Política de substituição dinâmica em Caches para Web

Roberto Ferreira Brandão
brandao@dcc.unicamp.br
Instituto de Computação
UNICAMP

Ricardo de Oliveira Anido
ranido@dcc.unicamp.br
Instituto de Computação
UNICAMP

Abstract

The use of caches in document transmission systems in the Web makes possible to decrease the load of requisitions submited to the transmission means and to servers of documents, reducing the user´s wait time when searching for documents.

In web cache systems, a copy of a requested document is stored in the cache and, if new requisitions to that document ocour, those requisitions can be answered using the copy stored in the cache.

One of the most important parameters related to web caches is the substitution policy. This is an algorithm that decides which document will be removed fromcache to store a new one when there is no more free space. This paper presents a new substitution policy that uses, dinamically, both size and time since last requisition, increasing the efficiency of the cache system.

Resumo

A implantação de caches em sistemas de transmissão de documentos via WWW possibilita uma diminuição da carga de requisições apresentadas aos meios de transmissão e aos servidores de documentos, diminuindo o tempo de espera do usuário dos serviços na busca por documentos. Em um sistema de caches para Web, uma cópia de um documento requisitado é armazenada em um cache de forma que, se houverem novas requisições a esse documento, essas poderão ser satisfeitas utilizando a cópia armazenada no cache.

Um dos parâmetros mais importantes relacionado com caches para Web é o de política de substituição, que é o algoritmo que decide qual documento deve ser removido do cache para que um novo possa ser armazenado, quando o espaço livre for insuficiente para armazenar o novo documento. Esse trabalho apresenta uma política de substituição dinâmica, que utiliza critérios tais como o tamanho e o tempo desde a requisição dos documentos para decidir quais documentos remover do cache para que seja armazenado um novo, aumentando dessa forma o desempenho do sistema de cache.

1 Introdução

No contexto de *World Wide Web*, cache significa um local de armazenamento temporário de documentos disponíveis na Web. A implantação de caches em sistemas de transmissão de documentos possibilita uma menor carga nos meios de transmissão e nos servidores remotos, apresentando ao usuário final dos recursos da Web um menor tempo de resposta às requisições feitas [Malpani95], [Abrams95].

Vários parâmetros influem no desempenho de sistemas de caches para Web [Melve98]. Um dos parâmetros mais importantes é o de política de substituição, que é o algoritmo que decide qual documento deve ser removido do cache quando o espaço de armazenamento disponível for insuficiente para armazenar um novo documento.

Desde o início da implantação de caches para Web, novas tecnologias têm sido pesquisadas de modo a otimizar a utilização dos caches. Dentre essas tecnologias pode-se citar o desenvolvimento de diversas políticas de substituição e a possibilidade de particionamento do espaço de armazenamento do cache [Murta99]. Esse particionamento permite uma melhor adaptação do cache aos perfis de documentos que são transmitidos através do cache.

4 Particionamento do cache

O particionamento do cache permite uma melhor adequação do espaço de armazenagem às diferentes classes de tamanhos dos arquivos. O particionamento permite evitar que muitos arquivos pequenos sejam removidos para o armazenamento de um único arquivo grande [Murta98]. Com uma configuração correta das partições é possível aumentar o *Hit Ratio* de um cache mantendo-se praticamente inalterado o valor do *Byte Hit Ratio.*

O particionamento possui a desvantagem de necessitar de um estudo aprofundado das classes de tamanhos dos arquivos de forma a definir uma correta divisão do espaço de armazenamento.

5 Política Dinâmica de Substituição

Com o objetivo de diminuir os efeitos da fragmentação em caches particionados, é proposta aqui uma política de substituição dinâmica, capaz de alternar dinamicamente a aplicação das políticas SIZE e LRU de modo a evitar que muitos arquivos sejam apagados para a entrada de apenas um arquivo, não apresentando o problema de fragmentação do espaço.

O modelo utilizado na política de substituição dinâmica aqui apresentada está representado esquematicamente na figura 1.

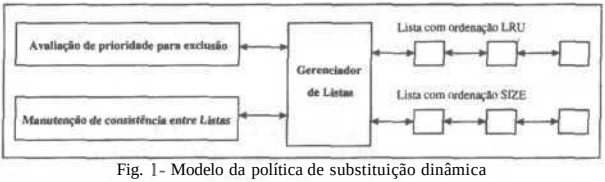

Fig. 1- Modelo da política de substituição dinâmica

Através da figura 1, observa-se que existem duas ordenações para os mesmos arquivos no cache. Uma fila é ordenada para aplicação da política SIZE enquanto que a outra é ordenada através da política LRU. Para decidir quais arquivos devem ser removidos para a entrada de um novo, é avaliado qual das políticas deve ser aplicada. Dessa forma, é possível utilizar um algoritmo de escolha baseado tanto no tempo desde o último acesso quanto no tamanho do arquivo, de forma a tomar uma decisão mais correta de qual arquivo remover.

A figura *2* apresenta os resultados em *Hit Ratio* da aplicação da política de substituição dinâmica, comparando-os com os resultados tanto do método do particionamento quanto com aplicações de políticas sem o particionamento do espaço de armazenagem do cache. A figura 3 apresenta a comparação dos resultados medidos em *Byte Hit Ratio.*

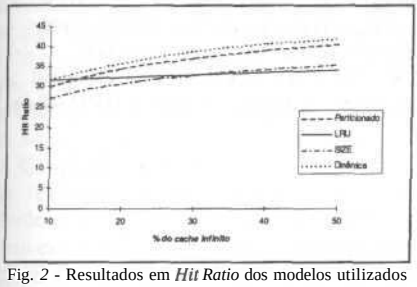

Fig. *2* - Resultados em *Hit Ratio* dos modelos utilizados

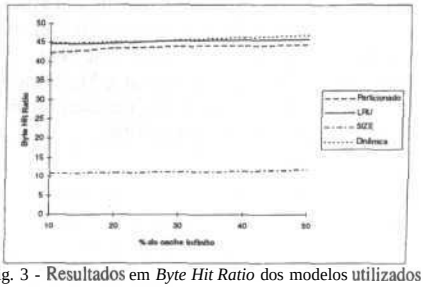

Fig. 3 - Resultados em *Byte Hit Ratio* dos modelos utilizados

Esse trabalho apresenta uma política de substituição dinâmica, que utiliza os critérios de tamanho e tempo desde a requisição dos documentos para decidir quais remover do cache para que seja armazenado um novo, proporcionando um melhor desempenho que o modelo do particionamento, propondo ainda uma solução para o principal problema do particionamento, que é o problema da fragmentação do espaço de armazenamento.

2 Simulações de caches para Web

Devido à dificuldade em se obter acesso a um sistema real de caches para fazer experimentos, a simulação de caches se torna a melhor opção no estudo do impacto da variação de parâmetros relativos a caches para Web. Para que sejam realizadas essas simulações foi implementado um simulador de caches para Web. A primeira versão desse simulador foi apresentada em [Brandão99]. O simulador utiliza como dados de entrada um arquivo contendo um histórico das requisições feitas a um cache real, de forma que a simulação seja baseada em requisições reais dos usuários. Dessa forma, é possível analisar o impacto das modificações de parâmetros relativos aos caches sem necessitar de um cache real para experiências.

Para configurar um sistema de caches a ser simulado, utiliza-se um arquivo de configuração que contém descritas, numa linguagem pré-definida, as características do sistema de caches a ser simulado. Para simular as requisições feitas aos usuários foi utilizado um arquivo que contém os históricos de requisições a caches reais. Esses arquivos são chamados de arquivos de *log (logfiles)* ou *traces. A* utilização de arquivos de log permite simular situações reais de cargas de requisições feitas a caches reais. Particularmente, os arquivos de *log* utilizados nas experiências descritas nesse artigo foram conseguidos em [Squid2000], sendo relativos à carga no período de 7 a 15 de janeiro de 2000 de um dos caches do sistema IRCACHE.

3 Políticas de substituição

Políticas de substituição são algoritmos executados quando o tamanho do documento que deverá ser gravado no cache excede o espaço livre do cache. São removidos então os documentos considerados mais dispensáveis até que o espaço seja suficiente para armazenar o novo documento [Kim98], [Willians96].

Existem três grandes questões quanto à remoção de documentos de um cache: o que remover, quando remover e quantos documentos devem ser removidos [Lorenzetti98]. A medida de eficiência de sistemas de cache para web é medida em *Hits* e em *Byte Hits.* O termo *Hit* indica que uma determinada requisição foi satisfeita usando um documento do cache. O termo *Miss* indica o contrário: um documento requisitado não foi encontrado no cache. O termo *Hit Ratio* indica o percentual de documentos requisitados que foram encontrados no cache enquanto que o termo *Byte Hit Ratio* indica o percentual de bytes requisitados pelos clientes que foram encontrados no cache.

As políticas de substituição que serão abordadas nesse artigo são a LRU e SIZE. Essas políticas são descritas resumidamente a seguir.

- LRU *(Least Recently Used)* - *É* removido o documento menos recentemente acessado. Dessa forma, mantém-se no cache os documentos que foram acessados mais recentemente, sendo esses os documentos com maior probabilidade de serem reacessados.
- SIZE - Quando essa política é aplicada a um conjunto de arquivos, é removido o maior documento do conjunto. Dessa forma, tenta-se evitar que muitos arquivos pequenos tenham que ser removidos para a inserção de um único arquivo grande.

A política de substituição dinâmica, entretanto, tem a desvantagem de exigir um maior poder computacional devido à necessidade de manutenção de duas filas de arquivos, avaliação da melhor opção para exclusão e manutenção de coerência entre as filas.

6 Conclusões e trabalhos futuros

A utilização da política dinâmica para gerenciamento do espaço de armazenagem de um sistema de cache permite uma maior taxa de acerto por documento (*Hit Ratio*) e uma maior taxa de acerto por byte *(Byte Hit Ratio)* que o modelo particionado ou a utilização das políticas de substituição em um cache não particionado.

A principal desvantagem da utilização da política dinâmica é a necessidade de manutenção de duas filas de arquivos, ordenadas de forma diferente, o que implica numa maior demanda por processamento e um maior consumo de memória. Além disso, é necessário o processamento do algoritmo de avaliação e escolha dos arquivos a serem removidos e do algoritmo de manutenção de consistência entre as duas filas, o que aumenta ainda mais a necessidade de alto poder de processamento do computador onde o cache está instalado.

Como perspectivas para o futuro pretende-se estudar mais profundamente a política de substituição dinâmica de forma a otimizar seu desempenho, procurando também diminuir a necessidade de processamento apresentada por ela. Além disso, estudos utilizando a política dinâmica em sistemas de caches cooperativos e distribuídos devem ser realizados.

7 Referências bibliográficas

[Abrams95] Abrams, M.; Standridge, C. R.; Abdulla, G.; Willians, S.; Fox, E. A. Caching proxies: Limitations and potentíals. 4th International WWW Conference, pp 119-133, dez/95.

[Brandão99] Brandão, R. F.; Anido, R. O. Cooperação entre caches para Web, Anais do 17º Simpósio Brasileiro de Redes de Computadores, pp 533-548, 1999.

[Kim98] Kim, H.; Chon, K.; Update policies for network caches. http://cosmos.kaist.ac.kr/salab/publication/technical-memo/SAL-TM-75, jul/98.

[Lorenzetti98] Lorenzetti, P.; Rizzo, L. Replacement policies for a proxy cache. URL: http://www.iet.unipi.it/~luigi/caching.ps.gz, jul/98.

[Malpani95] Malpani, R.; Lorch, J.; Berger, D. Making World Wide Web caching servers cooperate. 4th International WWW Conference, pp 107-117, dez/95.

[Melve98] Melve, I. 1997. Web caching architecture. URL: http://www.uninett.no/prosjekt/desire/arneberg/altsammen.html, jul/98.

[Murta98] Murta, C. D. 1998. Cache Particionado - Uma nova abordagem para cache na WWW. URL: http://wwwcache.ja.net/events/workshop/24/, ago/98.

[Murta99] Murta, C. D., Modelos de Particionamento de Espaço para Caches da World Wide Web. Tese de Doutorado. Departamento de Ciência da Computação, UFMG, 1999.

[Squid2000] Examples of Squid Log Files. URL: ftp://ftp.ircache.net/Traces/, jan/2000.

[Willians96] Willians, S.; Abrams, M.; Standridge, C. R.; Abdulla G.; Fox E. A.; 1996. Removal policies in network caches for World Wide Web Documents. *ACM SIGCOMM96*, pp. 293-305, agosto 1996.