# 5
# *Crawler* Focado

No presente capítulo são apresentadas e detalhadas as técnicas e heurísticas envolvidas no processo de coleta específica de dados baseado em *crawler* focado. A pesquisa sobre este assunto constitui o que é considerado, atualmente, o estado da arte em coleta de dados.

## 5.1. Definição

Trabalhar de forma a cobrir toda a *Web* pode ser bastante custoso, porém muitas aplicações não precisam indexá-la de forma integral. Muitos usuários adotam uma estratégia de filtragem por relevância e qualidade (DOM, CHAKRABARTI, & BERG, 1999). A relevância, por ser uma questão particular, pode ser usada para reduzir o contexto da busca de páginas por parte de um usuário. Desta forma, o usuário poderia manter a qualidade de suas consultas nesse universo reduzido de possibilidades. Por exemplo, suponhamos que um usuário seja interessado num tópico bastante específico. A tendência é que esse usuário mantenha uma lista de *sites* considerados por ele como mais importantes sobre esse tópico e navegue por outros *sites* que se encontrem nas redondezas daqueles. Dessa maneira apesar da *Web* estar passando por um crescimento estrondoso, dentro de uma determinada área de interesse esse crescimento é menor e a maioria dos usuários está constantemente realizando busca dentro de um determinado nicho. Transferindo essa tarefa para um rastreador automático podemos pensar no conceito de *crawlers* focados.

Um *crawler* focado é uma ferramenta automática que rastreia a *Web* seletivamente em busca de páginas relevantes a um critério definido *à priori* (DOM, CHAKRABARTI, & BERG, 1999). Definir o que é, ou não, relevante constitui a principal tarefa de um classificador, elemento obrigatório de qualquer *crawler* focado. Uma arquitetura relativamente mais complexa, com elementos

adicionais além do classificador pode ser vista em (DOM, CHAKRABARTI, & BERG, 1999). O *crawler* apresentado, ilustrado na Figura 38, utiliza uma taxonomia de tópicos com amostras para esses tópicos como sendo a base para seu aprendizado. A categorização de relevância é dada por um classificador, embutido no sistema. O usuário seleciona um tópico específico, ou seja, um nó na árvore de taxonomia, e pode ainda prover endereços a partir dos quais o *crawler* possa iniciar sua tarefa de rastreamento. É interessante que essa varredura possua algum tipo de estratégia de forma a selecionar muitas páginas de interesse sem que para isso necessite visitar muitos *sites*. O usuário pode ainda participar do refinamento do aprendizado, modificando a classificação automática de documentos indexados, mudando documentos de lugar na árvore de taxonomia. Essa possibilidade de configuração torna os critérios de relevância bastante personalizados e calibrados por requisitos específicos de uma ontologia determinada pelo usuário. Além disso, o *crawler* proposto por (DOM, CHAKRABARTI, & BERG, 1999) agrega outro elemento, o destilador, que opera periodicamente com o objetivo de priorizar os *links* a serem visitados. Ele determina medidas de centralidade das páginas rastreadas no intuito de estabelecer prioridade nas visitas.

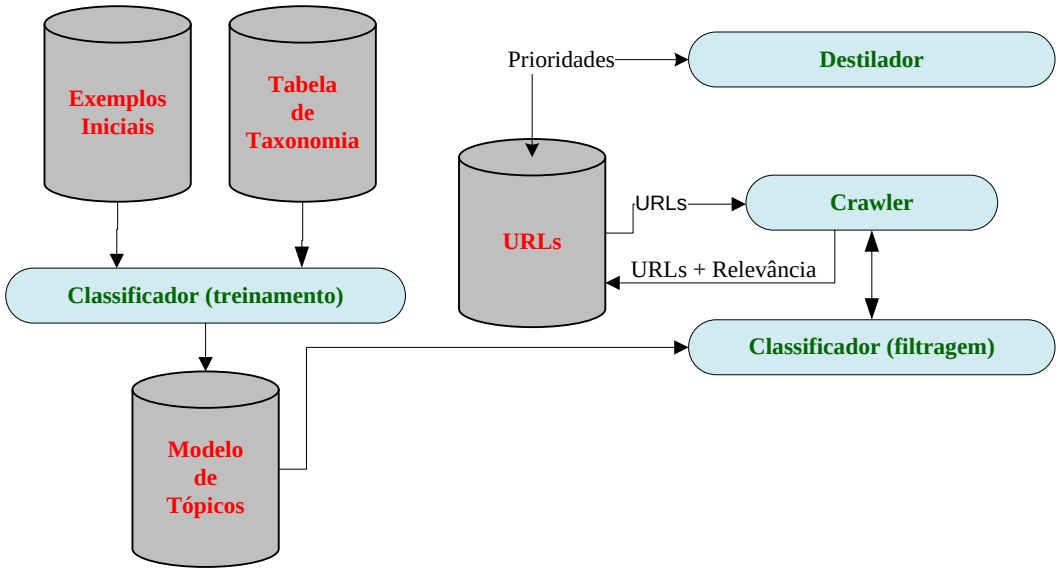


Figura 38 – Arquitetura de um *crawler* focado dotado do elemento Destilador

O *crawler* focado ideal recupera o conjunto maximal de páginas relevantes enquanto simultaneamente atravessa o número mínimo de documentos na Web (DILIGENT, COETZEE, LAURENCE, GILES, & GORI, 2000). A chave para o sucesso de um *crawler* focado é capacitá-lo a selecionar os *links* mais relevantes para prosseguir com as visitas subseqüentes na ordem pré-definida, com respeito ao tópico a ser buscado (ARDÖ, 2005).

## 5.2. Exames sobre pesquisas na área

A arquitetura e os critérios de decisão sobre relevância variam de acordo com a implementação. Diversas pesquisas propõem tipos distintos de abordagens para criação de *crawlers* focados. O *crawler* descrito em (DOM, CHAKRABARTI, & BERG, 1999) possui dois módulos, um classificador e um destilador, que julga a relevância da página e determina a ordem de visita dos *links*, respectivamente. A decisão é baseada sobre a relevância das páginas e sua potencial proximidade a outras páginas relevantes. (DILIGENT, COETZEE, LAURENCE, GILES, & GORI, 2000) usa a estrutura de *links*, também chamada de grafo de contexto, como uma forma de definir o percurso de visita às páginas relevantes. (NAJORK & WIENER, 2001) reporta que a estratégia *breadth-first* apresenta bom desempenho, para *crawlers* de baixa escala, quando o conjunto de sementes é apropriado. Algumas abordagens baseadas em técnicas de Aprendizado de Máquina também têm sido sugeridas como em (MCCALLUM, NIGAN, RENNIE, & SEYMORE, 1999), que adota aprendizado por reforço e aplica classificadores Bayesianos para o texto inteiro e texto âncora da página visitada. Já em (XU & ZUO, 2007) é dito que as decisões de visita a uma página confiam exclusivamente em indícios de relevância extremamente difíceis de serem explorados por abordagens tradicionais de Aprendizado de Máquina. Ele sugere uma perspectiva de aprendizado relacional motivada pelo fato de que todos os indícios de relevância de uma página não visitada possuírem natureza relacional: um *hyperlink* relaciona-se com a página em questão através de uma âncora que ocorre dentro de uma página baixada, e a âncora, por sua vez possui um relacionamento estrutural adicional com outros elementos HTML e etc. (XU

& ZUO, 2007) utiliza predicados para modelar os relacionamentos e alimentar uma base de conhecimento de forma a extrair regras de classificação baseadas em lógica de primeira ordem.

## 5.3. Estratégias

Para executar tarefas de varredura na *Web* de forma focada as mais diversas estratégias podem ser empregadas. A mais simples delas é conhecida como *breadth-first*, que não utiliza nenhuma heurística para decidir qual a próxima *URL* da pilha será visitada. Essa estratégia realiza uma visita tanto em largura como em profundidade no grafo e não difere da abordagem usada pelos *crawlers* comuns. Porém (NAJORK & WIENER, 2001) mostra que essa estratégia pode obter bons resultados para os casos de varredura em pequena escala e tendo como ponto de partida sementes cuidadosamente escolhidas. Isso pode ser explicado através de teorias de modelo de redes sociais, que mostram que a *Web* se constitui de pequenas comunidades, ou seja, de páginas que versam sobre o mesmo assunto (GIBSON, KLEINBERG, & RAGHAVAN, 1998). Essas comunidades web, verificáveis através de um grafo constituído pelos *links* pertencentes às páginas como sendo arestas e às páginas como nós são explicadas em grande parte pela existência de hubs, que são páginas que concentram uma grande quantidade de ligações para outras páginas e funcionam como bons pontos de partida para novas pesquisas dentro do tópico específico (KLEINBERG, 1998). É calcado na existência dessas comunidades que um *crawler* comum pode trazer resultados equivalentes aos de um *crawler* focado para buscas locais. A utilização de *breadth-first* produz resultados com baixa abrangência e por isso devem-se tentar estratégias mais eficientes do tipo *best-first*, que estabeleçam alguma heurística para ordenação e decisão das visitas aos *hyperlinks* das páginas web.

De forma didática pode-se dizer que para manter o escopo de varredura dentro do domínio desejado um crawler focado pode contar com dois tipos de algoritmos ou até mesmo com alguma forma de combinação das duas idéias. São elas: *Web Analisys*, que julga relevância e qualidade das páginas apontadas por

uma *URL* alvo; e *Web Search*, que determina a melhor ordem em que as *URL* alvo serão visitadas.

## 5.4. Web Analisys

Pode-se categorizar esse tipo de abordagem em duas vertentes distintas, uma com a análise apoiada no conteúdo da página (*content-based*) e outra calcada apenas nas ligações (*link-based*).

Nos algoritmos de análise de conteúdo, técnicas inteligentes de mineração de textos e indexação, com análise textual e extração de palavras chave são usadas para determinar se o conteúdo da página analisada é relevante ou não. Por exemplo, o confrontamento de palavras chaves com uma lista de termos representativa do domínio pode ser utilizado. Pesos diferentes podem ser atribuídos aos termos ou até mesmo rótulos HTML específicos, com intuito de estabelecer uma formulação mais precisa do conceito de relevância.

Outras abordagens têm dado grande importância à estrutura de ligações entre as páginas e dessa forma enfatizado a utilização dos *links* para estabelecimento de critérios de relevância. (DAVISON, 2000) mostra que pares de páginas, conectados através de um *link*, escolhidos aleatoriamente na Internet, possuem conteúdo textual de grande similaridade. Isso pode ser interpretado pelo fato de um autor de página expor ligações a outras páginas que ele julgue que sejam relevantes ao conteúdo da sua página, assim como acontece com as citações em pesquisas científicas. Dessa forma, análises com esse enfoque utilizam o texto das âncoras das páginas para estabelecimento dos critérios de relevância da página a ser visitada e o fundamento disso está no fato dos autores utilizarem esse texto como uma forma de descrever a página. (AMITAY, 1998) apresenta um estudo que leva em consideração também o texto próximo ao texto das âncoras.

## 5.5. Web Search

Essa categoria de algoritmos está preocupada com a melhor ordem de visitação e aqui se enquadram as estratégias *best-first* em oposição às *breadth-first*, que conforme já comentado, podem também ser utilizadas em *crawlers* focados. Isso é possível desde que a busca não se estenda muito em profundidade, já que na mesma medida que se afasta da semente, que deve ser cuidadosamente escolhida, introduz-se mais ruído aos resultados com respeito ao tópico a ser buscado. Essa abordagem pode ainda ser utilizada aliada a algum algoritmo de análise de conteúdo e nesse caso a página só seria indexada caso cumprisse os critérios de relevância. Isso não traz nenhuma melhoria de desempenho durante a varredura, porém mantém o índice restrito ao domínio do assunto buscado.

Em contraste às políticas baseadas em *breadth-first* têm-se as *best-first*, onde alguma heurística, normalmente baseada em algum método de *web analisys*, determina a melhor ordem para visitação. A pilha de *links* que o *crawler* irá visitar é então formada por aqueles supostamente mais promissores em termos de relevância sendo posicionados primeiro e os não-promissores indo para o final da fila e conseqüentemente tendo raras chances de serem visitados e até mesmo sendo descartadas dependendo da implementação. Porém alguns estudos mostram que essas políticas também acabam resultando em alguns problemas. (BERGMARK, LAGOZE, & SBITYAKOS, 2000) aponta o mesmo problema de busca local e conseqüente baixa abrangência na coleção final de documentos indexados, já que o espaço de busca está limitado à vizinhança das páginas visitadas, o que acaba restringindo as buscas dentro das comunidades relacionadas às sementes utilizadas.

## 5.6. Estratégias Adicionais

Com o objetivo de fugir dos limites de uma comunidade *web* algumas técnicas podem ser adicionadas resultando numa abrangência maior das páginas captadas. Uma delas é simplesmente permitir certo grau de ruído, ou seja, visitar páginas não relevantes a fim de possivelmente adentrar por uma nova comunidade

relacionada ao tópico buscado. Além disso, essas páginas visitadas e não relevantes não precisam ser indexadas. Detalhes dessas técnicas, conhecida como *tunneling*, podem ser encontrados em (BERGMARK, LAGOZE, & SBITYAKOS, 2000).

Outras técnicas incluem avaliação de métricas de co-citação exemplificada na Figura 39. Duas páginas possuem a relação de co-citação se forem apontadas por uma mesma terceira página. É bastante comum, até mesmo dentro de uma comunidade *web* páginas que compartilhem essa propriedade. Principalmente em domínios comerciais, onde normalmente uma empresa não faz referência a uma concorrente. Além disso, pesquisadores têm demonstrado que, freqüentemente, quando existem ligações entre páginas que pertençam a comunidades *web* diferentes, esses *links* possuem um único sentido, de uma comunidade para a outra, sendo muito pouco freqüente o inverso. E nos casos em que existe um nó externo às comunidades a ligação entre elas também costuma ser no sentido do nó para dentro das comunidades, sendo o inverso muito menos freqüente (TOYODA & KITSUREGAWA, 2001).

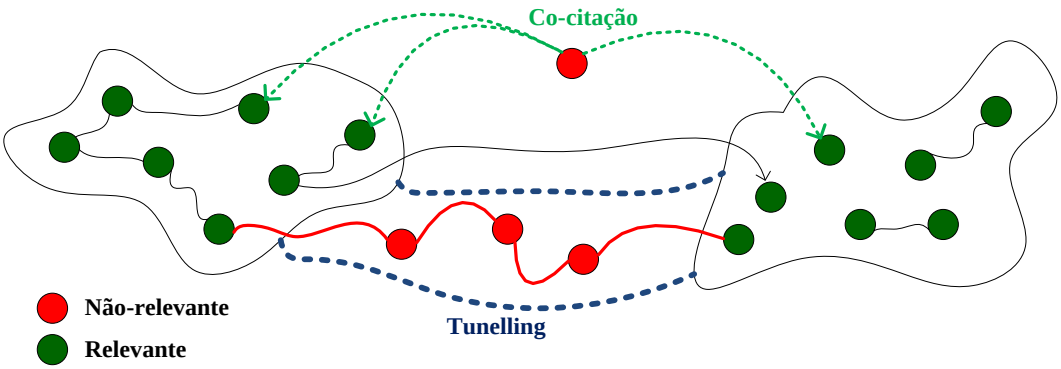


Figura 39 - Relacionamentos de co-citação em uma comunidade *Web*