

SquidPCB

Squid-cache Pornography Content Blocker

Fernando Lemes da Silva
Nr.USP 2371843

O que é o SquidPCB?



O SquidPCB é um filtro de conteúdo que trabalha em conjunto com o servidor proxy Squid-cache, realizando a análise e o armazenamento de informações sobre cada arquivo pertinente acessado através deste servidor proxy.

Essencialmente o programa é uma **base de dados** sobre arquivos presentes na internet com dados sobre a probabilidade de cada um destes arquivos possuir **conteúdo pornográfico**.

Motivação do projeto



A motivação deste projeto é a de permitir que administradores de rede consigam bloquear, ou ao menos dificultar, o acesso a conteúdo pornográfico na internet.

Os principais alvos deste controle são:

i) Crianças;
ii) Usuários em locais públicos;
iii) Funcionários em geral.

Ética e censura



Apesar da restrição de acesso poder ser considerada uma forma de censura, devemos levar em conta que a entidade que cede o acesso a internet é indiretamente **responsável** pelos seus usuários.

A censura em alguns casos é **necessária**, pois um estabelecimento eventualmente poderia ser responsabilizado por permitir que menores tenham acesso a sites pornográficos. Mesmo que os pais tenham o cuidado de utilizar de programas para proteção de seus filhos, quando fora de suas casas estes estarão a sujeitos a sites pornográficos e salas de bate papo altamente sujeitas a presença de **pedófilos**.

Solução proposta



A solução proposta por este trabalho envolve basicamente três partes:

i) Análise de arquivos quanto a conteúdo impróprio;
ii) Armazenamento de informações obtidas;
iii) Análise dos dados para pré-classificação de arquivos ainda não acessados.

Análise de arquivos



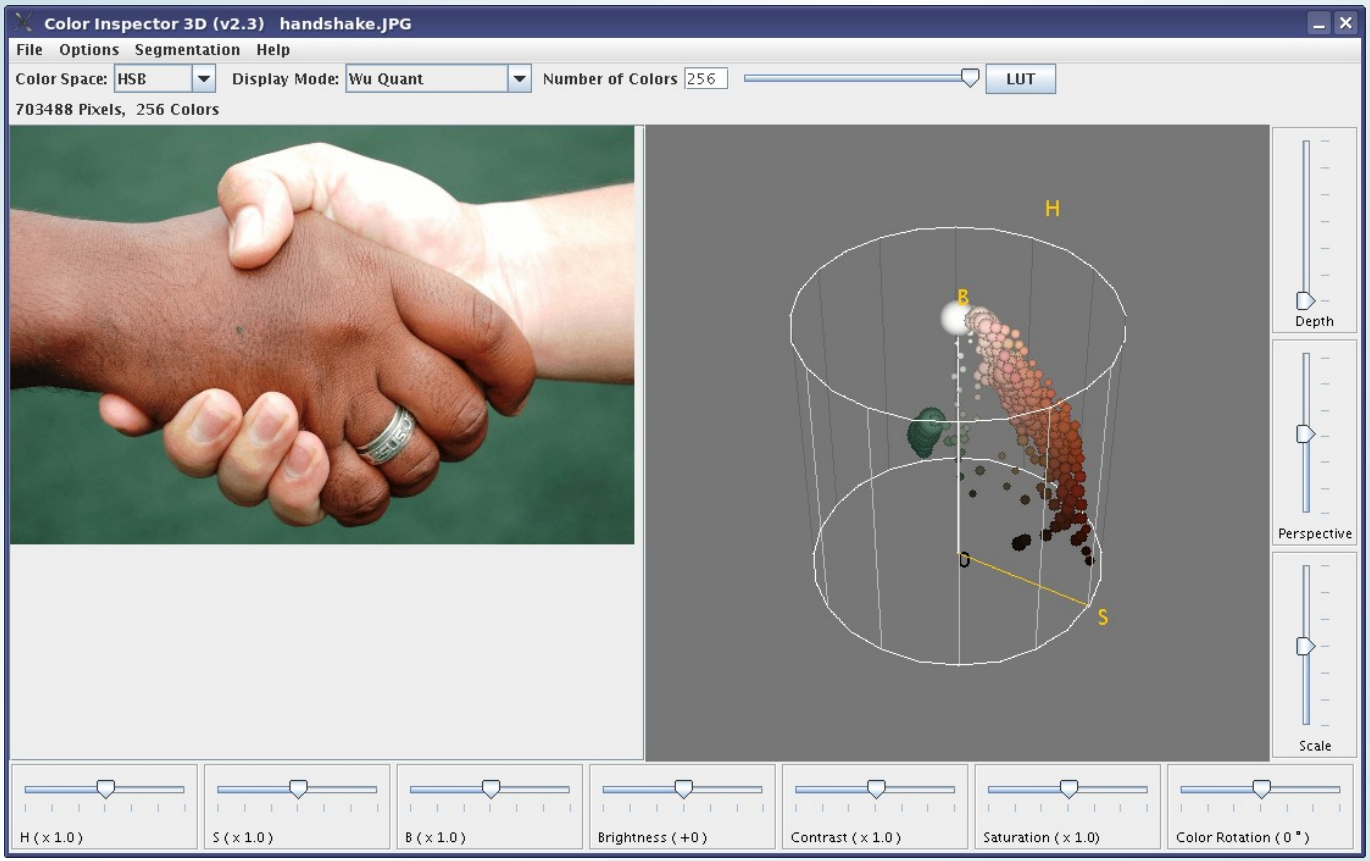
O SquidPCB analisa tanto arquivos de texto como imagens. Arquivo texto são analisados em **busca de palavras ou frases** que indiquem alguma probabilidade deste arquivo ser pornográfico.

Para cada palavra ou frase definida é atribuído um valor inteiro que será somado a um contador, o qual de acordo com um limite configurável indicará a probabilidade do arquivo ser pornográfico.

O algoritmo utilizado para esta função envolve um **autômato** determinístico que o percorre os arquivos em **tempo linear**, mesmo tendo milhares de itens em sua lista.

Análise de arquivos

Em relação as imagens, o SquidPCB determina a proporção dos pontos (pixels) “cor de pele” através da segmentação da visualização HSB destes.



Análise de arquivos

Em determinadas imagens o filtro produz um resultado bastante razoável. A foto abaixo possui 3,42% de seus pixels classificados como cor de pele.



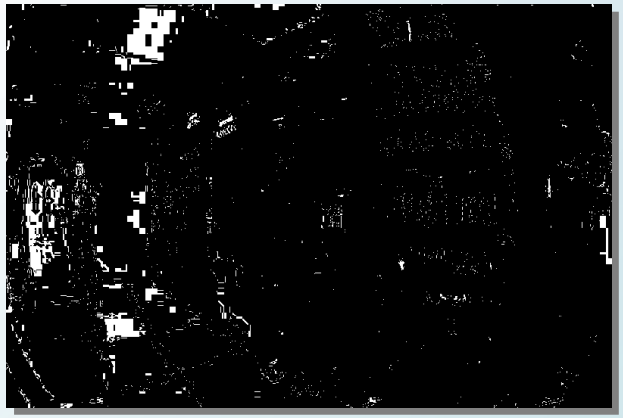


Foto por Gregory Maxwell sob licença GNU Free Documentation License.

Análise de arquivos

Abaixo temos uma imagem que de acordo com o filtro de imagens possui 45,19% de seus pixels contendo cor de pele.





Foto por Marcus Obal sob licença GNU Free Documentation License.

Análise de arquivos



O fato do filtro implementado neste projeto levar em conta somente a cor de pele certamente levará a vários resultados falsos-positivos.

Dado que a confiabilidade do analisador de textos é muito superior ao de imagens, por poder extrair com mais precisão características do texto, a estrutura de dados envolvida no projeto se torna importante por combinar os resultados de ambos os métodos.

Outros métodos de análise de imagens mais precisos também poderiam ser incorporados ao filtro de imagens, porém a quantidade de processamento requerida poderia ser muito grande. Nestes casos o filtro implementado neste projeto seria útil, realizando uma análise preliminar das imagens.

Armazenamento de informações



Um site fictício **www.sexsite.com.br** poderia conter os seguintes objetos que seriam acessados através da página index.html:

/index.html

/images/banner.jpg

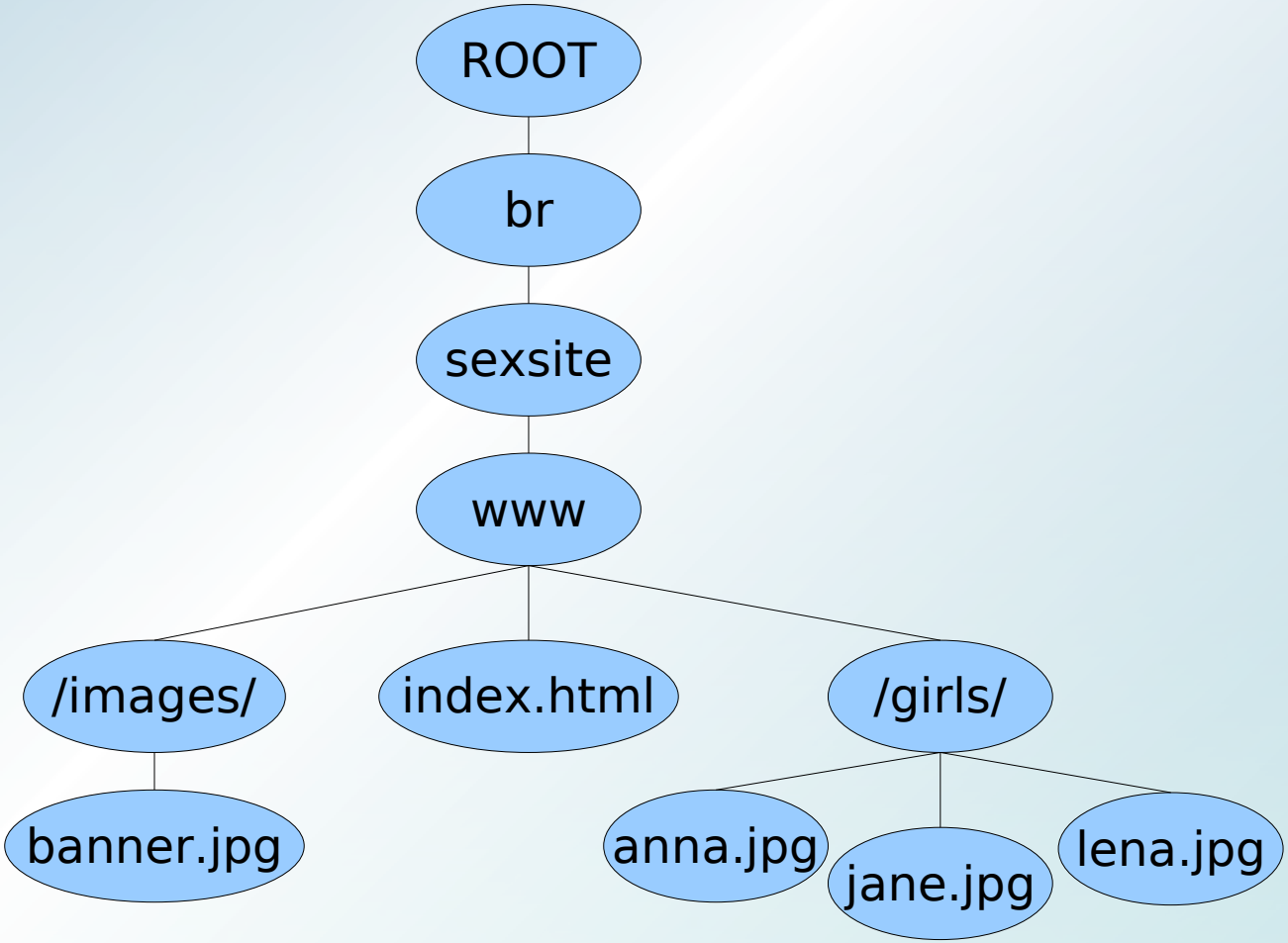
/girls/anna.jpg

/girls/jane.jpg

/girls/lena.jpg

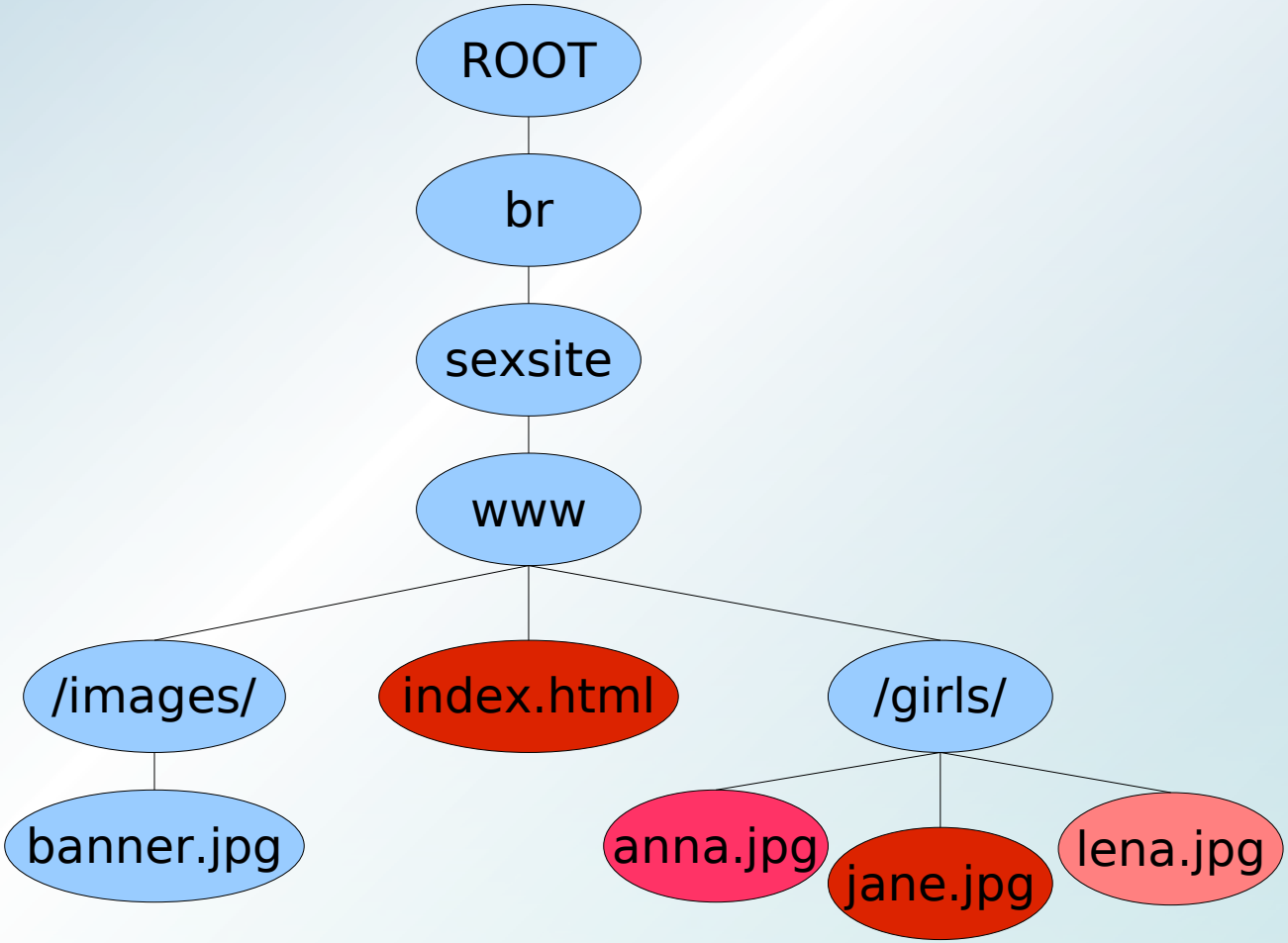
Armazenamento de informações

A árvore de dados produzida seria:



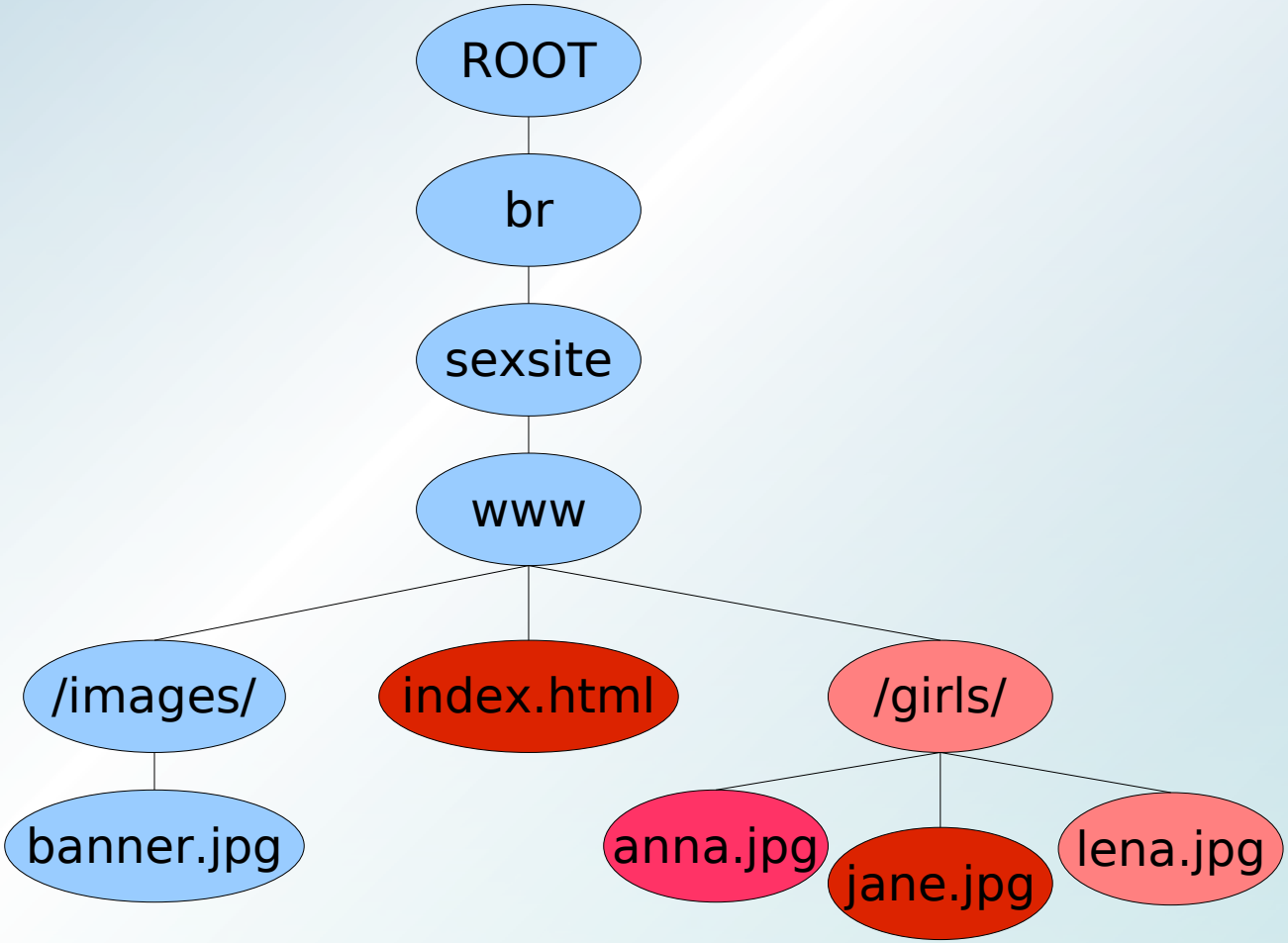
Análise dos dados

A árvore de dados produzida seria:



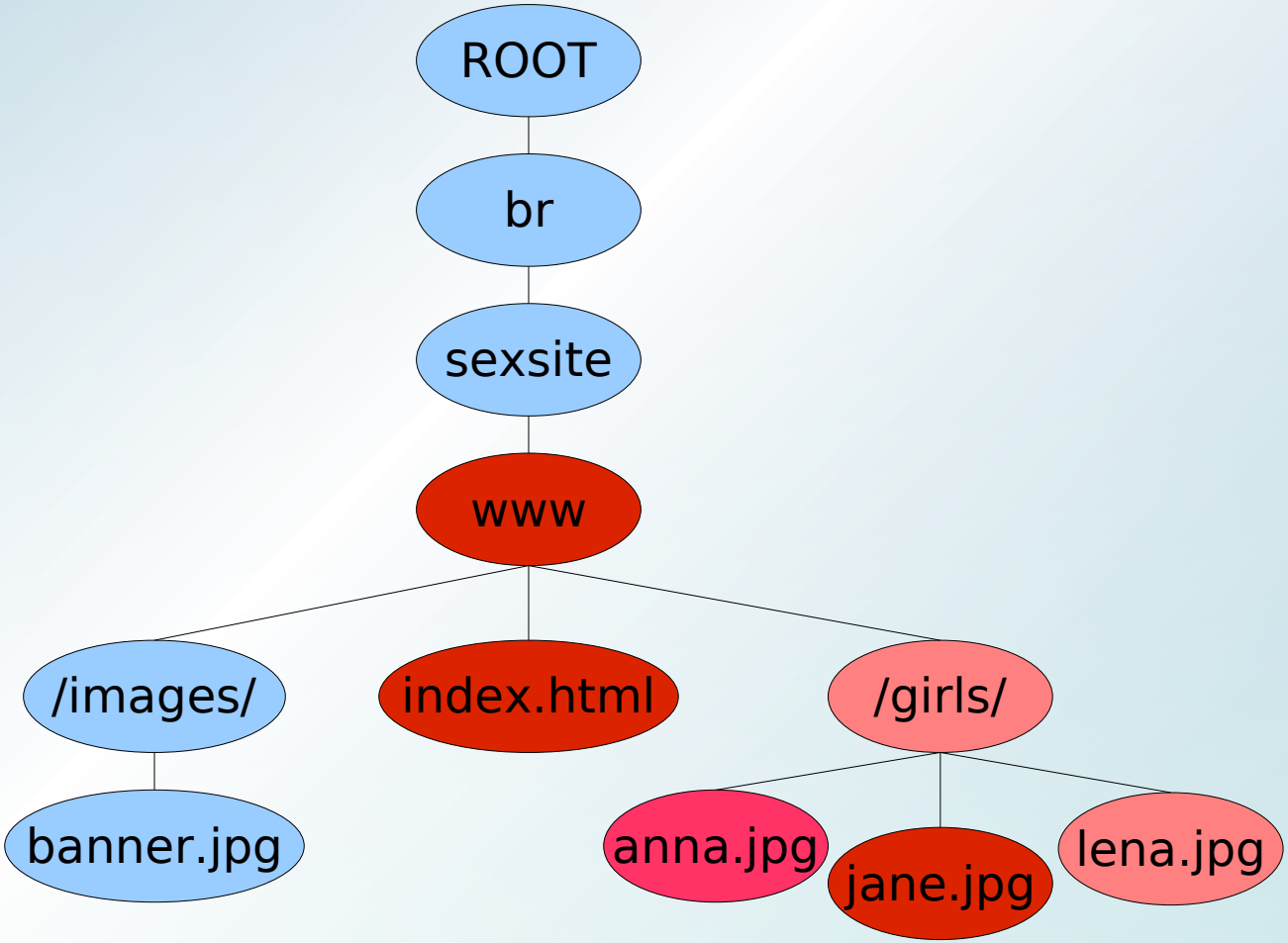
Análise dos dados

A árvore de dados produzida seria:



Análise dos dados

A árvore de dados produzida seria:



Conclusões



- Apesar dos esforços existentes em implementar controles de acesso, estes sempre estarão sujeitos a serem burlados;
- A educação do usuário, no sentido de estabelecer normas de conduta, locais e idade adequadas para acesso a este conteúdo, ainda é a melhor arma que temos para evitar a internet se torne além de uma ferramenta para disseminar informação, uma ferramenta para banalizar a pornografia.