Uma Arquitetura Proxy para a Tradução de Páginas WWW

Hermes A. L. Camelo[1] & Rafael Dueire Lins[2]

[1]Centro de Informática - UFPE - Recife - PE - Brasil
halc@cin.ufpe.br

[2]Departamento de Eletrônica e Sistemas
Centro de Tecnologia e Geociências - UFPE - Recife - PE - Brasil
rdl@cin.ufpe.br

Resumo. Este trabalho apresenta um servidor proxy para a tradução inglês-português de páginas da Internet. Este tipo de servidor é um filtro ativo e transformador de documentos da WWW, com grande aplicabilidade.

Abstract. This paper presents a proxy server for English to Portuguese translation of Web documents. This server acts as an active document filter, with great applicability.

1 Introdução

Com o crescimento acentuado da Internet, surgem problemas relativos à velocidade de acesso, segurança, conteúdo das informações, entre outros. Diversas propostas de solução são estudadas e a que tem conseguido um maior número de adeptos é o uso de um servidor proxy [2]: um servidor que assume papel intermediário entre dois pontos de comunicação, tais como um cliente e um servidor. Seu propósito primário é melhorar a qualidade da rede. Exemplos deste tipo de servidor incluem *firewalls* [4], roteadores programáveis [7], *gateways* para redução de tráfego [1] e filtros de conteúdo. Além disso, esse tipo de arquitetura promove compatibilidade com as mais diversas plataformas utilizadas pelo cliente, administração centralizada e a conveniência do ambiente de programação no nível de aplicação.

Dentre os serviços que podem ser oferecidos por um servidor proxy, a filtragem e/ou processamento de documentos da WWW vem despertando grandes possibilidades deste tipo de arquitetura. Além de guardar documentos, um proxy também é capaz de selecioná-lo e reprocessá-lo convenientemente [6][8], podendo ser utilizado para retirar partes do documento ou até mesmo bloquear certos tipos de páginas com material impróprio para menores.

Este trabalho apresenta uma arquitetura para a disponibilização de um servidor proxy para a tradução inglês-português [3] de páginas da WWW.

2 Arquitetura do Sistema

O sistema pode ser decomposto em três componentes: Servidor Proxy, Tradutor HTML e Tradutor. A interação entre eles pode ser vista na Figura 1 e funciona da seguinte forma: o usuário utiliza o Servidor Proxy como provedor de informações durante a navegação pela Internet. A cada pedido de novo documento, o proxy irá buscá-lo na rede, guardá-lo e passá-lo para o usuário acrescido de um ponto de ativação ao final do mesmo. A solicitação de tradução é feita através de um clique sobre o *link* presente no ponto de ativação. O Servidor Proxy irá

2.3 A Ferramenta de Tradução

A principal função deste componente é a tradução de textos escritos em Inglês para o Português Brasileiro. Para tanto, a ferramenta [3] dispõe de um dicionário com cerca de 40.000 raízes de uso coloquial e de um dicionário técnico na área de informática. As estruturas gramaticais de ambas as línguas também estão presentes, permitindo um processamento mais refinado do documento a ser traduzido. Como a tradução é realizada frase a frase, não é necessário aguardar pela tradução completa da página para começar a visualizá-la.

3 Análise Comparativa e Considerações sobre Desempenho

Como a ferramenta de tradução independe da forma utilizada para se obter o documento da Internet, o primeiro componente da arquitetura, o Servidor Proxy, pode ser substituído por sistemas que apresentem a mesma funcionalidade. Programas que funcionem como CGI [5], por exemplo, são fortes candidatos quando se deseja implementar este tipo de serviço *on-line.*

No caso do uso de um proxy, o tempo de aquisição das páginas não teria muita influência no processo já que o documento a ser traduzido provavelmente já estaria presente em uma *cache* local ao servidor. Este tempo de carga que é economizado pode ser notado em outras ferramentas similares como, por exemplo, o sistema BabelFish [11]. Neste último caso, o usuário informa a URL do documento que deseja traduzir e o sistema ainda terá de dispender tempo para obtê-lo da rede antes de gerar a tradução.

Outro ponto em que o uso de um script CGI para o tipo de sistema aqui proposto é deficiente está no uso de *cookies* [5]: pequenas informações enviadas por um servidor web para serem armazenadas em um navegador Internet. Como a solicitação do documento é feita diretamente ao CGI, o navegador não irá enviar os *cookies* armazenados para o servidor da página em questão. Com isso, alguns sites que exigem o uso de *cookies* durante a navegação, como, por exemplo, os serviços tipo *webmail*, poderão não ter suas páginas traduzidas através de um CGI. O mesmo problema não ocorre ao se utilizar um Servidor Proxy.

É importante salientar que a disponibilização de um servidor proxy para ser utilizado por qualquer usuário da Internet (e não somente para aqueles que estejam dentro da mesma sub-rede ou atrás de um *firewall*)fatalmente irá criar um gargalo no servidor, pois todo o acesso à rede se dará através dele, tornando o tempo de navegação bem mais lento e fugindo do objetivo principal deste tipo de arquitetura. Além disso, os usuários que já utilizam algum outro proxy poderão estar impossibilitados de desfrutar deste tipo de serviço. Nesses casos, um CGI se torna mais atraente e eficaz para o internauta.

4 Aplicações Correlatas

Esta mesma arquitetura também é interessante para aplicações que apresentem características de alterar o conteúdo do objeto original. A tradução ou leitura de páginas WWW é apenas uma delas. Outros exemplos de aplicações podem incluir:

- Remoção de conteúdo não desejado, como propagandas veiculadas através de *banners* ou textos que apresentem assuntos considerados ofensivos. Sistemas desse tipo já foram propostos e existem vários programas que removem propaganda de sites da *web* [12].
- Inclusão de propaganda. O proxy pode servir para a inclusão de *banners* por Provedores de Acesso. Neste novo conceito, a propaganda está no provedor e não na página acessada.

interagir com o Tradutor HTML que, por sua vez, irá filtrar a codificação HTML do documento original e enviar este novo documento para o Tradutor. Este último recebe o documento sem qualquer marca de formatação, processa-o, e retorna sua tradução. O Tradutor HTML irá então recompor a codificação do documento original no documento traduzido. Finalmente, o usuário recebe do Servidor Proxy o documento traduzido, em seu próprio navegador, com todas as características gráficas do documento original mantidas.

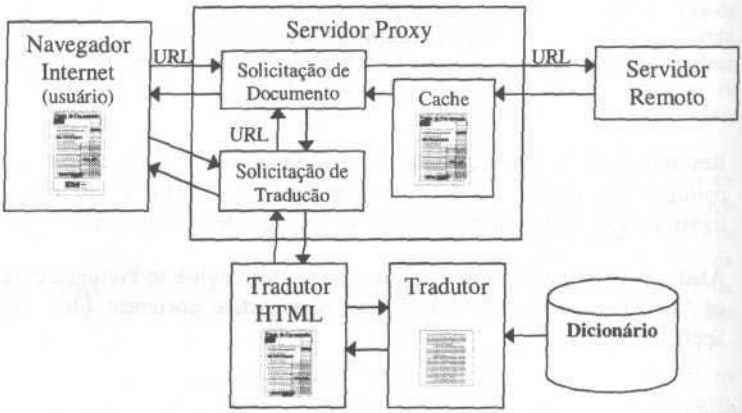

Figura 1: Ambiente de Tradução via Servidor Proxy

2.1 O Servidor Proxy

Dentre os servidores existentes atualmente, optou-se pelo Apache[1] pela simplicidade de uso, compatibilidade com o sistema operacional e facilidade de implementação (código aberto).

Enquanto o usuário navega pela rede, o Servidor Proxy automaticamente adiciona um ponto de ativação *(link)* ao final de cada página. Apesar de não fazer parte do documento original, o ponto de ativação (texto em HTML) é interpretado pelo navegador Internet como sendo parte integrante do mesmo e tratado como tal. Uma solicitação de tradução é reconhecida através de um padrão de URL previamente definido:

```
http://_tradutor_.adv?<URL_a_ser_traduzida>
```

O campo `<URL_a_ser_traduzida>` é preenchido com o endereço da página atual.

2.2 O Tradutor HTML

Este componente serve de interface entre o proxy e o Tradutor. Ele é capaz de filtrar a página original, gerando um texto sem formatação que será passado para a ferramenta de tradução. Ao receber o texto já traduzido, a formatação HTML do documento original será recomposta.

Vale observar que a própria formatação em HTML pode ser utilizada para indicar partes do documento que não precisam ser traduzidas. Por exemplo, os comentários não precisam ser traduzidos já que não serão visualizados pelo usuário. Também não são traduzidos elementos como endereços (presente entre as *tags* <address> e </address>) e *scripts* (presente entre as *tags* <script> e </script>).

---

[1] O sistema de tradução aqui apresentado inclui software desenvolvido pelo Apache Group para uso no projeto do servidor HTTP Apache (http://www.apache.org).

- Intermediação de salas de bate-papo (*chats*). O servidor faria a tradução de cada mensagem recebida para uma linguagem própria (UNL [9], por exemplo) e a repassaria para os demais integrantes da sala no idioma escolhido por cada um.

5 Conclusões

Este trabalho apresentou perspectivas de uso de servidores proxy na filtragem de documentos que circulam pela Internet. Uma ferramenta foi incorporada a um proxy para fornecer serviços de tradução inglês-português de páginas da WWW.

O uso de arquiteturas centralizadas aponta vantagens com relação à facilidade e à velocidade que se pode avaliar seu desempenho. Uma delas é a não necessidade de instalação física do produto (software) na máquina do usuário final. Como a tradução é realizada no servidor, nem o sistema operacional do cliente (usuário) nem o navegador Internet utilizado pelo mesmo influi no processamento, garantindo a compatibilidade com as mais diversas plataformas de software e hardware existentes. Além disso, é possível realizar atualizações e correções constantes e de forma transparente para os usuários, não sendo preciso disponibilizar nenhum arquivo para *download*.

O Servidor Proxy aqui descrito já está em funcionamento e disponível na URL `http://www.di.ufpe.br:9191`. O serviço de tradução é oferecido gratuitamente e os interessados em utilizá-lo encontrarão as instruções de como configurar o navegador Internet na própria página do projeto Tradutor [10].

Referências

[1] B.Badrinath, et al. Handling mobile hosts: A case for indirect interaction. In *Proc. 4th Workshopon Workstation Operating Systems*, pp. 91-97, IEEE, October 1993.

[2] C.Brooks, M.S.Mazer, S.Meeks, & J.Miller. *Application-specificproxy servers as HTTP stream transducers*. World Wide Web Journal, December 1995.

[3] H.A.L.Camelo, & R.D.Lins. Uma ferramenta de auxílio à tradução inglês-português. In *Actas do IV PROPOR*, pp. 43-52, Évora, Portugal, 1999.

[4] D.Chapman, & E.Zwicky. *Building internetfirewalls*. O'Reilly and Associates, 1995.

[5] J.D.Hamilton. *CGI programming101*. Paperback, 2000.

[6] A.Luotonen, & K.Altis. World-Wide Web Proxies. *Computer Networks and ISDN Systems* 27(2): 147-154, 1994.

[7] D.Tennenhouse, & D.Wetherall. Towards an active network architecture. *ACM Computer Communications Review* 26(2): 5-18, April 1996.

[8] B.Zenel, & D.Duchamp. A general purpose proxy filtering mechanism applied to the mobile environment. In *Proc. of Mobicom'97*,Hungria, 1997.

[9] *UNL - Universal Network Language*. UNU/IAS. <URL: `http://www.unl.ias.unu.edu`>

[10] Projeto Tradutor. <URL: `http://www.di.ufpe.br/~tradutor`>

[11] AltaVista Translations. <URL: `http://babelfish.altavista.com`>

[12] *Filtering and blocking banner ads and other unwanted material.* <URL: `http://www.junkbusters.com/ht/en/links.html#filtering`>