Proposta de um rastreador para recuperação de materiais ilícitos na Web

Roberto Wander Bezerra da Costa[1], **Sandro Raphael de Oliveira Paiva**[1], **Flavia Cristina Bernardini**[1], **Carlos Bazilio Martins**[1]

[1]Departamento de Ciência e Tecnologia — RCT
Pólo Universitário de Rio das Ostras — PURO
Universidade Federal Fluminense — UFF
Rio das Ostras, RJ, Brasil

{rwander,srpaiva}@gmail.com, fcbernardini@vm.uff.br, bazilio@ic.uff.br

Abstract. *We propose in this work a focused crawler for retrieval of illegal material in the Web. The content of the retrieved pages are analyzed in order to retrieve only illicit material. Our work were motivated by the fact that popular search engines do not retrieve this kind of material, as we could observe in our case study. Yet, as the Web has considerably grown, an efficient search is crucial. In this work, we describe the architecture of the proposed focused crawler, as well as we describe the main features of the focused search.*

Resumo. *Propomos neste trabalho um buscador focado para recuperação de materiais ilícitos na Web. O conteúdo das páginas será analisado com o intuito de se buscar somente material ilegal. Nosso trabalho foi motivado pelo fato dos buscadores disponíveis não conseguirem, em geral, recuperar esse tipo de material, conforme pudemos observar em um estudo de caso realizado. Ainda, como a Web tem crescido exponencialmente, uma busca eficiente é crucial. Neste trabalho descrevemos a arquitetura do buscador focado proposto, assim como descrevemos as características principais da busca focada.*

1. Introdução

Com o crescimento exponencial da informação disponível na Web, uma ferramenta de busca tradicional, mesmo baseada em sofisticados algoritmos de indexação de documentos, tem dificuldades de ser eficaz na recuperação de informações relevantes a seus usuários. As pessoas têm cada vez menos paciência e tempo para formularem perguntas em suas buscas por informações na Web. Consequentemente, é vital em muitas aplicações que o buscador encontre a informação desejada rapidamente. Assim, realizar buscas baseadas nas necessidades do usuário é um tema que tem sido investigado na área de busca na Web [Micarelli et al. 2007a, Webb et al. 2001].

Um tema ainda pouco explorado é a necessidade de busca por informações que poucos buscam, porém que há uma relevância para serem pesquisadas. É o caso de buscas de páginas relacionadas à venda de drogas ilícitas. Um dos problemas para procurar páginas na Web está relacionado ao tamanho da rede. A quantidade de páginas disponíveis é imensa, o que faz com que buscadores focados necessitem podar a busca, tipicamente realizada em largura, utilizando heurísticas para a poda [Pant et al. 2004, Menczer et al. 2004]. Para buscar páginas relevantes a um tema, além de todas as fases de

processamento das páginas em HTML que precisam ser realizadas para armazenar uma estrutura que represente a página [Menczer et al. 2004, Micarelli et al. 2007b], deve-se armazenar um conhecimento a respeito do tema que se quer pesquisar. Esse conhecimento pode ser desde uma série de termos que represente o tema até ontologias que representem o domínio. Há diversos critérios que podem ser levados em conta no quão próximo uma página está do tema que se busca. Além de se verificar a presença de palavras-chave no texto, pode-se atribuir uma pontuação extra, caso ao menos uma das palavras-chave estiverem no título, por exemplo [Menczer et al. 2004].

Para buscar páginas específicas a um tema, relacionadas a situações ilícitas, como, por exemplo, venda de uma droga controlada por receituário médico, não é desejável que sejam encontradas páginas que expliquem os efeitos do medicamento, mas sim páginas que vendam tais produtos ilicitamente. Os sistemas de busca descritos na literatura, focados ou não, utilizam métodos que disponibilizam páginas comumente acessadas pelos usuários [Micarelli and Gasparetti 2007]. Porém, as páginas que vendem produtos ilícitos geralmente não são referenciadas por sites tipicamente acessados ou não estão entre as páginas mais acessadas pelos usuários da internet, mesmo que esses usuários se sintam falsamente protegidos por acreditarem que suas identidades reais são desconhecidas [Demetriou and Silke 2003]. Assim, neste trabalho propomos um rastreador[1] cujo objetivo é encontrar páginas relacionadas a venda de drogas ilegais na Web. Para motivar o trabalho a ser realizado, descrevemos um estudo de caso, onde investigamos resultados de buscadores não focados quando tentamos realizar uma busca com palavras-chave relacionadas ao tema específico.

Este trabalho está organizado como segue: Na Seção 2, descrevemos as características gerais que um buscador na Web deve possuir, e na Seção 3 são descritas as características dos buscadores focados, também denominados rastreadores. Ambas as seções foram baseadas nos trabalhos [Micarelli and Gasparetti 2007, Menczer et al. 2004]. Na Seção 4, descrevemos a arquitetura do rastreador proposto. Na Seção 5, descrevemos o estudo de caso realizado. Finalmente, na Seção 6, fazemos as considerações finais deste trabalho, incluindo os trabalhos futuros a serem realizados.

2. Buscadores na Web

A Internet é o maior repositório e disseminador de informações da sociedade moderna. Sua abrangência está além de fronteiras territoriais, diferenças culturais ou divisões de classes. Entretanto, uma das linhas de estudos atuais está relacionada a buscar e encontrar de forma eficiente e satisfatória informações específicas, que atendam às necessidades de pesquisa de um determinado usuário. Isso se deve ao fato de a Internet não estar estruturada para classificar as informações que a compõe. Os buscadores na Web são softwares cujo objetivo é facilitar a busca por informações na Internet. Eles varrem a web, analisando e armazenando seu conteúdo, de forma enxuta e estruturada, para permitir que futuras consultas possam ser realizadas em curtos períodos de tempo. Entretanto essa tarefa não é simples. A seguir são listadas as principais dificuldades [Micarelli and Gasparetti 2007]:

Conteúdo Distribuído. Não existe um, ou pelo menos poucos, servidores que centralizem o conteúdo da Web. Ao contrário, as informações estão largamente dis-

[1]Traduzimos como rastreador os termos *crawler* ou *spider*.

tribuídas. Ainda, não existe um servidor central que possibilite informar com precisão o tamanho e o crescimento da Web. Um estudo realizado no início de 2005 revela que parte da Web considerada potencialmente indexada pelos maiores mecanismos de busca tinha em torno de 11,5 bilhões de páginas. Estudos anteriores sugerem que o tamanho da Web dobre a cada dois anos. Uma consideração amplamente aceita é que a velocidade da rede cresce menos que a capacidade de armazenamento, isso tanto para um usuário comum quanto para grandes comporações. Atualmente uma estimativa do tamanho da Web[2] informa que seu tamanho está em torno de 21,77 bilhões de páginas indexadas pelos maiores buscadores existentes.

Armazenamento do Conteúdo para Análise e Consultas. A quantidade de informações disponível é muito grande para ser armazenada de forma compacta e sem perda de conteúdo.

Alta Taxa de Crescimento e Atualizações. Uma página analisada hoje pode não existir mais amanhã ou ter sido completamente modificada. Alguns estudos tentam estimar qual a frequência com que as páginas da Web mudam durante um determinado período. Por exemplo, após o monitoramento de 700 mil páginas diariamente durante quatro meses, foi possível estimar que 40% dessas páginas sofreram mudanças em uma semana e 23% delas , localizadas no domínio `.com`, mudaram diariamente [Cho and Garcia-Molina 2003]. Para manter os resultados atualizados, os mecanismos de busca fazem cópias locais das páginas remotas com certa frequência. É possível desenvolver políticas de atualização que garantam baixo custo computacional e de recursos da rede através de estimativas da frequência de mudança das páginas. Uma destas técnicas de estimativa consiste em monitorar, regularmente, por um determinado período, a frequência de atualização da página através de agentes. Outra forma consiste em notificar o mecanismo de busca, por exemplo o Google SiteMaps, sobre a alteração ou inclusão de uma página. O webmaster deve informar através de um arquivo XML, a lista de páginas do Web site e qual a frequência de alterações. Essa proposta diminui o trabalho do buscador de ter que analisar e obter uma estimativa da frequência de atualização, otimizando o uso de recursos computacionais e da rede.

Organização ou Classificação do Conteúdo de Forma Estruturada. Sem isso, fica difícil informar, com precisão, que o resultado de uma pesquisa realmente está de acordo com o contexto semântico esperado pelo usuário. Existem propostas da W3C[3] para a implantação da Web 3.0 ou Web Semântica.

Devido às dificuldades relatadas acima, torna-se complexo o desenvolvimento de ferramentas de busca na Web. O projeto do buscador apresenta algumas particularidades, tais como recursos de rede eficientes, robustez e gerenciabilidade. Assim, rastreadores distribuídos constituem uma importante proposta para permitir que simples unidades computacionais formem uma infraestrutura interligada e colaborativa, aumentando a velocidade de recuperação de páginas da Web e a tolerância a falhas. Além disso, deve atender a políticas de recuperação de páginas de servidores Web, que especificam um intervalo mínimo entre duas consultas para o mesmo servidor, com o intuito de recuperar páginas e, ao mesmo tempo, não ocasionar danos na performance e disponibilidade do servi-

[2]Vide site `http://www.worldwidewebsize.com`.

[3]http://www.w3.org/2001/sw/

dor varrido. Outra política consiste na adoção de um tipo de protocolo de comunicação com o servidor a ser varrido, definido por um arquivo em formato texto, denominado `robots.txt`, no qual contém pastas e/ou arquivos que não devem ser recuperados pelo buscador [Micarelli and Gasparetti 2007].

Além dos problemas mencionados, varrer a Web não é algo trivial e pode ser até mesmo inviabilizado caso não sejam feitas boas escolhas para iniciar a varredura. A adoção do referido protocolo de comunicação limita a quantidade de páginas e conteúdos que podem ser indexados de um determinado servidor. Ainda, [Broder et al. 2000] discute a questão da conectividade. No trabalho mencionado, os autores analisaram cerca de 200 milhões de páginas e 1.5 bilhão de links. As páginas da Web analisadas foram divididas em quatro grupos distintos. O primeiro grupo é denominado Componente Fortemente Conectado (*Strongly Connected Component* — SCC), o qual é formado por 27% das páginas analisadas. Essas páginas têm a propriedade de poderem alcançar outras através de links diretos. O segundo e o terceiro grupo são chamados respectivamente de grupos de Entrada (IN) e Saída (OUT), correspondendo aproximadamente por 21% das páginas analisadas em cada grupo. As páginas do grupo IN podem alcançar as do SCC, porém não são alcançadas a partir dele. As páginas do grupo OUT são acessíveis a partir do SCC porém não têm link de retorno. O restante das páginas, algo em torno de 31%, não pode ser alcançado ou alcançar o SCC. Assim, algumas partes da Web não são facilmente encontradas. Se a busca iniciar de alguma página classificada como OUT, sua varredura estará limitada a 1/5 da Web. Na Figura 1 é mostrada a relação entre os grupos IN, SCC e OUT.

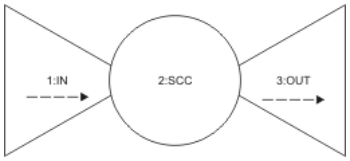

Figure 1. Relação entre os grupos IN, SCC e OUT de páginas da Web.

A abrangência da Web é usada para especificar todos os tipos de páginas que a compõe. Ocorre que algumas vezes seu uso é estendido para arquivos não textuais como arquivos de audio, filmes, etc. Para cada um deles haverá ferramentas de busca que terá problemas para determinar sua representação. Outras duas questões importantes relacionadas a buscadores na Web são a recuperação de páginas dinâmicas e de diferentes tipos de documentos tais como audio, video, etc [Bergman 2001]. Tais buscas não serão tratadas neste trabalho.

Uma das características importantes na implementação de um buscador consiste na existência de uma fila de links para serem recuperados, alimentada pela sucessiva extração de links das páginas processadas. A estratégia de download adotada pelo buscador afeta diretamente a ordem dessa fila. Essa estratégia consiste na adoção de métricas para ordenar os links de acordo com o grau de importância para o buscador em questão. Quatro conhecidas métricas são: Busca em Largura (Breadthfirst), Backlink Count, PageRank e HITS. Essas métricas estão descritas em [Menczer et al. 2004, Micarelli and Gasparetti 2007]. Na busca em largura tradicional, a Web é representada como um grafo, onde as páginas são os vértices e os hyperlinks são as arestas. Os links

vizinhos a um vértice são todos analisados e somente após o término da visita desses links são visitados os links subsequentes. Um link nunca é visitado duas vezes. A métrica Backlink Count consiste em atribuir um grau de importância para cada página a ser recuperada. Esse grau de importância é calculado por meio da função do número de páginas recuperadas que apontam para uma página a ser recuperada. Essa métrica influencia as métricas HITS e PageRank, algoritmos iterativos empregados no âmbito da Web para atribuir grau de relevância para cada recurso, contabilizando os links dentro de cada página. Para cada consulta realizada no buscador, poderá ser retornada uma infinidade de links para páginas a serem analisadas pelo usuário. HITS e PageRank tentam lidar com esse problema explorando a presença de links entre os documentos para descobrir novas informações a serem usadas no processo de ranqueamento.

Tendo em mente que as ferramentas de busca não são aptas a recuperar todo o vasto e crescente conteúdo da Web, a pesquisa deve ser focada nas melhores estratégias para priorizar a recuperação de páginas, já que os buscadores podem recuperar somente uma fração da Web em um determinado período. Portanto, é importante recuperar as páginas mais relevantes, que irão atender às principais necessidades de seus usuários. Na seção seguinte, discutimos questões relacionadas a um rastreador focado.

3. Rastreador Focado

Os buscadores tradicionalmente convertem a página Web em uma sequência de texto, extraindo os links contidos na página. Cada um desses links é utilizado para recuperar novas páginas. No entanto, buscadores focados exploram informações adicionais contidas nas páginas, como as âncoras de texto descritas nos links. Assim, uma das características que especificam um buscador focado é a forma como ele explora os links de uma página. Essa informação extraída é utilizada para indicar qual o benefício de se recuperar a página indicada no link. Isso ocorre porque não se conhece nada a respeito dessa página, sendo então essa análise feita com o intuito de evitar o uso desnecessário de processamento e rede, ao recuperar e analisar documentos irrelevantes. De forma geral, todos os buscadores focados são baseados no fenômeno do tópico de localidade. De acordo com esse fenômeno, páginas da Web são comumente agrupadas por tópicos, contendo vários links conectando uma página à outra.

Um tópico de localidade e âncoras ocorrem toda vez que uma página, referenciada por outra, pertence ao mesmo contexto. Isso ocorre pelo fato dos autores referenciarem em suas páginas links para outras, com o objetivo de permitir que seus usuários possam ir além na busca por informações ou conteúdo correlatos. Buscadores focados exploram o fenômeno da localidade recuperando grupos de páginas toda vez que encontram alguma página com conteúdo relevante. Um grupo é constituído por páginas que podem ser alcançadas pela página corrente através da extração de seus links. A busca é interrompida toda vez que uma página que não contenha conteúdo correlato é alcançada. Esse fenômeno é muito comum na Web. Uma página está mais facilmente relacionada com o conteúdo em questão através de páginas que apontam para ela do que, por exemplo, uma análise feita de forma aleatória para chegar a outras páginas. Além disso, alguns resultados mostram que páginas irmãs, ou seja, aquelas que são alcançadas por uma mesma origem, têm maior grau de similaridade em seu conteúdo de acordo com a proximidade do posicionamento dos links na página de origem [Davison 2000], como pode ser visualizado na Figura 2.

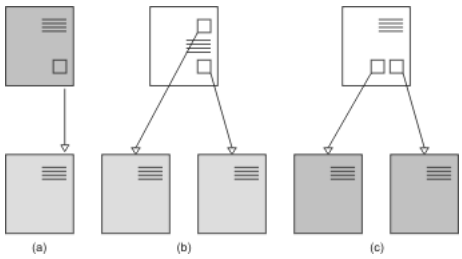

Figure 2. Relação entre páginas irmãs, alcançadas por uma mesma origem. (a) Uma única página alcançada. (b) Duas páginas alcançadas pela mesma página, porém com links distantes na mesma página. (c) Duas páginas alcançadas pela mesma página, porém com links próximos na mesma página. Neste caso, o grau de similaridade é maior.

A recuperação da informação na Web é realizada com base no conteúdo extraído das páginas, com o intuito de determinar a compatibilidade da informação com as necessidades de busca dos usuários. A hiperinformação, ou seja, a informação contida para estruturar a página Web, é usualmente ignorada. As informações extraídas da estrutura de links são utilizadas, pelos buscadores focados, com o intuito de navegar entre as páginas que tenham proximidade com o conteúdo de interesse. O fenômeno do tópico de localidade, as âncoras e os algoritmos HITS e PageRank são, sem dúvida, importantes tópicos a serem examinados nas buscas focadas.

Há 4 (quatro) importantes abordagens de busca focada propostas na literatura [Micarelli and Gasparetti 2007], as quais são:

Baseada em Taxonomias: Inclui buscadores compostos por um sistema que faz uso de dois métodos de itens que direcionam a busca. O primeiro é um classificador que avalia a relevância dos documentos de hipertexto, observando os tópicos selecionados, e o segundo é um identificador, que identifica nós de hipertexto que são considerados importantes pontos de acesso para outras páginas relevantes, através de alguns links. Após ter coletado as páginas de interesse, o usuário seleciona os melhores nós da árvore de categorias de um classificador treinado sobre uma dada taxonomia. A hierarquia ajuda a filtrar as páginas corretas durante a busca.

Tunelamento: Buscadores focados tendem a não seguir links cujos textos não têm ligação com o contexto da pesquisa em realização. Apesar de útil e efetivo, esse método pode impedir que sejam encontradas páginas relevantes. Isso se deve ao fato de que páginas de um mesmo tópico não necessariamente apontam umas para as outras. Uma solução é seguir um número limitado de páginas que não serviriam no contexto para encontrar páginas relevantes.

Buscador Contextual: Dado um conjunto de documentos considerados relevantes previamente, a pesquisa realizada por alguns mecanismos de busca são aptos a construir uma representação da distância de páginas recuperadas em relação ao conjunto de documentos relevantes. Assim, a Web é percorrida inversamente, no sentido de que as páginas analisadas são as que apontam para o conjunto de documentos relevantes. Toda a informação recuperada é armazenada em uma estrutura denominada grafo de contexto que mantém, para cada página, uma distância relativa, definida como o número mínimo de links necessários para alcançar cada página do conjunto inicial.

Web Semântica: Basicamente, o objetivo da busca utilizando a abordagem de Web Semântica é definir uma medida de relevância para mapear o conteúdo de uma página Web, utilizando uma ontologia do domínio fornecida pelo usuário. Casamentos textuais unidos às relações taxonômicas e léxicas entre super e sub-conceitos da ontologia, utilizadas para calcular uma relevância mais precisa do documento recuperado, são fatores fundamentais nessa abordagem.

Nossa proposta envolve a utilização das idéias centrais citadas nas quatro abordagens descritas acima, com exceção da última, Web Semântica. Em seu lugar pretendemos adotar uma abordagem baseada em Distâncias Semânticas, implementada através da representação do documento por espaço vetorial.

4. Arquitetura de um Rastreador para Busca de Materiais Ilícitos

Na Figura 3 é mostrada uma representação do funcionamento do sistema de busca especializado. Os componentes toCrawl e crawled são bases de dados que armazenam as URLs a serem percorridas e as já percorridas, respectivamente. O componente Base de Conhecimento representa o conhecimento sobre o contexto buscado, que é especializada à medida que os agentes colhem informações sobre as URLs.

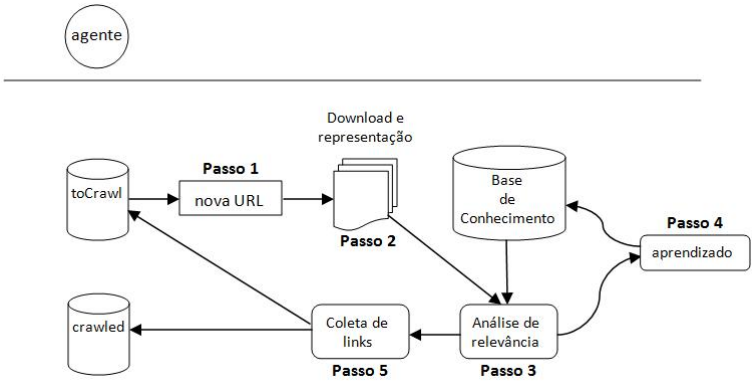

Figure 3. Representação do funcionamento do agente buscador.

O sistema de busca proposto neste trabalho é um sistema multi-agente, ou seja, um sistema composto por diversos agentes de software que realizam a tarefa de busca de páginas na Web. Vários agentes trabalham em paralelo e se comunicam com as bases de dados mencionadas, recebendo URLs da base toCrawl para serem percorridas, e alimentando a base crawled com novas URLs. No início do processo, o usuário indica para um agente de início uma URL para iniciar o processo de busca, conhecidas como sementes da busca. Esse agente de início recebe uma URL, indicada pelo usuário, e inicia a varredura pela URL, alimentando ambas as bases de dados toCrawl e crawled, até então sem informações. Ele difere dos outros agentes pois, em um primeiro momento, não colhe informações do banco de dados. Após essa fase de início, esse agente realiza o mesmo trabalho dos outros agentes, ou seja, uma URL é recuperada da base toCrawl para ser varrida, essa URL é armazenada em crawled, e os links resultantes do processamento dessa página são armazenadas em toCrawl. Na Figura 4, é ilustrado esse processo. No Passo 1, o agente buscador recupera uma URL do banco de dados toCrawl. No Passo

2, o agente buscador faz a requisição para o servidor na internet onde esta URL está armazenada. Esse servidor retorna o documento requisitado (seja um arquivo em formato HTML, ou em qualquer outro formato). Caso o formato do documento seja HTML, o documento é processado pelo agente buscador, e o resultado é representado através de um vetor de termos; caso contrário, o documento é descartado. Esse processamento de documentos em formato HTML, dentre outras ações, remove as *stopwords* (artigos, preposições, etc), TAGs do formato HTML, e mantém somente os radicais das palavras restantes. Ou seja, o vetor de termos é um vetor que contém somente os radicais, não repetidos, do texto obtido no documento. Deste modo, se em um texto ocorrem as palavras masculino e masculinidade, por exemplo, essas palavras seriam reconhecidas como duas palavras iguais — masculin — e possuiriam somente uma referência no vetor de termos, com frequência igual a 2 (dois), supondo que o radical masculin não conste no texto nenhuma outra vez.

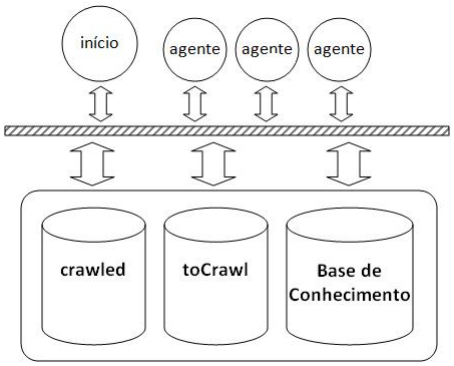

Figure 4. Processo de busca de páginas na Web.

No passo seguinte — Passo 3 — o agente compara o vetor obtido no Passo 2 com o assunto de interesse, utilizando a base de conhecimento. Essa base de conhecimento possui informações que direcionam o sentido da busca, contendo termos que podem existir em documentos que estejam no contexto buscado. O resultado da comparação é um valor que representa a relevância do documento analisado em relação à base de conhecimento, que representa o objeto da busca. Ou seja, quanto maior este valor, maior é a probabilidade do documento analisado estar no contexto do objeto da pesquisa. Caso este valor seja menor que um limite inferior aceitável (*threshold*), parâmetro a ser definido, o documento é descartado, pois provavelmente não está no contexto do objeto da busca. Vale ressaltar que os links para os quais este documento aponta serão também desconsiderados pelo mesmo motivo do documento ter sido descartado. Caso o valor seja maior que o limite inferior aceitável, as informações relevantes do texto são agregadas à Base de Conhecimento (como o vetor dos radicais mais encontrados, por exemplo), o que está representado no Passo 4. No Passo 5, o agente coleta os links existentes no documento analisado e alimenta o banco de dados toCrawl. Alimenta também com informações do documento analisado o banco de dados crawled. Após o agente passar por todos os passos (ou caso tenha descartado o documento no Passo 3), este volta a solicitar à base de dados toCrawl uma nova URL para dar continuidade à busca.

Pelo fato dos agentes serem independentes uns dos outros, eles podem ser dis-

tribuídos em vários computadores em redes diferentes, com acesso a Internet. Vários agentes trabalhando ao mesmo tempo podem aumentar significantemente a velocidade de obtenção dos resultados.

5. Um Estudo de Caso

O objetivo do sistema buscador que propomos neste trabalho é realizar buscas na internet por páginas que cometam alguns tipos de crimes, tais como a venda de drogas psicoativas. A maconha e o haxixe, substâncias provenientes da planta *Cannabis Sativa*, são substâncias que não podem ser comercializadas no Brasil. Porém, há páginas na internet que oferecem esse tipo de produto para ser comprado. Para fins de estudo de casos, foram buscadas páginas que continham palavras relacionadas à maconha, com os termos maconha, baseado e erva, tanto no sistema de buscas do Google[4], quanto no site de relacionamentos Orkut[5]. Utilizamos os mecanismos de busca em ambos os sites pelos termos mencionados, para analisar se conseguiríamos recuperar páginas contendo conteúdos ilegais — venda de narcóticos, nesse caso. Em cerca de duas horas de buscas manuais no Orkut, foram encontradas pouco mais de 40 comunidades que fazem algum tipo de apologia a este narcótico. Há discussões no Brasil a respeito da possível legalização da maconha. Assim, há algumas comunidades que são mais incisivas em relação à legalização, e ainda há outras comunidades com pessoas que, conforme informações da comunidade, declaram ser usuários. Já na pesquisa realizada no site de buscas da Google, obtivemos como primeiros resultados sites com informações relacionadas à essa droga psicoativa. Então, ao procurar por maconha, foram retornados sites como Wikipedia, BrasilEscola e Abril. Ao mudarmos os termos de busca para vende maconha, obtivemos diversos links de notícias. Alterando para baseado, obtivemos diversas informações menos direcionadas para o objetivo do trabalho. Somente ao procurar erva maconha vende, obtivemos links com alguma proximidade com o procurado. Em um dos sites, encontramos um vídeo que ensina como cultivar a planta Cannabis Sativa — `http://www.caixapreta.blog.br/?p=2305`. No entanto, como há uma grande discussão a respeito das propriedades medicinais da Cannabis Sativa[6], não se pode afirmar que o site mencionado comete algum tipo de crime. Para concluir, nosso estudo refletiu a dificuldade de encontrar páginas de venda de drogas ilegais. Por isso, acreditamos que utilizando ordenações inversas às praticadas pelos buscadores tradicionais, e agregando semântica em nosso sistema de busca, conseguiremos guiar de maneira satisfatória a busca por esse tipo de material.

6. Considerações Finais

Neste trabalho, fazemos uma breve descrição sobre o desenvolvimento de buscadores na Web, assim como descrevemos abordagens propostas na literatura para desenvolvimento de buscadores focados. Para podar a busca, diferentes métodos são propostos na literatura, porém geralmente os métodos propostos tendem a recuperar páginas da Web que são tipicamente acessadas pelos usuários. Entretanto, ao realizarmos um estudo com buscadores na Web em geral, não conseguimos recuperar facilmente materiais ilícitos relacionados

[4] `http://www.google.com.br`

[5] `http://www.orkut.com.br`

[6] Algumas informações a respeito dessas propriedades podem ser encontradas em `http://www.unifesp.br/dpsicobio/boletim/ed51/1.htm`.

ao uso e venda de drogas, somente apologias às mesmas. Assim, propomos um buscador focado com arquitetura de busca paralela para agilizar o processo, pretendemos também implementar em nosso buscador métodos para realização de busca por materiais ilícitos, levando em consideração a questão da busca por esse tipo de material se diferenciar do material rotineiramente procurado por usuários na internet. Atualmente, temos implementado um buscador para recuperar páginas e realizar manipulações básicas. Como continuação desse trabalho, implementaremos um sistema computacional que permita modificar facilmente a estratégia para definição de similaridade das páginas com o tema da busca. Pretendemos realizar estudos mais detalhados em relação à dificuldade de se localizar material do tipo que propomos encontrar.

References

Bergman, M. (2001). The deep web: Surfacing hidden value. *The Journal of Electronic Publishing*, 7(1).

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. In *Proceedings of the 9th World Wide Web Conference (WWW9)*, Amsterdam, Netherlands. `http://www9.org/w9cdrom/160/160.html`.

Cho, J. and Garcia-Molina, H. (2003). Estimating frequency of change. *ACM Transactions on Internet Technology (TOIT)*, 3(3):256–290.

Davison, B. (2000). Topical locality in the web. In *SIGIR'00: Proc. 23rd annual intern. ACM SIGIR conf. on research and development in information retrieval*, pages 272–279.

Demetriou, C. and Silke, A. (2003). A criminological internet sting: Experimental evidence of illegal and deviant visits to a website trap. *British Journal of Criminology*, 43:213–222.

Menczer, F., Pant, G., and Srinivisan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Transaction on Internet Technology*, 4(4):378–419.

Micarelli, A. and Gasparetti, F. (2007). *The Adaptive Web — Methods and Strategies for Web Personalization (LNCS)*, volume 4321, chapter Adaptive Focused Crawling, pages 231–262. Springer Verlag.

Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007a). *The Adaptive Web — Methods and Strategies for Web Personalization (LNCS)*, volume 4321, chapter Personalized Search on the World Wide Web, pages 195–230. Springer Verlag.

Micarelli, A., Sciarrone, F., and Marinilli, M. (2007b). *The Adaptive Web — Methods and Strategies for Web Personalization (LNCS)*, volume 4321, chapter Web Document Modelling, pages 155–192. Springer Verlag.

Pant, G., Srinivasan, P., and Menczer, F. (2004). *Web Dynamics*, chapter Crawling the Web, pages 1–25. Springer Verlag.

Webb, G. I., Pazzani, M. J., and B., D. (2001). Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):19–29.