MAC499 - Trabalho de formatura supervisionado

SquidPCB - Squid-cache Pornography Content Blocker

Fernando Lemes da Silva
Orientado por Prof. Dr. Roberto Hirata Jr.

O software

O SquidPCB é um software auxiliar destinado a filtragem de conteúdo pornográfico acessado através da web. Ele atua em conjunto com o software Squid-cache, um popular servidor proxy disponível como software livre.

Figura 1: Squid-cache Proxy Server Logo (http://www.squid-cache.org)

Motivação

A motivação deste trabalho está em impedir que as pessoas acessem conteúdo impróprio, seja na forma de texto ou imagem, por motivos que vão desde a proteção de crianças, as quais atualmente estão muito presentes e expostas na internet, até mesmo o controle sobre o que pode ou não ser acessado em locais públicos ou de trabalho.

Legalidade e censura

Apesar da internet ser considerada uma mídia extremamente democrática, sob determinadas circunstâncias é razoável entender que este tipo de material pode estar sujeito ao controle de um administrador de rede, o qual é responsável tanto pela manutenção deste canal com a rede mundial de computadores, quanto legalmente por eventuais crimes que sejam cometidos a partir de uma máquina que faz uso deste canal. O papel de censor, nos casos citados acima seria atribuído aos pais, que não devem permitir que seus filhos sejam influenciados, ou aos administradores de rede, sejam de *“lan houses”*, salas pró-aluno, ou grandes empresas.

Solução utilizada

Em relação a tecnologia utilizada para oferecer este controle de conteúdo, estamos falando basicamente de um servidor proxy, um software que recebe requisições de navegadores web para recuperar determinados objetos (páginas HTML, imagens, scripts, etc.) acessíveis geralmente através do protocolo HTTP, que antes de enviar o arquivo ao usuário que o solicitou o submete a uma análise para avaliação da probabilidade deste possuir conteúdo pornográfico.

Utilizando as informações de caminho e probabilidade de pornografia de um arquivo, uma estrutura de dados em forma de árvore é mantida de forma que é possível bloquear não só os arquivos que atingiram um determinado limiar, mas também caminhos que de acordo com a árvore tem uma grande probabilidade de conterem conteúdo similar.

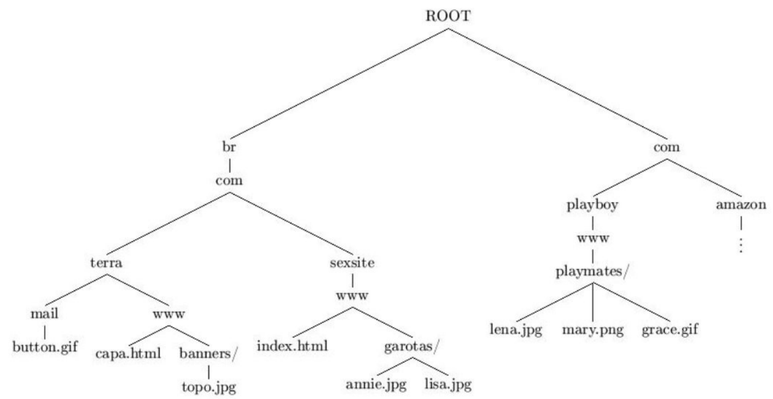

Figura 2: Árvore com nós que armazenam as URLs acessadas

As tecnologias implementadas para classificação de conteúdo são bastante simples, onde para um texto um autômato determinístico é utilizado para encontrar ocorrências de palavras ou frases que indicarão evidências sobre seu conteúdo, enquanto que para imagens é analisado a predominância de "cor de pele", a qual é facilmente destacada em uma visualização HSB (Matiz, saturação e brilho).

Através do gráfico ao lado podemos ver claramente que as cores referentes ao tom de pele são bem próximas, tanto para pele branca como para pele negra, o que nos garante uma boa precisão na seleção da pele na imagem.

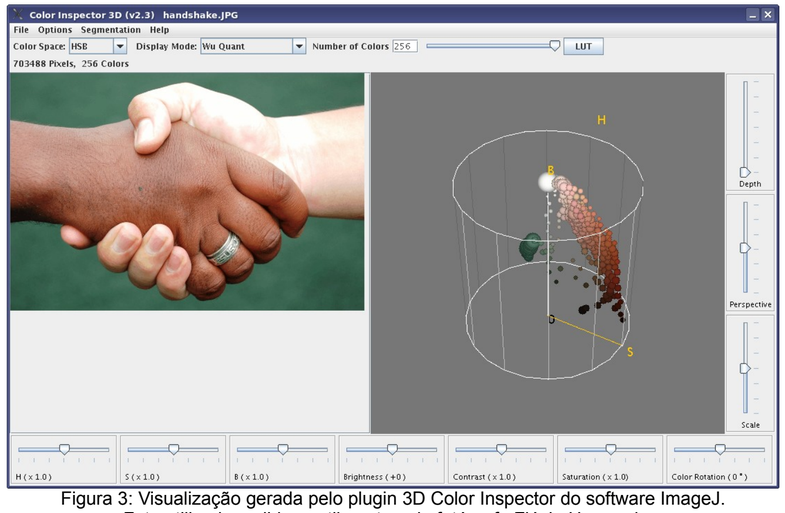

Figura 3: Visualização gerada pelo plugin 3D Color Inspector do software ImageJ.
Foto utilizada cedida gentilmente pelo fotógrafo Flávio Hernandes.

Apesar da segmentação da imagem pelas cores ser facilmente tratável, seria necessário utilizar algoritmos mais precisos, pois em testes até mesmo uma foto de um ursinho de pelúcia foi suficiente para gerar um resultado falso-positivo.

Da forma que o sistema foi implementado, os filtros podem ser facilmente estendidos, porém filtros mais elaborados poderiam comprometer o desempenho do sistema caso o servidor proxy tenha uma quantidade muito grande de acessos.

Conclusão

O trabalho proposto é uma tentativa de oferecer aos responsáveis uma forma de evitar o acesso a conteúdo impróprio, porém devida a quantidade de formas que existem para driblar estes controles, acredito que a maneira mais eficaz de obter os resultados desejados ainda é através da educação dos usuários sobre o assunto.