4
Recuperação de Informação

No presente capítulo são apresentados os fundamentos da área de Recuperação de Informação utilizados em Mineração de Textos, como por exemplo, os modelos de representação de documentos e as principais operações envolvidas nestes processos. A área de Recuperação de Informação na Internet também é abordada, com foco exclusivo para a *Web*.

4.1.
Introdução

Recuperação de Informação lida com a representação, armazenamento, organização e acesso a itens de informação (BAEZA-YATES & BERTIER, 1999) (MANNING, RAGHAVAN, & SCHÜTZE, 2007). O conceito de itens de informação, neste contexto, refere-se ao tratamento diferenciado que estes objetos, geralmente documentos textuais, recebem: todos possuem muita ou pouca relevância. Julga-se um documento relevante quando este supre a necessidade de informação do usuário. Relevância, a característica central de Sistemas de Recuperação de Informação, é o que distingue Sistemas de Recuperação de Informação de Sistemas de Recuperação de Dados.

Recuperação de Dados busca meios eficientes de recuperar objetos baseado em um critério simples: dado o conjunto de termos desejado, encontrar todos os documentos que atendam ao critério booleano determinado. E isto é suficiente para muitas aplicações, como por exemplo, Sistemas Gerenciadores de Bancos de Dados. Mas, para um usuário que deseja informações sobre um determinado tópico, a consulta baseada em termos nem sempre trará somente bons resultados, ou seja, nem sempre será relevante. A Tabela 4 apresenta algumas das diferenças entre Recuperação de Dados e Recuperação de Informação (RIJSBERGEN, 1979).

Características	Recuperação de Dados	Recuperação de Informação
Comparação	Exata	Aproximada
Dados	Fortemente estruturados	Fracamente estruturados
Inferência	Dedução	Indução
Modelo	Determinístico	Probabilístico
Ling. Consulta	Artificial	Natural
Esp. da Consulta	Completa	Incompleta

Tabela 4 - Comparação entre Recuperação de Dados x Recuperação de Informação

Usuários de Sistemas de RI estão mais interessados na recuperação de informação associada a documentos do que na recuperação dos termos presentes nestes. Com o crescimento do volume de publicações, ao longo dos anos, foram desenvolvidas técnicas específicas para a área de Recuperação de Informação com o intuito de atender às necessidades dos usuários.

A ferramenta mais importante para auxiliar o processo de recuperação de informação é denominada índice. Índices são estruturas de dados associadas à parte textual dos documentos, e, portanto, indicam o local onde a informação desejada pode ser localizada. Segundo (BAEZA-YATES & BERTIER, 1999), há aproximadamente quatro mil anos já são praticadas técnicas de catalogação manual por índices.

Recuperação de Informação, antes, interesse de poucos, agora, é uma das áreas que mais tem recebido atenção de cientistas e pesquisadores. Contribuiu principalmente para isto a explosão demográfica da *Web* que é de longe o maior acervo de dados do mundo (CHAKRABARTI, 2003). E na *Web*, prevalecem os documentos hipertextos que, em sua essência, constituem o objeto de estudo de RI: documentos textuais.

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configura como uma área de significância maior (CARDOSO, 2000).

Porém, além de muito sucesso, a *Web* também trouxe novos desafios para a área de RI. Por ser um ambiente onde impera a cultura liberal e informal de propagação de conteúdo, encontrar informação relevante na *Web* tem sido cada

vez mais difícil, motivando também uma grande pesquisa em torno da Recuperação de Informação na Internet, em especial, na *Web* (ver item "4.4").

4.2.
Histórico da área de Recuperação de Informação

4.2.1.
1ª Fase – Décadas de 50 e 60

Sistemas de Recuperação de Informação foram originalmente utilizados para gerenciar a explosão de conteúdo da literatura científica na segunda metade do século XX (RIJSBERGEN, 1979). Bibliotecas estão entre as primeiras instituições a adotarem Sistemas de RI.

Em suas primeiras versões, Sistemas de RI funcionavam como um simples catálogo eletrônico. O processo de indexação era basicamente manual e os documentos eram indexados somente pelos termos principais de um dicionário de sinônimos criado para este propósito: um dicionário *thesaurus* (ver item "3.3.2"). A idéia deste conceito é simples: permitir a indexação somente do termo principal sempre que o próprio termo ou termos sinônimos estiverem presentes em um texto, evitando assim, que a escolha de um sinônimo ou outro possa impedir a localização do documento. Já na década de 60, Sistemas de RI deram início ao processo de indexação automática, porém, somente título e abstract eram processados. Surgiram também os primeiros algoritmos de busca textual.

4.2.2.
2ª Fase – Décadas de 70 e 80

Neste período, houve grandes avanços na área tecnológica, o que resultou em aumento significativo do poder computacional da época, permitindo, também, a evolução de diversos sistemas, inclusive dos Sistemas de RI. Avanços como a indexação automática de todo o conteúdo e o desenvolvimento de funcionalidades adicionais de pesquisas foram possíveis.

RI – unida a área de Lingüística – iniciou os primeiros estudos de Processamento de Linguagem Natural possibilitando a criação de um sistema simples de perguntas-respostas (BAEZA-YATES & BERTIER, 1999). Foi também nesta fase que o modelo de representação de documentos mais utilizado foi criado: o Modelo de Espaço Vetorial.

4.2.3.
3ª Fase – Década de 90 em diante

Nesta fase, o grande crescimento da *Web* e a necessidade de informação relevante neste ambiente colocaram em foco novamente a área de RI. Inicialmente, técnicas tradicionais de Sistemas de RI foram utilizadas, porém, grandes foram os problemas encontrados na adaptação destas técnicas:

- Escalabilidade das soluções: escalabilidade, neste contexto, indica a capacidade de preparo para a manipulação de grandes quantidades dados, seja esta relacionada ao poder de processamento ou armazenamento.
- Velocidade de atualização das páginas-*web*: a incrível velocidade de modificação do conteúdo dos *web sites* torna difícil manter um índice operacional e coerente sem saber a freqüência de atualização dos documentos indexados.
- Velocidade de acesso aos documentos: em razão da sua distribuição geográfica mundial, a Web contém documentos nas mais diversas localidades. O acesso e indexação destes documentos exigem a disponibilidade dos mesmos, além do tempo necessário para que toda a informação neles seja transferida de um local para outro.

Atualmente, novas tecnologias estão sendo desenvolvidas para explorar as peculiaridades de um documento hipertexto e toda a sua relação n*a Web.*

4.3. Recuperação de Informação Clássica

Recuperar informação é o propósito básico de qualquer sistema Recuperação de Informação. Baseada em índices, a recuperação de informação nestes sistemas obedece à arquitetura ilustrada na Figura 26.

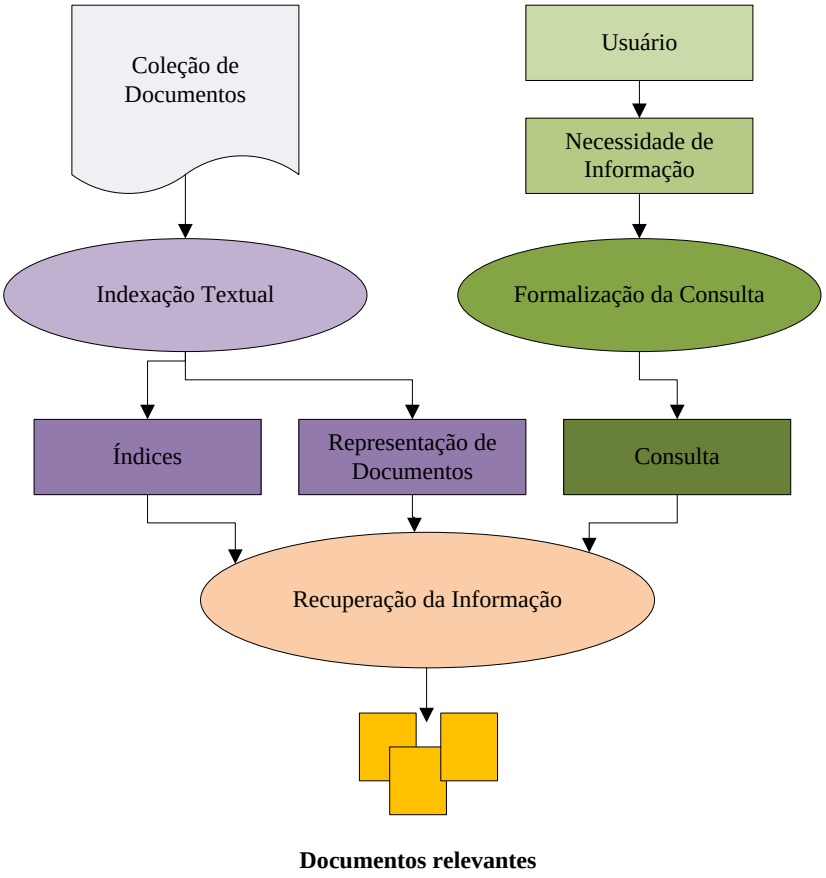

Figura 26 - Sistema Clássico de Recuperação de Informação

Neste modelo, duas entidades justificam a existência de um sistema de RI: a coleção de documentos, estes, geralmente textos, e o usuário com necessidade de informação. Os outros componentes decorrem destes.

A consulta é a representação formalizada da necessidade de informação do usuário em uma linguagem entendida pelo sistema. O processo de especificação da consulta geralmente é uma tarefa difícil. Há freqüentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada (CARDOSO, 2000). Essa distância é gerada pelo limitado

conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta.

Assim que formalizada, a consulta é processada junto aos documentos, que estão representados pelos seus respectivos modelos de representação textuais, e, em seguida, a resposta à necessidade de informação na coleção de documentos é exibida ao usuário. O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta (CARDOSO, 2000).

Os modelos de representação textuais utilizados em Sistemas de RI podem ser vistos como uma representação fortemente estruturada dos textos. E, como todo documento é considerado um conjunto de termos ou *tokens*, esta nova representação estruturada é baseada na presença ou ausência destes termos ou *tokens*.

Quando todo o conjunto de *tokens* de um documento é utilizado para representá-lo tem-se uma indexação textual completa ou *full text indexing*. Porém, embora a indexação textual completa seja aquela que forneça a visão lógica mais completa de um documento, nem sempre é possível utilizá-la, em razão do elevado custo computacional para o manuseio desta enorme quantia de dados, tornando necessário que um documento seja representado por um conjunto menor de *tokens*.

Como nem todas as palavras num texto não igualmente importantes para representá-lo semanticamente, para que seja bem representado por um conjunto menor de *tokens*, um documento pode ser submetido a sucessivos métodos de processamento textual, tais como remoção de *stopwords* e *stemming*, que visam eliminar conteúdo irrelevante do texto, permitindo que seja possível a representação lógica do mesmo. A Figura 27 ilustra algumas das possibilidades existentes em um processo de indexação (BAEZA-YATES & BERTIER, 1999).

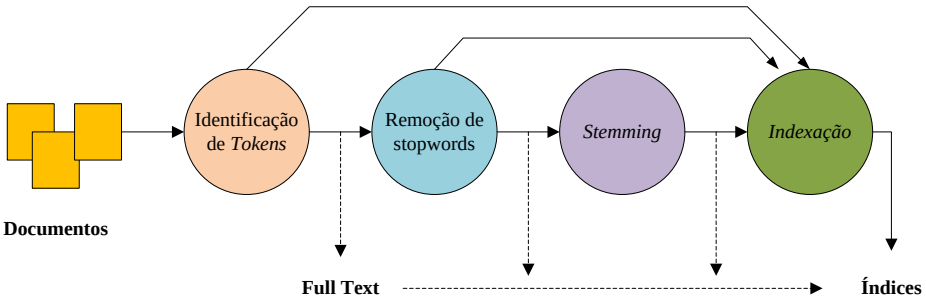

Figura 27 – Etapas possíveis no processo de Indexação de documentos textuais

Por utilizar seus próprios métodos de processamento textual, o interesse de Mineração de Textos na área de RI restringe-se às técnicas de representação e identificação de documentos. Muitos dos métodos de processamento textual utilizados em Mineração de Textos foram baseados naqueles utilizados em RI, e, portanto, foram abordados, sob o enfoque de Mineração de Textos, no item "3.2".

A seguir, serão apresentados dois modelos de representação de documentos utilizados, tanto em RI, como em Mineração de Textos: **Modelo de Recuperação Booleano**[18] e Modelo de Espaço Vetorial.

4.3.1. Modelo de Recuperação Booleano

Um dos primeiros modelos de pesquisa a ser adotado foi o Modelo de Recuperação Booleano ou, simplesmente, Modelo Booleano. Fundamentado na Álgebra Booleana e na Teoria dos Conjuntos, interpreta toda consulta como uma expressão lógica, permitindo até mesmo a utilização dos conectivos lógicos "e", "ou" e "não", e, portanto, possui critério de decisão simples para julgar a relevância de um documento: documentos relevantes são aqueles que contêm, ou não, os termos que satisfazem a expressão lógica da consulta.

Em virtude do critério de decisão binário deste modelo não existem meios para a realização de igualdade parcial da consulta com os documentos. Portanto, também não existem critérios de graduação de relevância dos documentos

[18] Do termo inglês, *Boolean retrieval model*.

encontrados, ou seja, não é possível ordenar documentos de acordo com a relevância individual de cada um.

Este modelo de representação é muito mais utilizado em Sistemas de Recuperação de Dados do que em Sistemas de Recuperação de Informação. É de fácil utilização para usuários que dominam lógica booleana, o que não ocorre na maioria dos casos.

Algumas das vantagens do modelo booleano são a excelente *performance* e a fácil implementação. Possui como principal desvantagem a dificuldade de se expressar a necessidade de informação por meio de uma expressão booleana. Outra característica ruim deste modelo é desconsiderar a freqüência de ocorrência dos termos em um texto.

4.3.2. Modelo de Espaço Vetorial

O Modelo de Espaço Vetorial busca abordagem geométrica para resolver problemas de representação de documentos. Documentos são representados como vetores em um espaço Euclidiano *t-dimensional* em que cada dimensão corresponde a um *token* da coleção de documentos (REZENDE, 2005), ou seja, cada *token* é um eixo deste espaço Euclidiano.

Neste modelo, vetores são representados pela forma $D_i = (t_1; t_2; t_3; ...; t_n)$, em que D_i é o *i-ésimo* documento de uma coleção, e t_n o *n-ésimo token* da coleção de documentos, ou seja, para cada documento da coleção existem n *tokens*-índices que os representa (SILVA A. A., 2007), conforme ilustrado na Figura 28 . Cada *token* desta coleção de documentos está associado a sua freqüência de ocorrência em cada documento, desta forma, para o documento D_i e para o token t_j, $w_{i,j} \geq 0$ representa essa associação e o tamanho do eixo$_j$ no vetor D_i. Quando o *token* j não ocorre no documento D_i, tem-se $w_{i,j} = 0$.

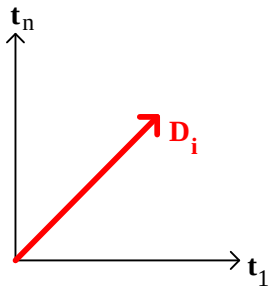

Figura 28 - Representação Vetorial do Documento Di no espaço n-dimensional

No Modelo de Representação Vetorial, o processamento de uma consulta é realizado através de um cálculo de similaridade entre cada documento da coleção e a própria consulta, ou seja, toda consulta é também representada de forma vetorial, e através de um cálculo de similaridade entre cada documento da coleção e a consulta, obtém-se uma lista dos documentos relevantes para aquela necessidade de informação.

O Modelo do Espaço Vetorial é o modelo de representação mais utilizado em Mineração de Textos (REZENDE, 2005). Contribuem para isto a sua forma de representação, intuitiva e prática, que torna possível:

- A ponderação de termos na representação dos documentos e processamento das consultas;
- A recuperação de documentos que não possuem todos os termos definidos na consulta;
- Ordenação do resultado baseada na relevância dos documentos.

Desvantagens deste modelo de representação são a necessidade de novo processamento da coleção de documentos quando esta é alterada e a ausência de relação semântica entre os *tokens* de uma coleção.

4.3.2.1. Freqüência dos termos

No Modelo de Espaço Vetorial, cada documento é representado por um vetor cujas dimensões são os termos presentes na coleção de documentos. Cada coordenada do vetor é um termo da coleção de documentos e possui valor

numérico que representa a freqüência de ocorrência deste termo no documento (LOPES, 2004).

A associação de valores numéricos as coordenadas dos vetores é conhecida como **atribuição de pesos**[19] e visa atribuir maior importância aos termos que são mais relevantes. A seguir, são citadas e explicadas as medidas de atribuição de pesos mais comuns:

- Binária: Quando um termo está presente em determinado documento, é atribuído o valor *true* ou um para indicar esta ocorrência. Quando um termo está não presente em determinado documento, é atribuído o valor *false* ou zero para indicar esta ausência. Por ser muito simples, esta medida de atribuição de pesos é raramente utilizada.
- Freqüência do Termo: **Freqüência do Termo**[20] ou **TF**[21] é definida como o número de ocorrências de um determinado termo em um documento (SALTON & BUCKLEY, 1988). Em geral, termos presentes em muitos documentos com alta freqüência não possuem caráter discriminatório para a diferenciação dos documentos de uma coleção e são considerados como uma *stopword*. É comum normalizar em um documento a freqüência de seus termos, pois, sem este artifício, os documentos mais extensos de uma coleção seriam privilegiados no processo de recuperação de informação. Na Figura 29 é ilustrado o cálculo normalizado da freqüência do Termo_i no Documento_j que possui k termos.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Figura 29 – Cálculo da medida TF em um documento

- TF-IDF: TF-IDF ou *Term Frequency – Inverse Document Frequency* é uma medida de atribuição de pesos que favorece termos

[19] Do termo inglês, *weighting*.

[20] Do termo inglês, *Term Frequency*.

[21] Acrônimo de *Term Frequency*.

que ocorrem em poucos documentos de uma coleção (SALTON & BUCKLEY, 1988). É utilizada para avaliar o quão importante é um termo para o documento em que ele ocorre, em relação a todos os documentos da coleção. A medida TF-IFD de um termo, ilustrada na Figura 30, é a combinação de sua medida local (TF) e global (IDF).

$$\boldsymbol{tf - idf_{i,j} = tf_{i,j} \times idf_i}$$

$$idf_i = \log \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|}$$

$$|D|, \textit{número total de documentos da coleção};$$
$$\left|\{d_j : t_i \in d_j\}\right|, \textit{número de documentos em que o termo } t_i \textit{ ocorre};$$

Figura 30 – Cálculo da medida TF-IDF em um documento

4.3.2.2. Cálculo de Similaridade

No Modelo do Espaço Vetorial cada documento é representado por um vetor de n dimensões, em que cada dimensão é um termo distinto e presente em algum documento da coleção. A cada termo é atribuído um peso como forma de identificar a importância deste no documento e para isto são utilizadas as medidas de atribuição de pesos mencionadas acima.

Uma das técnicas mais utilizadas para obter um valor de similaridade entre documentos ou entre documentos e consultas decorre naturalmente deste modelo de representação: é através do cosseno do ângulo formado pelos vetores de representação destes objetos (BAEZA-YATES & BERTIER, 1999).

O cálculo do cosseno do ângulo entre dois vetores é ilustrado na Figura 31. Quanto mais perto de um o valor do cosseno, mais ortogonais são os vetores comparados, o que significa que existem poucos termos comuns entre os documentos. Quanto mais perto de zero o valor do cosseno, mais paralelos são os vetores comparados, o que significa que existem muitos termos comuns entre os documentos.

$$\cos\theta = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\|\overrightarrow{v_1}\| \times \|\overrightarrow{v_2}\|}$$

Figura 31 – Similaridade entre dois documentos pela medida do Cosseno

4.4. Recuperação de Informação na Internet

4.4.1. Crawlers

Na *Web*, a coleta de dados pode ser realizada de forma automatizada através de *web crawlers* (HEATON, 2002). Um *crawler*, também conhecido como *web spider* ou *web robot*, é um robô que percorre a *Web* de forma automática e metódica. Dentre as ferramentas automáticas que trabalham na *Web* existem algumas distinções feitas por alguns autores, como em (HEATON, 2002). *Spiders*, *bots* e *aggregators* são todos chamados de agentes inteligentes, que executam tarefas na *Web* sem intervenção humana. *Bots* são programas que podem recuperar informações de locais específicos na Internet. *Spiders* são *bots* específicos que vão até a *Web* e identificam múltiplos *sites* com informações sobre um tópico escolhido e recuperam a informação. A*ggregators* são *bots* específicos capazes de reunir dados similares de múltiplos *sites* e consolidá-los em uma única página. De certa maneira todas as referidas ferramentas executam o trabalho de rastreamento e podem ser chamadas, de forma genérica, de *crawlers*, sendo este um termo mais utilizado como referência aos *spider* típicos, ou seja, os robôs responsáveis pela varredura de toda a Internet. Essa varredura é realizada a partir de endereços *Web* fornecidos como semente, seguindo o caminho fornecido pelos *links* encontrados nas páginas rastreadas. O nome *Spider* é uma alusão à aranha, que percorre um caminho na teia, ou seja, no grafo de *sites* da *Web*.

Este processo de percorrer a WWW é chamado de *web crawling* ou *spidering*. Muitas ferramentas oferecidas na *Web*, em particular as **máquinas de busca** (ver item “4.4.5”), utilizam o método de *spidering* como um meio de obter conteúdo. Freqüentemente, todo o conteúdo coletado por um *web crawler* é armazenado para que outros componentes possam fazer uso dele. Especificamente

no caso de máquinas de busca, este conteúdo será indexado para posterior uso no processo de consulta.

Crawlers também podem são utilizados em tarefas mais simples, como por exemplo, na verificação da hierarquia topológica de um *web site* ou para obter estatísticas do mesmo. Utilizando uma das ferramentas desenvolvidas para esta Dissertação, foi realizado um *crawling* no *site* do Departamento de Engenharia Elétrica da PUC-Rio. Com isto, foi constatado que existem 20 *links* inválidos em todo o *site*, além de outras estatísticas presentes na Tabela 5. A análise do w*ebgraph* deste *site*, ilustrado na Figura 32, permite também a visualização dos *links* das páginas deste *site*.

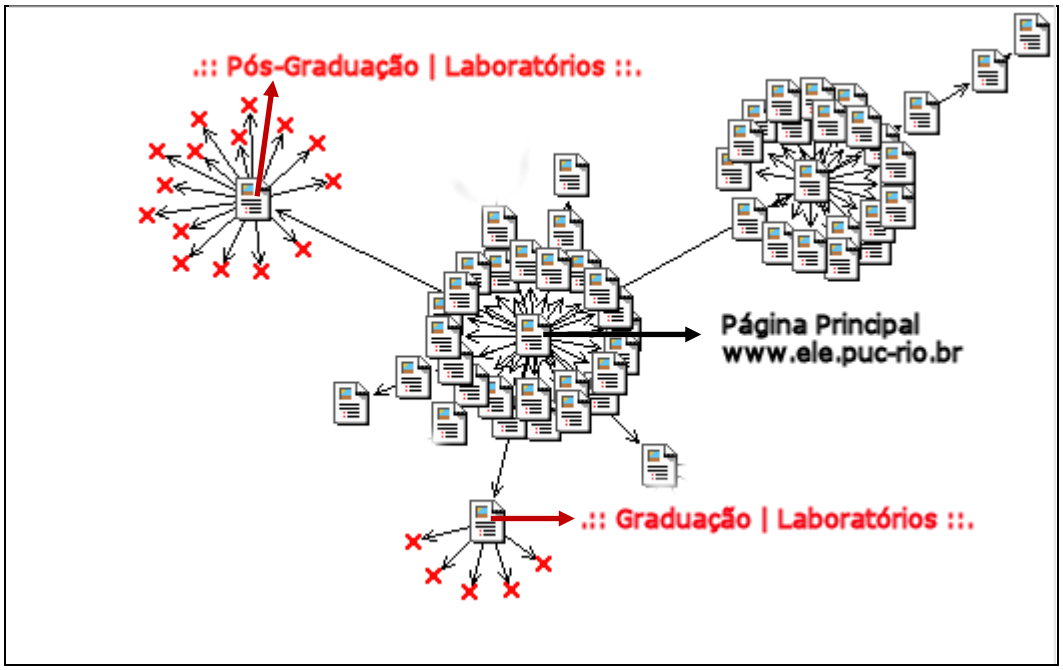

Figura 32 – *Webgraph* do *site* do DEE/PUC-Rio

Total *Links* Testados	793
Total páginas processadas	56
Links inválidos	20
Tamanho total *site* (HTML)(Mb)	4
Páginas processadas / seg	37,20

Tabela 5 – Estatísticas de *crawling* do *site* do DEE

O algoritmo de funcionamento de um *crawler* é simples: o processo de rastreamento é iniciado pela visita a um conjunto de ***urls***[22] fornecidos à priori e que recebem a denominação de sementes. Ao visitar um *url*, o *crawler* identifica todos os **hiperlinks**[23] desta página e adiciona todos eles a uma pilha de *URLs* a serem visitadas. Uma vez visitada todas as sementes, o crawler passa a visitar os *sites* armazenados na pilha de *URLs*. A partir de então, o processo pode seguir ininterruptamente. É o que está resumido na Figura 33.

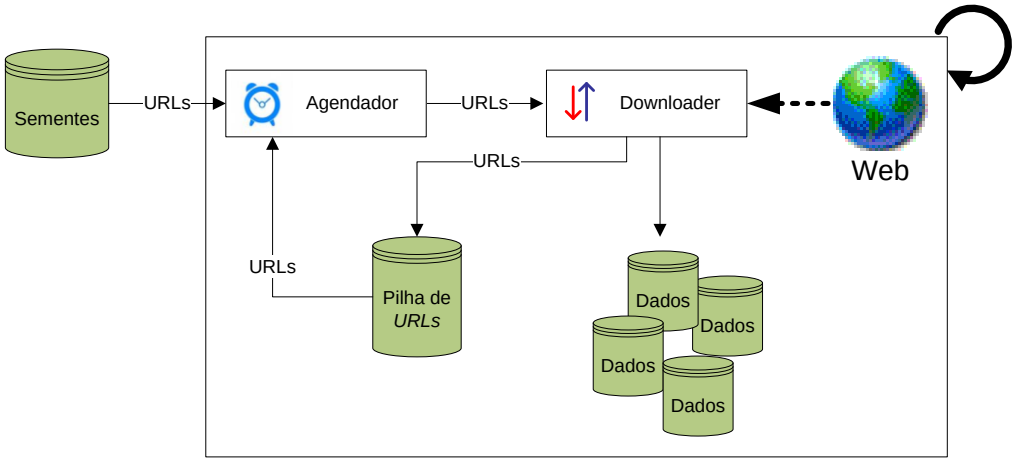

Figura 33 - Funcionamento de um crawler simples

A *World Wide Web* ou, simplesmente, *Web* é o mais ambicioso serviço de comunicação de dados e informações mediadas pela Internet (TANENBAUM, 2003). É formada por uma rede de documentos multimídia conectados por hiperlinks. Processa-se na forma gráfica, utilizando, para isso, um programa especial denominado navegador ou *browser* que permite a navegação entre as informações disponíveis nos computadores da rede. Os documentos multimídia são coloquialmente chamados de páginas e podem estar na forma de vídeos, sons, hipertextos e figuras. Navegar ou surfar na *Web* refere-se o ato de seguir as hiperligações de um documento para outro.

[22] Localizador Uniforme de Recursos. Acrônimo de *Uniform Resource Locator*.

[23] Do termo inglês, *hyperlink*. Também conhecido somente por *link*.

4.4.2. URL

O sistema de endereçamento de algumas redes, entre elas a Internet, é baseado em uma sintaxe definida para este propósito. Esta sintaxe é denominada de *Universal Resource Locator* ou, simplesmente, *URL*, e segue a seguinte estrutura: "protocolo://nome_do_computador/caminho_de_diretórios/recurso".

O campo protocolo designa o protocolo de acesso ao recurso. Poderá ser HTTP, FTP, entre outros. O campo nome_do_computador designa o servidor que disponibiliza o recurso desejado. O campo caminho_de_diretórios especifica o caminho de diretórios até o recurso. E, finalmente, o recurso ou arquivo a ser obtido.

4.4.3. Hiperlink

Um hiperlink é o apontamento ou referência, em um documento *web*, a outro documento ou recurso. Como tal, pode-se vê-la como análoga a uma citação na literatura. Ao contrário desta, no entanto, o *link* pode ser combinado com uma rede de dados e um protocolo de acesso adequado e assim ser utilizado para direto ao recurso referenciado (RICOTTA, 2007). E é este mecanismo que garante o funcionamento dos *crawlers*.

4.4.4. Políticas de *Web Crawling*

Existem três características importantes na *Web* que criam um cenário no qual o *web crawling* é muito difícil (SHKAPENYUK & SUEL, 2002):

- Grande volume de dados;
- Alterações de conteúdo muito rápidas;
- Geração de páginas com *URLs* dinâmicas.

O grande volume implica na priorização do conteúdo que será visitado e, futuramente, indexado, pois, o *crawler* só pode capturar uma pequena porção do universo de páginas da *Web*. Com uma estimativa de que menos de trinta por

cento de toda a *Web* seja visível (FUNREDES, 2007), isto é, esteja indexada, torna-se necessário definir o que será indexado. A maior parte das políticas de *crawling* das máquinas de buscas atuais baseia-se no critério de popularidade de um *site*. Quanto mais popular for um *site*, maior será a possibilidade deste ser indexado.

A enorme freqüência de alteração de conteúdo implica na dificuldade de determinar o tempo necessário para que um *site* seja revisitado. Como os recursos de armazenamento e processamento são limitados, quando comparados às necessidades da *Web*, visitar novamente um *site* é deixar de visitar um *site* novo. Por outro lado, manter um índice desatualizado é um grande convite a evasão dos usuários.

A geração de páginas com *URLs* dinâmicas é fruto das recentes tecnologias de *scripts server-side.* Esta tecnologia, além de outras funcionalidades, permite que se tenha para um mesmo conteúdo diversos endereços *URLs*. Por exemplo, uma simples galeria de fotos pode oferecer a visualização de um mesmo conteúdo de diferentes maneiras, como por exemplo, pela ordem de publicação das fotos, ou pela ordem de popularidade destas, o que resulta em mais de um *URL* para um mesmo conteúdo. Esta combinação matemática cria um problema para os *crawlers*, pois eles devem realizar infinitas combinações em mudanças de script para conseguir um conteúdo único (RICOTTA, 2007).

O comportamento de um *web crawler* é a combinação das seguintes políticas de seleção, re-visitação, cortesia e paralelização. A política de seleção interfere no modo com que as páginas a serem capturadas são selecionadas. A política de re-visitação fornece informações sobre a periodicidade com que se deve verificar por atualizações de um *site* já coletado. A política de cortesia reflete as precauções tomadas por um *crawler* para que este não sobrecarregue os servidores dos *web sites* que estão sendo coletados. A política de paralelização é responsável pela coordenação dos *web crawlers* distribuídos que atuam em um mesmo objetivo.

Como o objeto de interesse deste trabalho é a seleção de conteúdo na *Web*, apenas as políticas de seleção serão abordadas. Para ver mais sobre as outras políticas, consulte (RICOTTA, 2007) e (CASTILLO, 2004).

4.4.4.1. Políticas de Seleção de Conteúdo

Devido ao gigantesco tamanho da *Web*, até os maiores *search engines* cobrem apenas uma pequena porção dos documentos públicos existentes neste ambiente. Estima-se que menos do que trinta por cento de toda a *Web* esteja indexada (FUNREDES, 2007). Por razões de recursos de hardware e rede limitados, é desejável que toda esta pequena fração visita e indexada não seja composta de páginas aleatórias, mas, sim, de páginas populares e relevantes.

Isto requer a utilização de heurísticas para determinar quais páginas devem ser priorizadas no processo de visitação. Definir a importância de uma página *web* não é uma tarefa trivial, pois, muitos fatores estão envolvidos neste processo, como por exemplo, sua popularidade em termos de *links* externos que apontam para a própria, quantidade média de visitas que recebe e até mesmo a sua *URL*. Criar uma boa política de seleção possui uma dificuldade adicional: deve funcionar com parte da informação, pois o conjunto completo de páginas *web* não é conhecido durante a operação de *crawling* (RICOTTA, 2007).

As estratégias de seleção de conteúdo mais utilizadas são *breadth-first* e cálculo parcial de *pagerank*. Ambas as estratégias partem do princípio de que a *Web* é um grande grafo direcionado gerado pelas estruturas de *links* das páginas *web:*

- Cada página web é um vértice;
- Cada *hyperlink* entre páginas é um arco direcionado.

Breadth-first ou busca em largura primeiro é um algoritmo de procura de árvore usado para realizar uma busca ou travessia em uma árvore ou grafo. Funciona da seguinte maneira: inicia a busca pelo nó raiz e a cada vez que desce um nível na árvore, explora todo este nível antes de descer outro nível em busca do objetivo.

A seleção de conteúdo na *Web* baseada em *breadth-first* ou busca em largura primeiro tem funcionamento semelhante ao do algoritmo utilizado em árvores: toda vez que um *URL* é retirado da pilha, explora-se todo o conteúdo de mesmo nível (diretório), para então explorar conteúdo em níveis abaixo. O funcionamento deste algoritmo está ilustrado Figura 34.

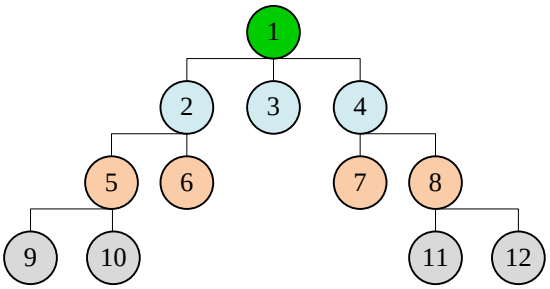

Figura 34 - Ordem de visitas dos *sites* utilizando a estratégia *breadth-first*

A estratégia de seleção de conteúdo baseada em *pagerank* é uma analogia a técnica acadêmica de citações literárias: quanto mais citações (*links* direcionados para) um livro (*site*) ou artigo tiver, mais importante ele é. Essa técnica fornece uma aproximação numérica da importância ou qualidade de uma página: essa importância se dá pelo número de *links* de qualquer lugar da *Web* para aquela página, sendo que, *links* apontados por páginas mais importantes valem mais pontos do que *links* de páginas menos importantes. Por exemplo, na Figura 35, o *web site* C possui menos *links* apontados para ele do que o *web site* E, porém, o *web site* C recebe apontamento de um *web site* de elevada pontuação (*site* B), e por isso recebe valor maior de *pagerank*.

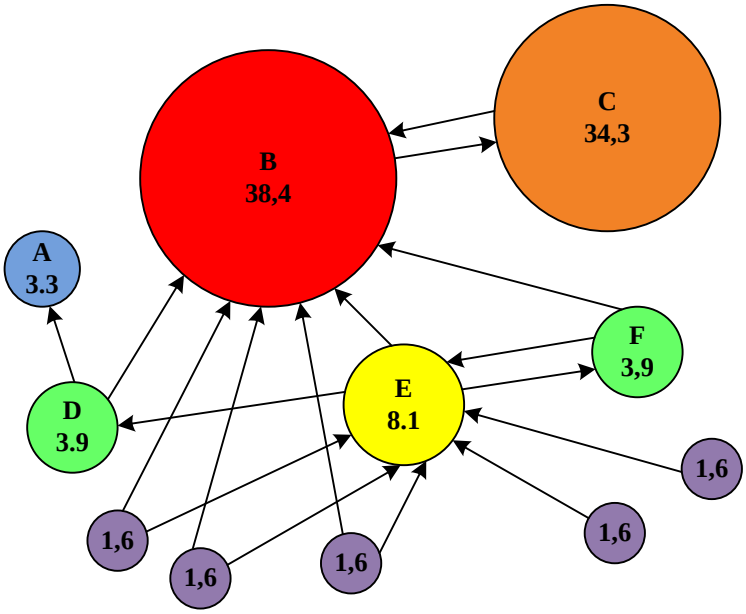

Figura 35 – Pontuação do algoritmo de *pagerank*

Outra estratégia utilizada também é chamada de *crawling* de caminho ascendente. É utilizada quando *crawlers* pretendem capturar todos os recursos disponíveis de um *web site* particular. Em (COTHEY, 2004) é introduzido o *crawler* de caminho ascendente que deve ascender de qualquer nível da *URL* que pretende capturar. Por exemplo, quando dado um *URL* semente "http://www.ele.puc-rio.br/POS/2008/index.html", ele tentará capturar "/POS/2008/", "/POS/", e "/." Constatou-se que o *crawler* de caminho ascendente foi muito efetivo em encontrar recursos isolados (COTHEY, 2004).

(RICOTTA, 2007) relata que muitos *crawlers* de caminho ascendente são conhecidos como *softwares* de colheita, pois eles usualmente "colhem" ou colecionam todo o conteúdo - talvez uma coleção de fotos em uma galeria - de uma página específica ou host.

Porém, quando o objetivo é coletar dados específicos, a estratégia mais utilizada é a de *crawler focado: web crawlers* que se propõem a baixar páginas similares ou relevantes a uma determinada página. Este tipo de estratégia será abordada com detalhes no capítulo "5".

4.4.5. Máquinas de Busca

Máquinas de Busca ou *search engines* foram o primeiro tipo de ferramenta usada para consultar a *Web* (BAEZA-YATES & BERTIER, 1999). Um *search engine* é um *web site* especializado em buscar e listar páginas da *Web* a partir de palavras-chave indicadas pelo utilizador. As máquinas de busca, inicialmente, eram baseadas na busca simples por palavras ou sentenças presentes nos documentos *web*. Posteriormente, foi incluída a exploração da estrutura de *links* para aumentar a qualidade das respostas.

Surgiram logo após o surgimento da Internet, com a intenção de prestar um serviço importante: a busca de qualquer informação na *Web*, apresentando os resultados de uma forma organizada, e também com a proposta de realizar a tarefa de uma maneira rápida e eficaz. Atualmente, diversas empresas disputam a liderança deste segmento de serviço. Entre as maiores e mais conhecidas empresas

encontram-se o Google, o Yahoo, o MSN, o Baidu, e mais recentemente a Amazon.com com o seu mecanismo de busca A9.

Embora cada máquina de busca utilize suas próprias técnicas, o processo de consulta é realizado como demonstrado na Figura 36. O usuário submete ao processador de consultas a necessidade de informação por meio de palavras-chave. Então, o processador de consultas irá a sua base de dados realizar o casamento de termos e, baseado em algumas heurísticas de *rank*, retornará o resultado ao usuário.

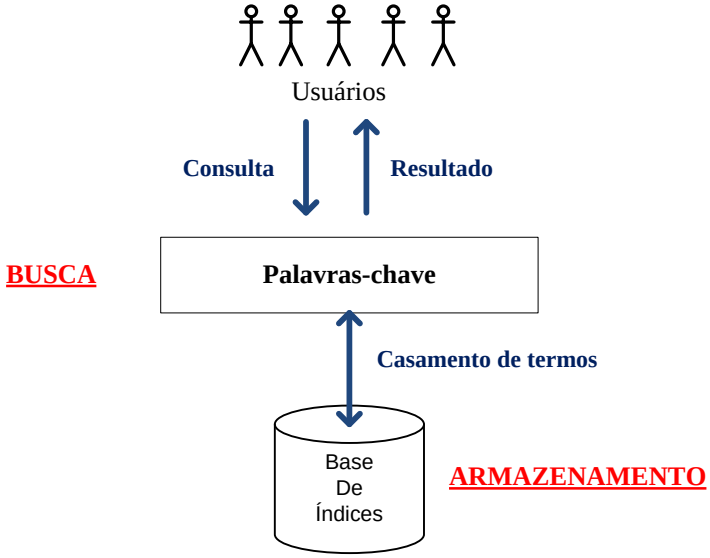

Figura 36 - Processo de consulta em uma máquina de busca

A recuperação de informação no complexo ambiente da *Web* é relativamente facilitada pelos *search engines* que coletam e indexam uma parte da imensa quantidade de páginas disponíveis neste ambiente. Para facilitar a seleção dos itens recuperados, a maioria dos mecanismos de busca realiza um ordenamento dos resultados, utilizando algum algoritmo que tenta predizer a relevância de cada item para a necessidade de informação do usuário. As primeiras referências são presumivelmente mais relevantes do que as últimas.

Como cada máquina de busca utiliza seu próprio algoritmo para a coleta e indexação de páginas, para uma mesma expressão de busca, os resultados apresentados pelos diferentes mecanismos podem variar consideravelmente. Pode-se supor, então, que a combinação de vários mecanismos de busca pode aumentar a área de cobertura da Web e, conseqüentemente, permitir obter resultados mais completos do que um mecanismo de busca tomado isoladamente.

De acordo com (BRIN & PAGE, 1998) e, mais recentemente, com (WEN, 2006), a arquitetura comum de uma máquina de busca moderna é ilustrada na Figura 37. Todo o processo começa com *crawlers* que, baseados em estatísticas como a freqüência de modificação de um *site*, iniciam o trabalho de rastreamento na *Web*. Baseado no que foi coletado, *crawlers* atualizam as estatísticas dos *sites*, enviam as páginas recuperadas para o *Parser* e mantém, temporariamente, uma cópia em cachê de alguns *sites*. O *parser* é responsável por interpretar a linguagem de codificação do documento e remover *tags* de marcação visual e estrutural. Uma vez extraído o conteúdo informativo de um *site*, este é enviado ao construtor de índices para que possa ser atualizado ou incluído na base de dados (índices invertidos). *Links* identificados pelo *parser* são enviados ao *crawler* para que possam ser recuperados em algum momento (incluídos em uma pilha) e para o construtor do *Web Graph*. A maior inovação em relação às máquinas de buscas de primeira geração é a análise do *Web Graph*. Algoritmos como o *pagerank* são utilizados para identificar o nível de importância de um *site* na *Web* baseado na quantidade e importância de apontamentos que este recebe.

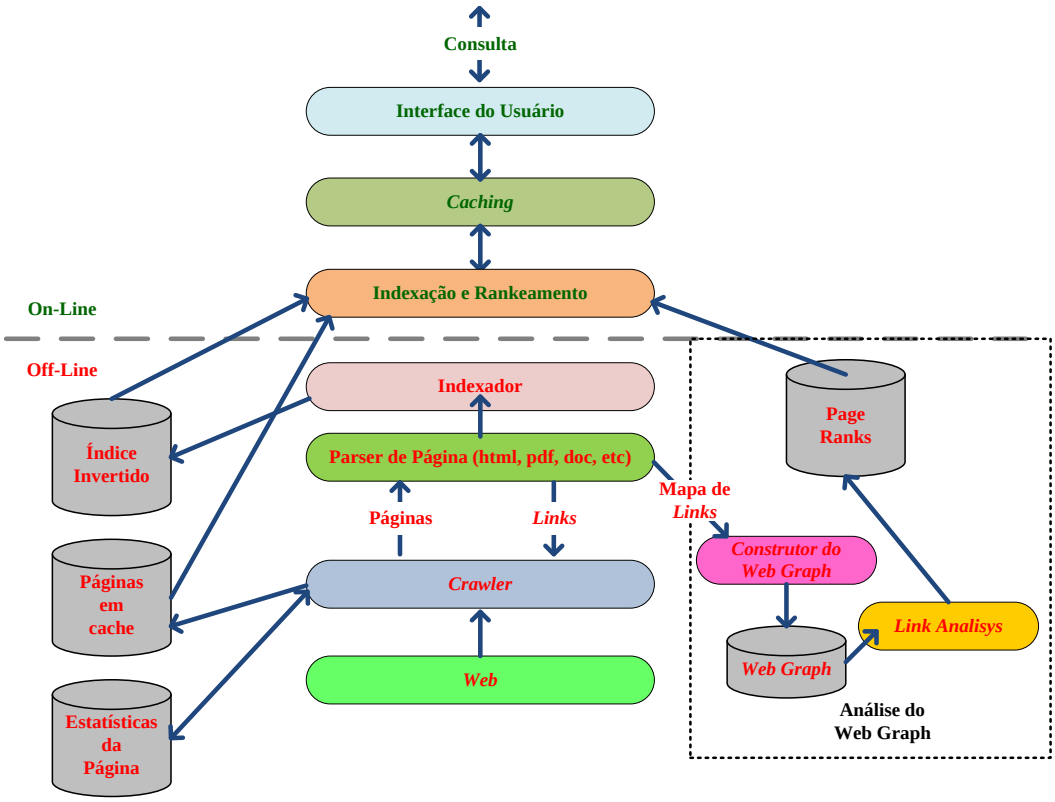

Figura 37 – Arquitetura de uma máquina de busca moderna

Na parte *on-line*, máquinas de busca mantêm em *cache* uma lista das consultas mais solicitadas. Quando uma consulta não armazenada em *cache* é solicitada, realiza-se uma pesquisa na base de dados destas máquinas de busca, geralmente estruturada sob a forma de índices invertidos e recupera-se a lista de *sites* que atendem àquela necessidade de informação. Em seguida, informações sobre a análise do *web graph* são utilizadas para ordenar os *sites* que atendem à necessidade de informação do usuário pelo seu nível de importância.