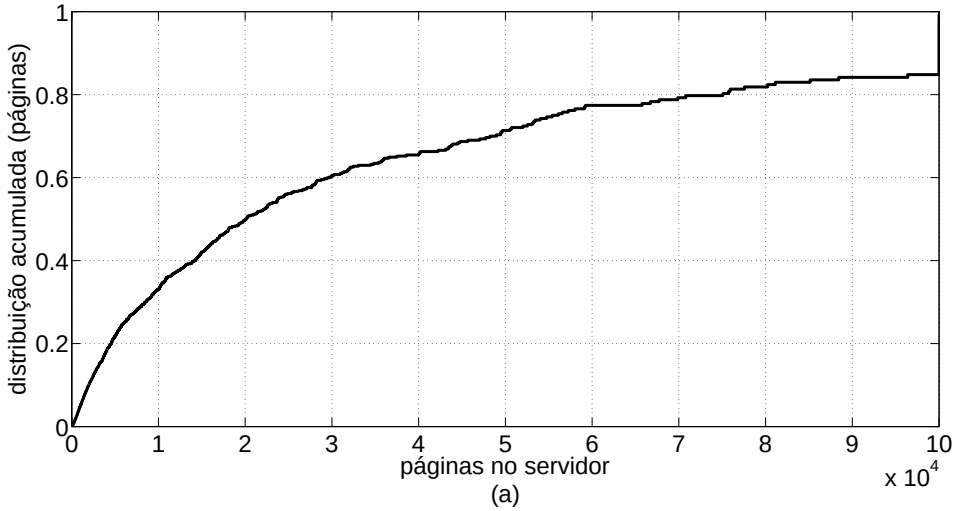

Figura 3.2: Concentração de páginas por servidor. Para cada quantidade x de páginas por servidor, o gráfico fornece a proporção das páginas hospedadas em servidores com até x páginas (repositório WEBBASE).

3.2 Número de Páginas por Servidor

As páginas hospedadas em servidores com muitas páginas podem ser mais afetadas pela restrição de *politeness*. A restrição de *politeness* estabelece um tempo mínimo entre requisições a um mesmo servidor, e portanto restringe a soma das frequências de revisitação das páginas de um mesmo servidor. Além disso, pode ocorrer competição entre as páginas em dada região do escalonamento, visto que a restrição de *politeness* limita o número de requisições feitas em um determinado intervalo de tempo.

Podemos observar na Figura 3.1 que o repositório WEBBASE apresenta alta concentração de páginas em poucos servidores. Para a construção do gráfico da Figura 3.1 os servidores foram ordenados em ordem decrescente de número de páginas, e o percentual de páginas do repositório nos primeiros x servidores foram computados. Por exemplo, podemos observar que cerca de 90% das páginas estão hospedadas em apenas 20% dos servidores. De fato, Castillo et al. (Cas04b) observaram uma lei de potências no número de páginas por servidor nos experimentos realizados com um repositório de páginas chilenas. Esta lei de potências é menos evidente no repositório WEBBASE, possivelmente devido ao menor número de servidores observados. Entretanto, a alta concentração das páginas em poucos servidores pode ser observada.

Na Figura 3.2 temos a proporção das páginas hospedadas em servidores com até x páginas. Podemos observar que 95% das páginas estão em servidores com mais de 1.000 páginas.

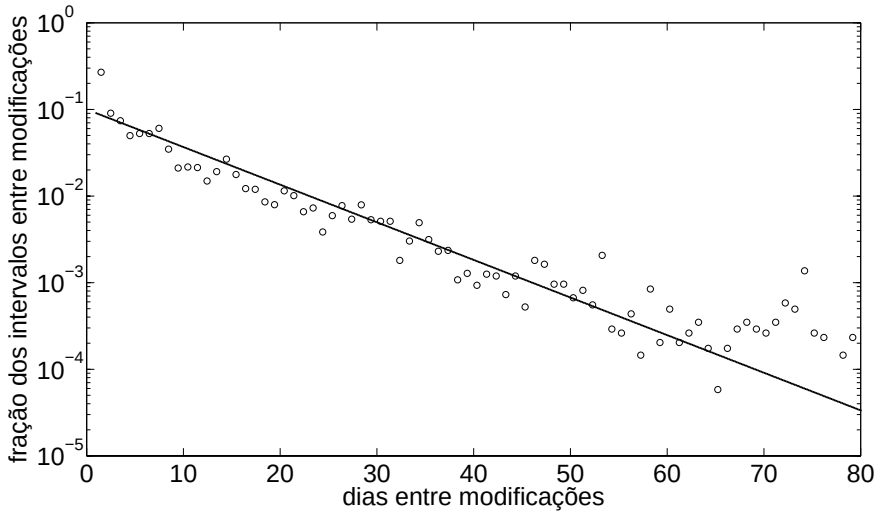

Figura 3.3: Histograma das durações dos intervalos entre modificações das páginas que se modificam em média 1 vez a cada 10 dias (repositório WEBBASE).

3.3 Modificações segundo um Processo de Poisson

A verificação da hipótese de que as páginas são modificadas segundo um processo de Poisson é feita nesta seção utilizando os mesmos passos utilizados em (Cho03a). Ou seja, coleta-se todos os intervalos entre modificações das páginas que apresentaram uma determinada taxa de modificação estimada. Em seguida, verifica-se graficamente como a duração destes intervalos se ajustam a uma distribuição exponencial.

A taxa de modificação utilizada para testar o processo de Poisson precisa estar em uma região considerada bem estimada. Como vimos, devido à discretização dos instantes de monitoramento e um período de monitoramento de 275 dias, taxas de modificação próximas de zero e próximas de uma modificação por dia não são bem estimadas. Portanto, foi escolhida a taxa de uma modificação a cada 10 dias.

O histograma das durações dos intervalos entre modificações das páginas cuja taxa de modificação foi estimada em 0,1 modificações por dia é apresentado na Figura 3.3. Como as frequências estão em escala logarítmica e os pontos se ajustam bem a uma reta, não refutamos a existência de um processo de Poisson nas modificações destas páginas. A frequência acima da esperada para os intervalos com mais de 70 dias entre modificações foi provocada por páginas com baixa taxa de modificação que, em um dado momento do período de monitoramento, passaram a se modificar diariamente. A principal explicação é a inclusão nas páginas de algum elemento com atualização permanente, como um *frame* contendo notícias.

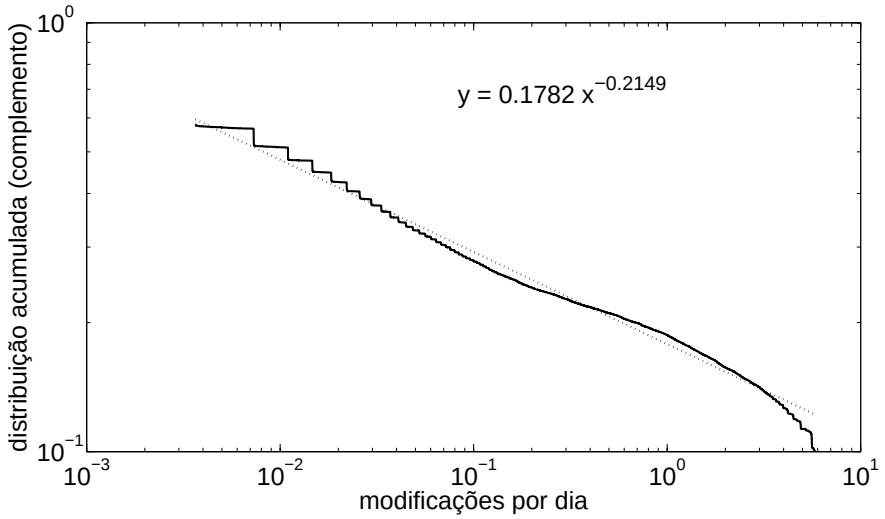

Figura 3.4: Distribuição das taxas de modificação das páginas - lei de potências (repositório WEBBASE).

3.4 Distribuição das Taxas de Modificação

As taxas de modificação das páginas monitoradas foram obtidas utilizando os estimadores de máxima verossimilhança propostos em (Cho03b). A distribuição destas taxas estimadas é apresentada na Figura 3.4. Este gráfico sugere a existência de uma lei de potências na distribuição das taxas de modificação das páginas do repositório WEBBASE.

Nas simulações realizadas ao longo da tese, cada página não monitorada recebeu uma taxa de modificação sorteada aleatoriamente dentre as taxas de modificação das páginas do mesmo servidor que foram monitoradas. Se M_s é o conjunto de páginas do servidor s que foram monitoradas, então cada página em M_s tem uma estimativa de taxa de modificação baseada nos dados coletados no monitoramento. Por outro lado, as páginas do servidor s que não estão em M_s não possuem estimativa da taxa de modificação, mas recebem a taxa de modificação de uma página de M_s escolhida aleatoriamente. Desta forma, mantemos em cada servidor a distribuição das taxas de modificação observada no conjunto de páginas monitoradas do servidor.