Como os Mecanismos de Busca da Web Indexam Páginas HTML

Fernando Campos
campos@dcc.ufmg.br

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Marcello Peixoto Bax
bax@eb.ufmg.br
www.bax.com.br

Escola de Ciência da Informação
Universidade Federal de Minas Gerais

14 de Abril de 2000

1. Introdução

As máquinas de busca são hoje um dos métodos mais utilizados para recuperação de informação na Internet. Várias pesquisas são feitas nesta área, procurando incrementar os processos automáticos de indexação e classificação da informação utilizados, a fim de melhorar a relevância e a velocidade da recuperação de informações na *Web*. [1] O grande volume de informação disponível nos índices das máquinas de busca, hoje em dia, contribui para a geração de respostas muito extensas e abundantes, mas muitas vezes de baixa qualidade. [2]

Nota-se que estes processos estão longe da perfeição e dependem, em grande parte, do preparo prévio dos documentos a serem indexados, uma tarefa para especialistas. Se a "página" (documento HTML referenciado por uma URL[1]) a ser indexada contiver suas informações organizadas de maneira que as máquinas possam acessá-las e "compreendê-las", os processos de indexação e classificação são então potencializados, resultando em índices de melhor qualidade.

Para se produzir uma página com estas características é necessário ter conhecimento de como as máquinas de busca realizam esta indexação, ou seja, quais fatores são considerados por elas no momento em que analisam uma página. [3]

Os termos "mecanismos de busca" e "máquinas de busca" muitas vezes são utilizados de maneira confusa, ou como sinônimos. Neste documento consideramos o segundo como um termo genérico que contempla tanto os mecanismos de busca quanto os diretórios. Ambos os termos se referem às ferramentas utilizadas para recuperação de informação na Internet, mas que funcionam de maneira diferente. Mecanismos de busca, como *AltaVista* [4], criam seus índices automaticamente. Elas

1 URL – "Uniforme Resource Locator", modo de endereçamento formal de um recurso na Web.

percorrem continuamente a *Web* visitando *sites* e indexando suas páginas. As pesquisas são feitas utilizando-se estas informações colhidas. Já os diretórios, como *Yahoo*[5], constroem seus índices através de descrições de páginas fornecidas pelas pessoas no momento de submete-las. As pesquisas se baseiam nas informações fornecidas, e não no conteúdo real das páginas. Assim, a indexação dos mecanismos de busca é feita de maneira diferente daquela dos diretórios e os fatores importantes para a classificação das páginas também diferem.

Quando se procura por alguma coisa na Internet usando uma máquina de busca, ela instantaneamente realiza uma pesquisa nos milhões de páginas que ela conhece (aquelas já indexadas) e apresenta como resultado aquelas que têm alguma relação com o tópico. Geralmente este resultado é apresentado por ordem de relevância, e freqüentemente constitui-se de uma lista de centenas, às vezes milhares, de URLs. Na maioria dos casos, apenas os 10 resultados mais relevantes são exibidos na primeira página do resultado. Denominar-se-á tal lista de *top ten*.

Qualquer pessoa ou organização que possua uma página na *Web* (ou um *website*) deseja estar na lista *top ten*. Isto porque a maioria dos usuários dos mecanismos encontra o resultado desejado na primeira página de resposta da pesquisa (ou no máximo na segunda) e não visita as páginas subseqüentes. Para um *website* estar listado na décima primeira posição, pode significar deixar de receber a visita de muitos usuários. [6]

Neste artigo procura-se apresentar os principais fatores analisados pelos mecanismos de busca no momento da indexação de uma página e que devem, portanto, ser considerados no momento de se produzir documentos ou páginas HTML. O documento está estruturado da seguinte forma, na Seção 1 é apresentada uma contextualização do problema da indexação de páginas HTML feita pelos mecanismos de busca. A seção 2 introduz o conceito de palavra-chave e explica sua importância dentro do processo de indexação de documentos. A seção 3 explicita como deve ser feita a estruturação de um documento HTML de maneira que este possa ser melhor indexado por uma máquina de busca. A seção 4 apresenta as considerações finais deste texto e explica o propósito maior de sua produção. A seção 5 contém as referências bibliográficas utilizadas.

O texto assume bons conhecimentos de HTML por parte do leitor.

2. Palavras-chave estratégicas

Quando se pesquisa em uma máquina de busca por páginas de um determinado assunto, utiliza-se palavras-chave relacionadas com este assunto. Imagina-se que as palavras usadas são as mais indicadas para descrever o conteúdo procurado. Palavras-chave são palavras que possuem a capacidade de descrever semanticamente o conteúdo de uma página, sob o ponto de vista dos usuários dos mecanismos de busca.

Quais termos as pessoas pensariam em utilizar para pesquisar por uma página de determinado assunto em uma máquina de busca? Para quais termos pesquisados o detentor de uma página gostaria de ver uma referência a ela no *top ten*? Estes termos são as palavras-chave estratégicas de uma página.

Por exemplo, digamos que alguém possua uma página dedicada ao assunto "cultura africana". Sempre que alguém digitar "cultura" e "África" para realizar uma pesquisa, o detentor da página certamente gostaria que ela estivesse no *top ten*. Então estas são exemplos de palavras-chave estratégicas para esta página.

Cada uma das principais páginas de um *website* deve ter palavras-chave estratégicas próprias, que reflitam o conteúdo da página. Por exemplo, digamos que o dono da página sobre "cultura africana" também possua uma página sobre "culinária africana". "Culinária" e "África", por exemplo, são palavras-chave para esta outra página.

As palavras-chave estratégicas podem ser combinadas para formar frases-chave, a fim de aumentar a especificidade semântica do termo. Normalmente o tamanho do resultado de uma pesquisa pela palavra "cultura" é muito maior do que o resultado de uma pesquisa por "cultura africana", porque existem mais páginas sobre o tema genérico "cultura" do que sobre o tema mais específico "cultura africana". Devido a estas diferenças de tamanho e especificidade dos resultados, a qualidade da resposta no primeiro caso (pesquisa por "cultura") é menor do que no segundo (pesquisa por "cultura africana"), do ponto de vista de quem procura por uma página sobre "cultura africana".

3. Posicionamento das palavras-chave

Os mecanismos de busca são bem menos eficientes que os humanos para analisar semanticamente uma página HTML e determinar qual é o assunto tratado assim como sua relevância. [9]

O que os mecanismos fazem para inferir estas informações é procurar pela ocorrência do(s) termo(s) pesquisado(s) em posições estratégicas na página. Baseando-se no número de ocorrências do(s) termo(s) e na sua localização, determina-se o grau de relevância da página para este(s) termo(s).

Ao se construir um documento HTML, deve-se procurar posicionar as palavras-chave estratégicas nestas posições cruciais, para garantir que os mecanismos conseguirão "entender" o significado da página. Estas posições serão apresentadas e analisadas nas sessões seguintes.

3.1 Título

Imagine que alguém numa biblioteca solicite livros sobre o Japão. Caso o bibliotecário não possua conhecimento sobre quais são os (bons) livros sobre o assunto, faz

sentido supor que ele iniciará a busca procurando por todos os livros que contenham a palavra “Japão” no título.

Mecanismos de busca operam de maneira similar. Eles consideram o título da página (texto entre os *tags* <title>...</title>) o local mais importante para a determinação do assunto da página. Páginas com palavras-chave no título são supostas serem mais relevantes, dentro de um mesmo tópico, do que outras páginas que não as possuem naquela posição. Algumas páginas de ótimo conteúdo sobre um determinado assunto podem obter classificações ruins principalmente por falharem neste ponto. Portanto, o título deve conter as palavras-chave estratégicas da página.

Não existe um valor padrão para o número máximo de caracteres aceito pelos mecanismos de busca para este campo. Cada mecanismo possui seu próprio valor. No entanto, pode-se considerar um valor médio de 1000 caracteres.[4] Não significa que o título será ignorado caso o valor máximo seja excedido. Este número indica apenas o tamanho da porção do título que será indexada. É recomendável ocupar-se as primeiras posições do título com as palavras-chave estratégicas, para evitar que sejam ignoradas pelos mecanismos no momento da indexação.

3.2 Topo da página

Os mecanismos de busca consideram o cabeçalho da página (textos entre os *tags* <head>..</head>, <h1>..</h1>, <h2>..</h2> ou ..) e as primeiras linhas de texto, uma região importante para a classificação. Consideram que uma página relevante para um determinado assunto conterá palavras relacionadas com este assunto desde o seu começo.

Pode-se pensar que a máquina de busca vê o texto da página como alguém que lê um artigo em um jornal. O primeiro parágrafo de um artigo sempre diz (ou deveria dizer) quais são os seus pontos principais. A máquina de busca “lê” então esta porção de texto para tentar entender sobre qual assunto se refere a página. Portanto, as palavras-chave estratégicas devem aparecer tanto no cabeçalho da página, quanto no primeiro parágrafo de texto.

3.2.1 Tabelas

A utilização de tabelas para construir uma página pode “empurrar” o texto inicial para posições mais baixas da página, tornando as palavras-chave de seu conteúdo menos relevantes. Isto ocorre porque os mecanismos (como as versões antigas de *browsers*) lêem as tabelas de maneira fragmentada, por coluna. Por exemplo, vamos supor uma página de duas colunas típica, onde a primeira coluna possui os *links* de navegação e a segunda possui o texto contendo as palavras-chave:

home página1 página2 página3	**Cultura Africana** Cultura Africana....

página4

Mecanismos de busca enxergariam esta página assim:

home
página1
página2
página3
página4
Cultura Africana

Cultura Africana

Pode-se observar como as palavras-chave foram movidas para a parte mais baixa da página. Isto provavelmente irá prejudicar a classificação desta página numa busca por “cultura africana”, pois os mecanismos de busca considerarão como primeiro parágrafo os links e não o texto que contém as palavras-chave.

Infelizmente, não existe uma forma definitiva de contornar este problema. Pode existir um compromisso entre a boa diagramação (*design*) da página e sua boa colocação nos índices dos mecanismos. Caso a não utilização de tabelas prejudique o *design* da página, é recomendável continuar utilizando-as. Não significa que a colocação da página dentro dos resultados de pesquisa estará completamente comprometida com o seu uso. Existem maneiras de minimizar o problema, as quais serão abordadas mais à frente.

3.2.2 JavaScript

Grandes porções de código JavaScript no topo da página também podem causar o mesmo efeito ruim das tabelas. Os mecanismos podem indexar o código primeiro, atribuindo-lhe maior relevância do que o texto dos primeiros parágrafos que contêm termos importantes. Por exemplo, foi feita uma pesquisa com o termo “+document +write” (um típico pedaço de código JavaScript é “document.write”) no AltaVista, com o objetivo de encontrar alguma página que tivesse sido indexada com este texto (“doument.write”). A língua do resultado da pesquisa foi definida para português, para diminuir a probabilidade de encontrar páginas que contivessem as palavras da língua inglesa *document* e *write* em seu conteúdo, e não na porção de código JavaScript. Foi obtido o seguinte resultado:

1. Obrasil.com
O seu portal brasileiro nacional. E-mail grátis, páginas grátis, e muito mais....
URL: www.obrasil.com/
Last modified on: 13-Mar-2000 - 32K bytes - in Portuguese (Win-1252)
[Translate] [More pages from this site] [Related pages]

O seguinte trecho está localizado no topo do documento fonte desta página:

```
<SCRIPT LANGUAGE="JavaScript">
<!--
...
document.write('<a onclick="submit();">');
...
-->
```

Não há nenhuma ocorrência das palavras *document* e *write* na página fora da porção de código JavaScript. Pode-se concluir que o AltaVista indexou alguma porção de código desta página.

Um bom artifício para evitar que o código JavaScript seja indexado é colocá-lo dentro de um *tag* comentário. Esta é a prática padrão para impedir que *browsers* que não entendam JavaScript o vejam. Porém esta tática pode falhar, caso o símbolo “>” seja utilizado no código. O que se supõe é que nada entre “<!--“ e “-->” será indexado por uma máquina de busca que ignore comentários. No entanto, o caracter “>” pode ser interpretado como o fechamento do *tag* comentário (isto pode ser comprovado com o uso de um *browser* antigo). Então, tudo a partir deste ponto será tratado como texto HTML. Outro problema é que alguns mecanismos indexam texto entre *tags* comentário.

Uma alternativa complementar ao uso de comentários é mover o JavaScript o máximo possível para o final da página. Isto pelo menos dará a certeza de que o código não será o primeiro texto encontrado.

Uma solução mais apropriada é incluir arquivos externos (“.js”) que contêm o código JavaScript fazendo referência a eles na página HTML. Dessa forma somente *browsers* capazes de entender JavaScript carregam o código. Os mecanismos não irão importar esta informação.

3.3 Freqüência

A freqüência[2] de ocorrência de um termo em uma página é outro fator importante na determinação de relevância de uma página para este termo. Os mecanismos calculam a freqüência das palavras pesquisadas e analisam sua relevância considerando este valor.

Os mecanismos mantêm em seu índice o número de vezes que cada palavra foi indexada. Desta forma, elas conseguem determinar quão rara é uma palavra. Palavras pouco raras costumam ser chamadas de *stop words* e não são consideradas para

2 Frequência = número de aparições da palavra no texto / número total de palavras no texto.

efeito de cálculo de relevância. Palavras raras recebem um peso maior no cálculo da relevância.[10]

Estes dois valores, freqüência da palavra na página e sua freqüência no índice são combinados para determinar a relevância da página para esta palavra.

As palavras-chave estratégicas devem portanto estar presentes no conteúdo da página. Isto significa que o texto escrito em uma página é necessário. Páginas constituídas somente por figuras não são entendidas pelos mecanismos de busca, simplesmente porque elas não conseguem ler figuras. Alguns (poucos) mecanismos indexam textos contidos no tag <ALT> e em comentários, que são um bom artifício para inserir o texto escrito correspondente a uma figura. No entanto, para se ter segurança é melhor utilizar diretamente o texto HTML sempre que possível.

3.3.1 Spam

O texto de uma página deve ser sempre visível em um *browser*. Algumas pessoas tentam enganar os mecanismos repetindo várias vezes as palavras-chave na página e formatando-as com um tamanho de fonte muito pequeno ou com uma cor idêntica à cor do fundo, tornando estas palavras invisíveis em um *browser*. Os mecanismos são capazes de identificar esta prática (*spam*) e, além de não indexar este tipo de texto, penalizam a página com pontos negativos no momento do cálculo da relevância.

Outra prática que pode ser considerada *spam* é a repetição indiscriminada de palavras-chave no texto. Os mecanismos são capazes de ler uma frase (seqüência de palavras) e identificar se esta seqüência representa uma frase válida ou se é apenas um conjunto desconexo de palavras.

3.4 Meta Tags

O HTML permite que sejam especificados meta dados em um documento, ou seja, informações sobre o documento além do seu conteúdo, através do tag <META>. Este tag pode ser utilizado para incluir pares nome/valor que descrevem propriedades do documento, como por exemplo o autor, uma lista de palavras-chave, etc. Estas informações são invisíveis em um *browser*, e portanto ao visitante da página.

Muitos mecanismos consideram estas informações tanto no momento em que indexam uma determinada página, quanto quando calculam sua relevância para efeito de classificação. Os principais meta tags considerados são **description** e **keywords**.[7][8]

3.4.1 O Meta Description

O meta tag description é utilizado para incluir uma descrição do conteúdo da página dentro do documento HTML.

Esta descrição é utilizada pelos mecanismos de busca que suportam meta *tags* no momento em que eles apresentam a página como resultado de uma pesquisa. Por

exemplo, fazendo uma pesquisa no AltaVista pelo termo “africa” foi encontrado como resultado:

1. Orientation Africa
Gateway to Africa on the Web: art, literature, business, trade, culture, politics, music, news, events, sports, travel, food and discussions...
URL: af.orientation.com/
Last modified on: 4-Mar-2000 - 4K bytes - in English
[Translate] [More pages from this site] [Related pages]

O código fonte desta página, contém o seguinte tag meta description:

```
<META NAME="Description" CONTENT="Gateway to Africa on the Web: art, literature, business, trade, culture, politics, music, news, events, sports, travel, food and discussions">
```

A descrição apresentada pelo mecanismo de busca para a página e o conteúdo do meta tag são idênticos.

Caso a página não possua o meta tag description, os mecanismos utilizam as primeiras linhas de texto como uma descrição da página. Isto pode ser um problema para páginas que não possuem nenhum texto, como as constituídas somente por figuras ou as que possuem somente a definição de um *frameset*. Estas páginas não terão descrição quando aparecerem em algum resultado de pesquisa. Páginas que possuem os problemas referentes ao uso de tabelas ou JavaScript descritos anteriormente terão uma descrição confusa, constituída por estes textos que aparecem no topo da página.

O meta description também é considerado no momento do cálculo da relevância de uma página para um determinado termo. Este tag pode ser visto como uma porção adicional de texto na página que é invisível, ou seja, não aparece nos navegadores.

3.4.2 O Meta Keywords

O meta tag keywords é utilizado para especificar as palavras-chave que descrevem o conteúdo da página ou do *site*. Este meta tag é indexado pelos mecanismos de busca e é utilizado no cálculo da relevância da página.

O meta keywords deve ser explorado para incluir todas as palavras-chave que tenham alguma relação com a página, mesmo as que não aparecem no seu corpo. O meta pode ser utilizado para incluir sinônimos de palavras-chave, plurais irregulares (por exemplo, “person” e “people”), palavras-chave menos importantes, combinações de palavras para formar frases, etc.

A utilização para a inclusão de frases é particularmente importante. Considerando como exemplo a página “Orientation Africa**”** utilizada no tópico anterior

(meta tag description), vamos supor dois possíveis meta tags keywords para esta página que possuem as mesmas palavras em seu conteúdo.

```
<META NAME="Keywords" CONTENT="AFRICAN DIRECTORY, TRAVEL AFRICA, AFRICAN BUSINESS, AFRICAN CUISINE">

<META NAME="Keywords" CONTENT="AFRICAN, DIRECTORY, TRAVEL, AFRICA, BUSINESS, CUISINE">
```

Se alguém procura por “african business”, os mecanismos atribuirão maior relevância à página se ela contiver o primeiro tag. Isto ocorre porque as palavras aparecem neste tag com a mesma proximidade e na mesma ordem em que foram escritas na busca. Em outras palavras, é recomendável incluir frases que possuem uma boa probabilidade de serem usadas nas pesquisas no meta tag keywords.

3.4.3 Tamanho

Não existe um tamanho máximo padrão para os meta tags description e keywords. Este valor é diferente para cada mecanismo de busca. Normalmente, pode-se considerar um valor médio de 1000 caracteres para o meta keywords e 200 caracteres para o meta description.

Ultrapassar estes limites não significa que toda o meta tag será ignorado. Significa apenas que o mecanismo de busca não utilizará a informação excedente, de acordo com seus próprios limites.

O conteúdo dos meta tags é invisível para os usuários. Como já foi citado, páginas que não possuem texto em seu corpo não podem ter seu conteúdo analisado. O meta tag keywords pode então ser utilizado para incluir na página algumas palavras-chave que descrevem seu conteúdo. O meta description pode ser utilizado para incluir uma descrição à página. Páginas que possuem os problemas causados pelo uso de tabelas ou JavaScript podem se beneficiar destes tags para incluir texto de conteúdo relevante no topo da página.

Os meta tags são úteis e sua utilização em páginas HTML é importante, eles ajudam a aumentar a relevância de uma página. Porém, não devem ser considerados como uma solução mágica; simplesmente incluir um meta tag não garante que a página subitamente irá alcançar o *top ten*. O algoritmo de classificação e indexação de um bom mecanismo de busca considerará todos os fatores discutidos ao longo deste documento (e talvez outros não citados) em conjunto, e não isoladamente. Tornando-se assim menos vulnerável às diversas técnicas de spam.

4. Considerações finais

O presente documento é produto de um estudo sobre o comportamento dos mecanismos de busca feito para o projeto INDEXA [11], do qual os autores são

integrantes. O objetivo final do projeto é a construção de uma séria de ferramentas de software que permitam, àquelas organizações que produzem e disponibilizam informação na Web, fazerem uma análise automatizada de quão bem preparados estão seus documentos antes que estes sejam submetidos aos mecanismos de busca.

O resultado será uma aplicação que é capaz de analisar as informações de uma página HTML e propor modificações e ajustes, com o objetivo de otimizar os processos de indexação utilizados pelos cinco mais populares mecanismos de busca na *Web*.

5. Bibliografia

[1] *Survey of information Retrieval*. Available from World Wide Web: http://www.cs.jhu.edu/~weiss/ir.html

[2] *Accessibility and Distribution of Information on the Web*. Available from World Wide Web: http://www.wwwmetrics.com

[3] *Search Engine Watch*. Available from World Wide Web: http://www.searchenginewatch.com

[4] *AltaVista*. Available from World Wide Web: http://www.altavista.com

[5] *Yahoo*. Available from World Wide Web: http://www.yahoo.com

[6] *Search Engine Optimization: The Science… And The Art*. ClickZ, Nov. 19, 1998. Available from World Wide Web: http://www.searchz.com/Articles/1119981.shtml

[7] *Back to Basics: META Tags*. WebDeveloper, Nov. 1998. Available from World Wide Web: http://www.webdeveloper.com/categories/html/html_metatags.html

[8] *A Dictionary of HTML META Tags*. Available from World Wide Web: http://vancouver-webpages.com/META

[9] *Cooperative Hierarchical Indexing Coordination (TF-CHIC).* Available from World Wide Web: http://www.terena.nl/task-forces/tf-chic/

[10] Dirk van Eylen. *AltaVista ranking of query results*. Available from World Wide Web: http://ping4.ping.be/~ping0658/avrank.html

[11] *Projeto INDEXA*. Available from World Wide Web: http://www.eb.ufmg.br/bax/