6
Experimentos com um Repositório de Artigos da Wikipedia

Uma dificuldade encontrada ao realizar experimentos com o repositório WEBBASE é a ausência do histórico de modificações das páginas. Por esta razão, a identificação da forma como as páginas se modificam é feita através de observações periódicas. Porém, devido à restrição de *politeness*, páginas hospedadas em servidores com muitas páginas têm uma frequência baixa de monitoramento. Uma baixa frequência de monitoramento, durante um período pré-definido de observação, pode implicar em uma identificação imprecisa do padrão de modificação da página.

Uma alternativa é realizar experimentos com um repositório de artigos do Projeto Wikipedia (Wik07), pois neste caso é disponibilizado o histórico de modificações dos artigos. Com uma única requisição ao servidor da Wikipedia podemos obter informações sobre as últimas 500 modificações de um artigo, como data, hora, e usuário modificador. Desta forma, se vamos utilizar um *crawler* para manter o nível de atualização de um repositório de artigos da Wikipedia, temos informações mais precisas sobre o padrão de modificação das páginas. Além disso, para simular qual teria sido o nível de atualização do repositório ao se utilizar uma dada política de revisitação durante um período passado, podemos utilizar os instantes reais de modificação dos artigos ao invés de observações de uma variável aleatória com distribuição estimada, como ocorre nas simulações com o repositório WEBBASE.

Entretanto, conforme observado em (Alm07) e nos experimentos realizados neste capítulo, não podemos afirmar que os artigos da Wikipedia são modificados segundo um processo de Poisson. Ou seja, as modificações dos artigos da Wikipedia não atendem à Suposição 1.4, e portanto os resultados teóricos apresentados nos Capítulos 4 e 5 podem não ser válidos para este repositório.

Por outro lado, como as distribuições das durações dos intervalos entre modificações possuem "memória", podemos avaliar o nível de atualização do repositório obtido por uma política que utiliza informações sobre a última modificação conhecida de cada artigo, chamada política GREEDY. Neste caso, a decisão sobre qual página revisitar em cada requisição ao servidor é feita no

momento da requisição por um algoritmo guloso, e não em um planejamento completo de revisitações em um horizonte futuro de operação do *crawler*, como ocorre em (Wol02). Utilizando o instante da última modificação conhecida temos um cálculo mais preciso do ganho em se revisitar cada artigo, tomando assim uma decisão mais bem informada em cada requisição ao servidor. Além disso, a etapa de alocação de recursos torna-se desnecessária, permitindo que a política se adapte mais rapidamente a modificações nas distribuições das durações dos intervalos entre modificações. Uma desvantagem desta abordagem é abrir mão de decisões globais em troca de melhores decisões locais, o que pode prejudicar a qualidade final do escalonamento.

A Seção 6.1 descreve a construção e principais características do repositório de artigos da Wikipedia utilizado neste capítulo. A política GREEDY é apresentada com mais detalhes na Seção 6.2, bem como uma variação da política MERGE (Seção 5.3) utilizada para efeito de comparação. Finalmente, a Seção 6.3 apresenta os resultados experimentais.

6.1 Repositório WIKIPEDIA

O repositório WIKIPEDIA é composto por uma amostra aleatória com 10.000 artigos sorteados dentre os artigos em inglês da Wikipedia que sofreram pelo menos 300 modificações em 2007. Esta seção apresenta a construção e algumas características deste repositório.

6.1.1 Construção

O Projeto Wikipedia (Wik07) disponibiliza *dumps* de seus artigos na página `http://static.wikipedia.org/downloads`. O *dump* de junho de 2008 era o mais recente no período de construção do repositório WIKIPEDIA. Apenas as páginas em inglês deste *dump* foram utilizadas, totalizando 5.453.838 artigos.

Cada *dump* fornece o conteúdo dos artigos em uma determinada data, mas não disponibiliza o histórico de modificações dos artigos. Utilizando o *crawler* Wget (Wge07), foram coletadas todas as modificações de artigos ocorridas em 2007. Cada artigo da Wikipedia possui uma página com o histórico de suas modificações, contendo a data, hora e minuto de cada modificação, e o usuário que realizou a modificação. Por exemplo, a página `http://en.wikipedia.org/w/index.php?title=Computer_science&action=history` contém as modificações sofridas pelo artigo com título "Computer Science".

De modo a ter uma boa estimativa das distribuições dos intervalos entre modificações dos artigos, o repositório WIKIPEDIA é composto apenas por artigos que sofreram pelo menos 300 modificações em 2007. Dentre os artigos em inglês presentes no *dump* de junho de 2008, 21.458 sofreram pelo menos 300 modificações em 2007. Deste conjunto de artigos, 10.000 foram selecionados aleatoriamente para compor o repositório WIKIPEDIA.

6.1.2 Número de Servidores

De acordo com a página `http://en.wikipedia.org/wiki/Mirrors_of_Wikipedia`, recomenda-se que os *mirrors* da Wikipedia sejam atualizados através do *download* dos *dumps* disponibilizados no site do projeto:

> *"The appropriate way to run a mirror is to download a dump of the compressed 'pages-article' file and the images from* `http://download.wikimedia.org/`*, and then use a modified instance of MediaWiki to generate the required HTML, along with above mentioned copyrights information."*

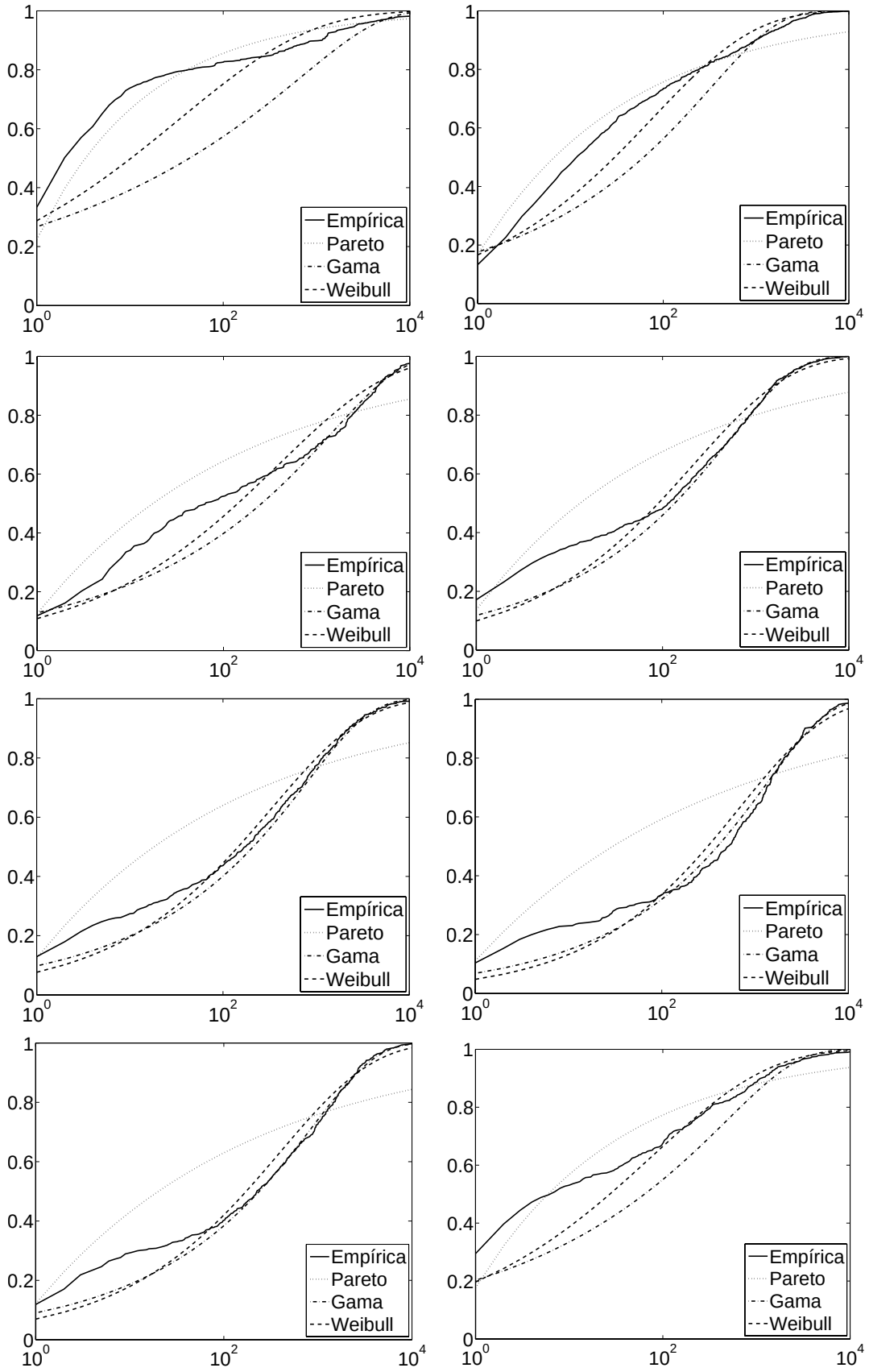
Ou seja, os *mirrors* não são notificados quando os artigos são modificados, nem são encorajados a fazer *downloads* periódicos de artigos para manter suas cópias atualizadas. Desta forma espera-se que vários artigos nos *mirrors* estejam obsoletos[1]. Além disso, os *mirrors* que são atualizados através dos *dumps* não podem fornecer informações sobre o histórico de modificação dos artigos, pois estas informações não são disponibilizadas nos *dumps*.

Durante os experimentos, o servidor `en.wikipedia.org` teve apenas um endereço IP retornado pelo servidor DNS. Assim, os experimentos realizados neste capítulo assumem que as requisições realizadas pelo *crawler* são enviadas a um único servidor.

6.1.3 Distribuição das Durações dos Intervalos entre Modificações

A Figura 6.1 apresenta a distribuição empírica do tempo entre modificações de 10 artigos do repositório WIKIPEDIA escolhidos aleatoriamente. Temos também em cada gráfico as distribuições Pareto, Gama e Weibull que melhor se ajustam aos dados, utilizando estimadores de máxima verossimilhança para os parâmetros das distribuições. As distribuições Gama e Weibull são muito utilizadas para representar a duração de uma tarefa ou o tempo até ocorrer a falha de um equipamento (Law91). As distribuições

[1]Nenhum *dump* da Wikipedia foi disponibilizado em 2009.



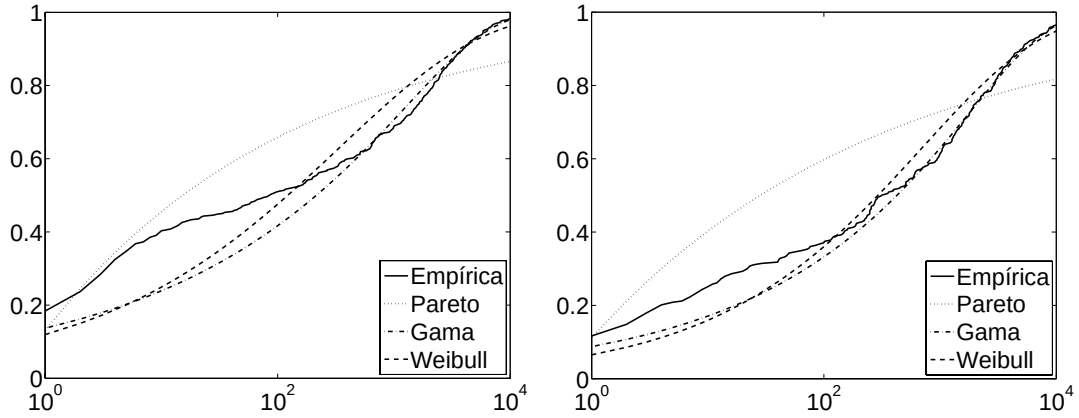


Figura 6.1: Distribuição acumulada empírica do tempo entre modificações consecutivas (em minutos) de 10 artigos escolhido aleatoriamente dentre os artigos do repositório WIKIPEDIA. Os gráficos mostram também as distribuições Pareto, Gama e Weibull ajustadas com estimadores de máxima verossimilhança.

Pareto e Weibull já foram utilizadas para modelar o tempo entre modificações de páginas *Web* (Wol02). A distribuição Exponencial é um caso particular das distribuições Gama e Weibull (Law91).

Podemos observar na Figura 6.1 que as durações dos intervalos entre modificações dos artigos na amostra não são bem ajustadas às distribuições Pareto, Gama e Weibull. Assumindo que mais de 26% dos artigos no repositório WIKIPEDIA podem ser modelados por pelo menos uma destas distribuições, a chance de sortear 10 artigos com ajuste ruim a todas elas é inferior a $(1 - 0,26)^{10} < 0,05$. Isto implica em uma chance inferior a 5% de que mais de 26% dos artigos no repositório WIKIPEDIA podem ser modelados por pelo menos uma destas distribuições.

Portanto, a distribuição empírica é utilizada neste capítulo para modelar as durações dos intervalos entre modificações. Como temos mais de 300 modificações por artigo, a representação completa da distribuição empírica é ineficiente. Por esta razão, as durações dos intervalos entre modificações de cada artigo são agrupadas em um histograma com 10 intervalos. As larguras dos intervalos do histograma crescem exponencialmente para produzir mais uniformidade no número observações nestes intervalos.

6.2 Políticas de Revisitação

Uma política de revisitação para o repositório WIKIPEDIA deve determinar o instante de revisitação de cada artigo, respeitando um intervalo de tempo mínimo entre requisições consecutivas ao servidor da Wikipedia (restrição de *politeness*). Os experimentos realizados neste capítulo assumem

que as requisições são igualmente espaçadas por servidor, portanto os instantes onde são feitas requisições ao servidor da Wikipedia são previamente definidos, restando apenas determinar qual artigo revisitar em cada um destes instantes.

6.2.1 Política MERGE

A política de seleção de páginas MERGE é definida na Seção 5.3, e consiste em utilizar a mesma ordem de revisitações realizadas pela política ótima para o problema sem a restrição de *politeness*. Entretanto, não podemos empregar a alocação de recursos apresentada no Capítulo 5, visto que as modificações de artigos no repositório WIKIPEDIA não ocorrem segundo um processo de Poisson.

A alocação de recursos proposta em Wolf et al (Wol02) permite determinar o número ótimo de revisitações de cada página em um horizonte finito de operação do *crawler* de modo a minimizar o *staleness* (Wol02) do repositório, dado que a restrição de *politeness* não precisa ser respeitada (Seção 2.1.3). Uma vantagem da alocação de recursos de Wolf et al é o fato dela permitir qualquer distribuição de probabilidades para o intervalo entre modificações de uma página, e não necessariamente um processo de Poisson como em (Cho03a). Conforme discutido na Seção 2.1.3, a métrica *staleness* pode ser considerada uma generalização do *freshness*, visto que para o caso em que o *freshness* está definido (modificações segundo um processo de Poisson) o *staleness* vale 1 menos o *freshness* quando utilizamos a política igualmente espaçada por página. Note que o *freshness* da política igualmente espaçada por página é utilizado como função objetivo na formulação do limite superior POLITE, de onde obtemos o algoritmo OPT_POLITE para alocação de recursos. Portanto, nesta seção vamos substituir o algoritmo OPT_POLITE pela alocação de recursos proposta em (Wol02).

Nos experimentos com o repositório WEBBASE realizados nos Capítulos 4 e 5, foi possível melhorar a alocação de recursos inserindo restrições que mantêm a frequência total de requisições a cada servidor em no máximo o inverso do tempo mínimo permitido entre requisições consecutivas a um mesmo servidor. No caso do repositório WIKIPEDIA assumimos que todas as requisições são enviadas a um único servidor, tornando desnecessária a utilização de restrições adicionais. Portanto, a alocação de recursos é feita exatamente como proposta em (Wol02).

Figura 6.2: Quando um artigo i é avaliado no instante t, conhecemos o instante $u_i(t)$ da última revisitação deste artigo, e o instante $b_i(t)$ da sua última modificação antes de $u_i(t)$. O instante $A_i(t)$ da primeira modificação depois de t é uma variável aleatória.

6.2.2 Política GREEDY

A métrica *staleness* assume nenhum conhecimento sobre as modificações sofridas pelas páginas, embora os instantes de todas modificações passadas estejam disponíveis no caso da Wikipedia. A política GREEDY utiliza o instante da última modificação conhecida de cada página para decidir qual artigo revisitar em cada requisição ao servidor. Ou seja, esta política abre mão de uma solução global (revisitações igualmente espaçada por página) para revisitar o melhor artigo em cada requisição ao servidor (guloso), visto que a decisão local é mais bem informada (dispõe do instante da última modificação).

A Figura 6.2 ilustra um cenário onde deseja-se selecionar um artigo para revisitação no instante t e avaliar a qualidade da escolha de cada artigo i. O instante da última revisitação ao artigo i antes do instante t é denotado por $u_i(t)$. Quando o artigo i foi revisitado no instante $u_i(t)$ conheceu-se o instante da última modificação antes de $u_i(t)$, denotado por $b_i(t)$. O instante da primeira modificação do artigo i depois do instante t não é conhecida no instante t, e portanto é uma variável aleatória denotada por $A_i(t)$.

Se o artigo i sofreu alguma modificação entre os instantes $u_i(t)$ e t, então a atualização do artigo i no instante t manterá este artigo atualizado no intervalo de tempo entre t e $A_i(t)$. Caso não tenha havido modificação do artigo i entre os instantes $u_i(t)$ e t, a revisitação deste artigo no instante t não traz nenhum melhora ao seu *freshness*. Considerando que todos os artigos tem a mesma importância no cálculo do *freshness* do repositório, o melhor artigo a ser revisitado no instante t é aquele que fornece o melhor valor esperado para o ganho em se revisitá-lo no instante t. Ou seja, a política GREEDY atualiza no instante t o artigo que maximiza

$$\mathrm{E}[\text{ganho}] = \mathrm{E}[A_i(t) - t \mid \varepsilon_i(t)] \times \Pr[\varepsilon_i(t)] + 0 \times (1 - \Pr[\varepsilon_i(t)]),$$

onde $\varepsilon_i(t)$ é o evento que indica que o artigo i está modificado no instante t. O evento $\varepsilon_i(t)$ indica que alguma modificação ocorreu entre $u_i(t)$ e t. Logo, a

probabilidade deste evento ocorrer é dada por $\Pr[u_i(t) - b_i(t) < X_i < t - b_i(t)]$ onde X_i é a variável aleatória que representa a duração de um intervalo entre modificações do artigo i. De acordo com a Suposição 1.3, as durações dos intervalos entre modificações de um artigo são observações independentes e identicamente distribuídas.

Definição 6.1 *A Política* **GREEDY** *seleciona no instante t o artigo que maximiza a expressão*

$$\mathrm{E}[A_i(t) - t \mid \varepsilon_i(t)] \times \Pr[\varepsilon_i(t)],$$

onde $\varepsilon_i(t)$ é o evento que indica que o artigo i está modificado no instante t, e $A_i(t)$ é o instante da primeira atualização depois do instante t.

Uma forma de se obter uma aproximação para o valor de $\mathrm{E}[A_i(t) - t \mid \varepsilon_i(t)]$ é através de simulações do processo de renovação que produz os intervalos entre modificações do artigo. Utilizando a distribuição dos intervalos entre modificações, inicialmente sorteia-se um valor para X_i entre $u_i(t) - b_i(t)$ e $t - b_i(t)$. Em seguida, novos valores para X_i são sorteados em todo o intervalo de valores possíveis para o tempo entre modificações consecutivas do artigo, até que a soma dos valores sorteados seja maior que $t - b_i(t)$. Desta forma obtemos uma observação para $A_i(t) - t$, dado que o evento $\varepsilon_i(t)$ ocorreu. Repetindo este processo e calculando a média dos resultados obtidos, temos uma estimava para $\mathrm{E}[A_i(t) - t \mid \varepsilon_i(t)]$. Nos experimentos realizados neste capítulo este processo é repetido 3 vezes por artigo. Isto implica em 30.000 estimativas para cada revisitação (seleção de artigo).

6.3 Experimentos

O objetivo desta seção é avaliar o *freshness* do repositório WIKIPEDIA no ano de 2007 ao se empregar as políticas MERGE ou GREEDY. Assume-se que o *crawler* começa a operar às 0h do dia 01/01/2007, e neste momento todos os artigos estão desatualizados. São utilizados na simulação os instantes reais de modificação dos artigos.

Conforme observado na Seção 6.1.3, as distribuições dos intervalos entre modificações de cada artigo são aproximadas por histogramas com 10 intervalos. As modificações de 2007 são utilizadas para produzir estes histogramas, ou seja, assume-se que os histogramas da maioria dos artigos sofrem poucas modificações com o tempo. Se esta hipótese é verdadeira, então estes histogramas fazem com que as políticas forneçam um *freshness* semelhante ao que seria obtido caso os histogramas fossem construídos com

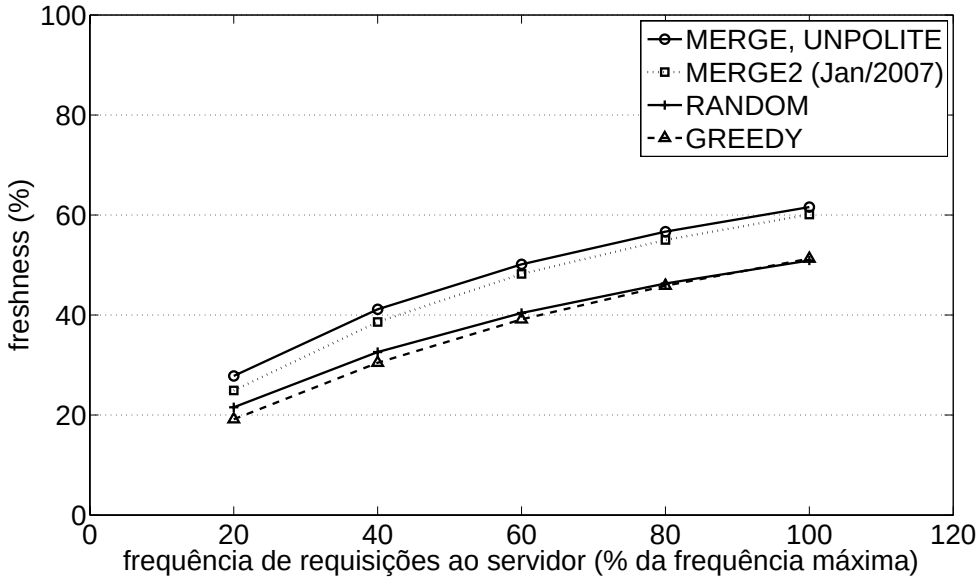


Figura 6.3: *Freshness* do repositório WIKIPEDIA após a execução das políticas MERGE, MERGE2, RANDOM e GREEDY, variando a frequência de requisições ao servidor.

dados anteriores a 2007. Para testar esta hipótese, avaliou-se também o comportamento da política MERGE ao se construir os histogramas utilizando apenas as modificações ocorridas em janeiro de 2007. Esta variação é chamada de MERGE2.

Para avaliar quanto perde-se com a restrição de *politeness*, avaliou-se também a política de Wolf et al (Wol02). A diferença para a política MERGE está no fato de que as atualizações são mantidas igualmente espaçadas por artigo (solução ótima), ao preço de produzir várias violações à restrição de *politeness*. Esta política é chamada aqui de UNPOLITE.

É importante comparar a política GREEDY com uma política que também não realiza a alocação de recursos proposta em (Wol02), mas seleciona aleatoriamente o artigo que será revisitado em cada requisição ao servidor. Ou seja, em cada requisição cada artigo tem a mesma chance de ser escolhido. Esta política chama-se RANDOM. A comparação entre as política GREEDY e RANDOM permite verificar se vale a pena tomar decisões locais mais bem informadas, visto que uma estratégia gulosa pode comprometer muito a qualidade global da solução.

A Figura 6.3 apresenta o *freshness* obtido pelas políticas MERGE, MERGE2, GREEDY e UNPOLITE; variando a frequência de requisições ao servidor da Wikipedia. Assumindo um tempo mínimo de 15 segundos entre requisições consecutivas, a frequência máxima é de 4 requisições por minuto. Cada política de revisitação na Figura 6.3 utilizou 20%, 40%, 60%, 80% e 100%

desta frequência máxima.

A diferença de *freshness* entre as políticas MERGE e UNPOLITE foi inferior a 0,4%, indicando que os ajustes para respeitar o *politeness* (revisitações igualmente espaçadas por servidor) praticamente não afetaram o *freshness* do repositório.

A diferença de *freshness* entre as políticas MERGE e MERGE2 diminui com o aumento da frequência de revisitação, indo de 2,9% à 1,5%. Esta diferença também é pequena, indicando que as modificações sofridas nos histogramas dos artigos entre fevereiro e dezembro de 2007 produziram pequeno impacto no *freshness* do repositório.

A diferença de *freshness* entre as políticas MERGE e GREEDY é de $10,30 \pm 0,95\%$, sem uma relação clara com a frequência de revisitação. Ou seja, mesmo utilizando decisões locais mais bem informadas (conhecendo o instante da última modificação de cada artigo), a política GREEDY produz *freshness* inferior à política MERGE (que toma decisão global assumindo nenhum conhecimento sobre as últimas modificações). Por outro lado, como a política GREEDY não tem uma etapa de alocação de recursos, ela pode ir atualizando o histograma de cada artigo revisitado, tendo uma resposta mais rápida às variações de comportamento dos usuários que modificam os artigos. Mas como percebido na diferença entre as políticas MERGE e MERGE2, não espera-se que esta característica traga grande melhoria de *freshness* no caso do repositório WIKIPEDIA.

Podemos observar também que o *freshness* fornecido pela política GREEDY ficou um pouco abaixo do fornecido pela política RANDOM, com uma diferença mais evidente para frequências menores de revisitação. Esta diferença chega a 2,4% quando o *crawler* realiza requisições utilizando 20% da frequência máxima. Isto mostra que a utilização de mais informação não compensa o que se perde com o uso de uma estratégia gulosa. Uma maior uniformidade nas durações dos intervalos entre revisitações de cada artigo tendem a promover um melhor resultado global, e isto é realizado melhor pela política RANDOM que pela política GREEDY.

Utilizando um PC Intel Core 2 Quad 2,4GHz, a heurística GREEDY consome em média 144 milissegundos para selecionar uma página. O MERGE consome em média 7 segundos na alocação de recursos, e 166 nanossegundos para selecionar uma página.