

A BIBLIOTECA DE BABEL

Introdução

Jorge Luis Borges, em seu conto "A Biblioteca de Babel", descreve um espaço tão grande que não pode ser percorrido e onde toda informação do mundo (e a desinformação também) estão à disposição de todos, inclusive a história passada e futura de suas vidas e todas as suas diferentes vidas possíveis ou não. A coleção desta biblioteca é tão vasta e avassaladora que encontrar algo de valor nela é quase impossível.

Algumas pessoas tem comparado a Internet com esta biblioteca de Borges e com razão. Existem atualmente milhões e milhões de documentos *online*, um grande número deles de acesso gratuito, e milhares são acrescentados a cada momento. Recuperar aqueles que nos interessam entre tudo o que existe nesta dinâmica e vasta coleção, reconhecer o que é relevante e o que é bobagem, o que é verdadeiro e o que é falso, enfim é, para muitos, quase como achar uma agulha no palheiro: produto do acaso ou da sorte.

Diferente de uma biblioteca, entretanto, a esmagadora maioria dos documentos colocados na Internet não obedecem a nenhuma classificação sistemática, como o código de Dewey, onde as obras são catalogadas por assunto, num determinado código. A disponibilização de documentos ainda é feita, em geral, por pessoas que não são profissionais da área de informação e documentação. Por esta razão, encontrar informação e garantir sua validade não é fácil e agora muitos profissionais da área de T.I. se debruçam sobre a necessidade de organização da *web*.

Existem iniciativas, como o "Dublin Core" e uso de *meta-tags* (como veremos mais tarde), que representam uma primeira tentativa de organização do caos mas, atualmente (2003), a pesquisa na Internet pode ser feita apenas por três recursos: a navegação aleatória, uma dica de alguém, ou através do uso dos *sites* de catalogação e indexação. Este texto é sobre esta última opção, sobre como encontrar informação pública na *web*, de forma inteligente e independente.

BUSCANDO INFORMAÇÃO

Sumário VERSUS Indexes ou “Índice Sinóptico” VERSUS “Índice Remissivo”

Semelhante aos livros, existem 2 estratégias de buscar informação de modo sistemático:

1. Os *sites* catalogadores, que corresponderiam ao sumário (também chamado comumente de índice) de um livro (o grande livro da Internet, que organizam os *sites* em tópicos (capítulos) hierárquicos e suas diferentes sub-seções). Ex. Yahoo e Cadê?, onde apenas os *sites* manualmente submetidos e cadastrados é que são incluídos nos diretórios, por assunto.
2. Ou os *sites* indexadores, *search-engines*, cujos robots ou *crawlers* percorrem incessantemente todos os *sites* e atualizam a meta informação a respeito do conteúdo dos mesmos, para um *search-engine* específico, que salva os dados em ordem alfabética, exatamente como num índice remissivo de livro. A diferença é que, no livro, como a lista é pequena, podemos vê-la toda e, ao acharmos o tópico desejado, basta abrir a página indicada junto ao mesmo. As listas dos *search-engines*, por outro lado, são acessadas através de formulários de buscas por palavra(s) que indicam a informação desejada, criando um índice remissivo personalizado. A lista das páginas onde reside a informação é apresentada (menor ou maior dependendo do número de *hits* bem sucedidos que a busca ocasionou) e o acesso às mesmas dá-se através dos *links*. Diferente dos *sites* catalogadores, os *search-engines* lançam seus robots regularmente, indexando os *sites* independentemente da ação humana. Exs: altavista e google.

Quando acessar um ou outro?

A resposta era mais uma questão de bom-senso. No início, páginas importantes eram indexadas manualmente (i.e. valiam o esforço de descrevê-las e detalhá-las para o formulário de submissão do *site* catalogador). Para uma “varredura” completa, *os sites* indexadores levantavam todas as possibilidades existentes mas, com frequência, apresentavam um volume de resultados impossível de manejar.

Hoje em dia confiamos nos indexadores como o Google e assemelhados, pois quase ninguém mais indexa manualmente no Yahoo e outros catalogadores. A solução? Refinar as estratégias de busca usando os operadores lógicos e consultar mais de um *search-engines*, uma vez que estes possuem periodicidade e estratégias de indexação diferentes, como veremos adiante.

A LÓGICA DA BUSCA

Operadores lógicos

A lógica booleana (inventada por Boole), consiste em construir afirmações lógicas utilizando os chamados “operadores lógicos”, que aproximam ou separam os elementos, possibilitando refinar a busca e recuperar dados pertinentes.

1. **AND** Como mostra o diagrama abaixo, a palavra AND engloba a área comum dos dois universos, isto é, aquela que contém ambos os termos. Qualquer documento que contenha apenas um dos termos, é excluído. Muitos *search-engines* aceitam o sinal de + para fornecer o mesmo resultado (+cães +gatos).



2. **OR** Como mostra o diagrama abaixo, o operador lógico OR engloba ambos os universos, não apenas as áreas comuns. Isto é, serão igualmente recuperados todos os documentos que contenham apenas a palavra “cães”, apenas a palavra “gatos”, OU ambas as palavras.



3. **NOT** Como mostra o diagrama abaixo, o operador lógico NOT exclue todos os “gatos” do universo “cães” e também os “cães” que aparecem no universo “cães e gatos”. Muitos *search-engines* aceitam o sinal de - para fornecer o mesmo resultado (+cães -gatos). Não esqueça de colocar qualquer um destes sinais sem espaço algum com a palavra que vem depois.



Operadores de texto ou “ de proximidade”

1. **NEAR** Alguns *search-engines* e bases de dados, neste momento, utilizam este operador de proximidade. Isto quer dizer que ambos os termos definidos devem estar próximos um do outro, geralmente na mesma frase. Isto permite refinar a busca de forma muito eficiente. Diferente do AND, que requer apenas a presença de ambos os termos em qualquer lugar do documento, o NEAR cria um *link* conceitual entre ambos.

2. **FOLLOWED BY** - Poucos *search-engines*, como OpenText, oferecem este operador que liga dois termos ou frases de modo que um preceda o outro na ordem determinada. Na maioria dos *search-engines* isto equivale a usar ambos os termos entre aspas. Ex. Porto FOLLOWED BY Alegre. Ou “Porto Alegre”. É possível colocar diversas palavras entre aspas ou mesmo uma frase inteira: “Festival de Cinema de Gramado”. Outros usam o “_”, como em: Porto_Alegre.

2. **ADJ (adjacent)** - Utilizados para termos juntos e na mesma ordem. Dog ADJ cat produz resultados diferentes de cat ADJ dog.

4. Os bons *search-engines* permitem também que se use *wildcards** ou palavras truncadas, para abranger o maior número de variações em torno de um radical. Por exemplo, caminh* abrangerá resultados com “caminho”, “caminhada”, “caminhão”, etc. Em geral são buscados até 3 caracteres após o asterisco.

SEARCH ENGINES

Altavista (www.altavista.com)

Desenvolvido pela Digital Corp. é um dos mais poderosos e flexíveis *search-engines* globais, atualmente. Os indexes são atualizados diariamente, a frequência e a proximidade de palavras significativas são registradas, e formam a base da ordem do *display* do resultado da busca. Em 97 já possuía 31 milhões de registros para *webpages*, de 620.000 servidores por todo o mundo. Indexa também os 4 milhões de artigos postados pelos grupos da Usenet diariamente. Seu *site* é acessado 30 milhões de vezes a cada dia. Seu *search-engine* permite a utilização de operadores lógicos e de proximidade, bem como o uso de termos truncados.

Excite! (www.excite!.com)

Desenvolvido pela Excite Inc., usa um *web crawler* e oferece *reviews* de *sites* em uma grande variedade de categorias. Excite! se autoproclama como sendo a melhor ferramenta de busca na *www* com mais de 50 milhões de *sites* indexados desde 1997. Este *search-engine* utiliza uma tecnologia de Inteligência Artificial "ICE" (*Intelligent Concept Extraction*) para estabelecer relações entre os termos das páginas indexadas. O *search-engine* lida também com frases coloquiais e utiliza *fuzzy logic* para encontrar resultados relevantes. Por isto o Excite! é muito útil para os novos usuários utilizarem, porque procura compensar as buscas mal-formuladas e monta listas por relevância.

HotBot! (www.hotbot.com)

Desenvolvido por Inktomi Corp., o hotbot é o *search-engine* da revista WIRED e "afirma" possuir o maior e mais completo index de documentos da *www*, utilizando elementos de inteligência artificial para recuperar informações através de uma grande variedade de opções, acessíveis através do *search control panel*.

Infoseek (www.infoseek.com)

Desenvolvido pela Infoseek Corp., este *search-engine* foi criado em 1994 e, naturalmente, proclama possuir o mais vasto diretório de *sites* organizado. Sua atuação o coloca em destaque nas lista de performance de indexadores, anualmente. Usuários de Windows podem adicionar a capacidade de busca do Infoseek na barra do menu, fazendo o *download* do *software* sugerido no *site*.

Google (www.google.com)

Um dos *search engines* mais eficientes e populares no momento e com uma interface despojada e elegante, o Google, como o Altavista, oferece também a capacidade de busca na barra do menu do browser. Oferece também a opção de "pesquisa de Imagens", que é de fácil compreensão, com mais de 390 milhões de imagens indexadas e disponíveis para visualização. O Google também tem uma opção de tradução automática, mas isso nem sempre funciona adequadamente.

O coração do sistema de *ranking* é o software PageRank(TM), um sistema para dar notas para páginas na *web*, desenvolvido na Universidade de Stanford. A classificação das páginas (*PageRank*) confia na natureza excepcionalmente democrática da Web, usando sua vasta estrutura de *links* como um indicador do valor de uma página individual. Essencialmente, o Google interpreta um *link* da página A para a página B como um voto da página A para a página B. Mas o Google olha além do volume de votos, ou *links*, que uma página recebe; analisa também a página que dá o voto. Os votos dados por páginas "importantes" pesam mais e ajudam a tornar outras páginas "importantes."

Sites importantes, de alta qualidade, recebem uma nota de avaliação maior, que o Google grava a cada busca feita. Naturalmente, uma página importante não significa nada se não combinar com a sua busca. Assim, o Google combina os resultados de alta qualidade com o *search* que está se realizando para que o resultado seja o mais relevante possível. O Google pesquisa quantas vezes a palavra procurada aparece nas páginas e examina todo o aspecto delas (e conteúdo das páginas ligadas a ela) para determinar o melhor resultado para a busca realizada.

Vivisimo (Vivisimo.com) & Northernlight (www.northernlight.com)

Dois dos *search-engines* de última geração, acrescentam inovações importantes no gerenciamento de conteúdos, partindo do princípio que não basta encontrar uma longa lista de informação em potencial, mas é também necessário criar *sub-rankings* com as regularidades encontradas, organizando a informação de modo significativo. Estes *search-engines*:

1. Fazem uma segunda pesquisa no universo encontrado e criam sub-grupos com as regularidades, dividindo a busca em diretórios com sub-tópicos, subgrupos "inteligentes" dentro de cada *search* realizada.
 2. Apresentam apenas a URL mais relevante de cada *site*, colocando as demais num diretório separado.
 2. O Vivisimo ainda oferece abrir cada *site* numa nova janela, ou *frame*, etc..
- Uma opção ótima para espiar em que consiste cada uma delas e tomar a decisão de explorá-las ou não.

TodoBr (www.todobr.com.br)

Indexador brasileiro com várias opções interessantes na busca avançada. Por exemplo, localiza documentos por estado (UF) onde se encontram.

META-INDEXADORES

Dogpile (www.dogpile.com)

Diferentes indexadores encontram resultados diferentes, em virtude de uma série de variáveis, como a periodicidade dos *updates* da lista, critérios de *ranking**, alcance, etc. . Os meta-indexadores, como o Dogpile, buscam em todos os demais, que sejam de acesso público também, como as anteriormente citadas. Eles não são os proprietários da(s) base(s) de dados acessada(s) apenas buscam, organizando os resultados e removendo os redundantes.

Outros exemplos de meta-indexadores:

<http://www.MetaCrawler.com/> OU <http://www.surfwax.com/>

Uma lista bem extensa e a descrição de cada um pode ser encontrada em:

<http://mayura.sjp.ac.lk/faq/meta.htm>

* -> Critérios de *ranking* (e sua manipulação) é assunto de nossa próxima aula.