Rodrigo Ferreira de Carvalho

Webdesign Goal: Uma metodologia de auxílio no desenvolvimento de sistemas inteligentes para a busca de Informações na *Web*.

Bauru
2003

10304029 - Arquitetura de sistemas de computação
- Sites da Web - Desenhos
- Sistemas operacionais distribuídos

(A) Número de chamada
Clas. T/UNESP
PHA C327w
Tombo n.º 1.181

Rodrigo Ferreira de Carvalho

Webdesign Goal: Uma metodologia de auxílio no desenvolvimento de sistemas inteligentes para a busca de Informações na *Web*.

BIBLIOTECA UNESP CAMPUS BAURU

Dissertação apresentada à Faculdade de Arquitetura Artes e Comunicação, da Universidade Estadual Paulista "Julio de Mesquita Filho", Campus de Bauru, para a obtenção do título de Mestre em Desenho Industrial.

Orientador: *Prof Dr. João Fernando Marar*
Co-orientador: *Prof Dr. Olympio José Pinheiro*

Bauru
2003

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Carvalho, Rodrigo Ferreira de

Webdesign Goal: Uma metodologia de auxílio no desenvolvimento de sistemas inteligentes para a busca de Informações na *Web*. Rodrigo Ferreira de Carvalho. Bauru 2003.

Dissertação – Mestrado – Faculdade de Arquitetura, Artes e Comunicação – Universidade Estadual Paulista.

1. *Web Design;*
2. *Search Engines;*
3. Sistemas Distribuídos;
4. Arquitetura de Informação.

Rodrigo Ferreira de Carvalho

Webdesign Goal: Uma metodologia de auxílio no desenvolvimento de sistemas inteligentes para a busca de Informações na *Web*.

COMISSÃO JULGADORA

DISSERTAÇÃO PARA A OBTENÇÃO DO GRAU DE MESTRE

Prof. Dr. João Fernando Marar

Presidente e Orientador

Prof. Dr. José Carlos Plácido da Silva

2º Examinador

Prof. Dr. Gilberto Prado

3º Examinador



Agradecimentos

Agradecimentos aos meus pais, Domingos Carvalho Netto e Eurides Ferreira de Carvalho, que por vezes, preocupados, questionavam-me se eu não iria descansar.

A Patrícia Bellin Ribeiro (namorada) que me acompanhava durante os estudos nos finais de semana.

A minha sobrinha Letícia M. Carvalho na época com três anos, que algumas vezes entrava no meu quarto para "ajudar" o desenvolvimento, querendo clicar o mouse em qualquer lugar da tela do computador...

Ao Colégio Técnico Industrial por ter reduzido minha grade de aulas, colaborando para o desenvolvimento do Mestrado.

Ao meu Orientador Prof. Dr. João Fernando Marar, pela determinação e rigor científico.

Ao meu Co-orientador Prof. Dr. Olympio José Pinheiro, pelo incentivo e paciência.

Ao Prof. Luiz Vitor Martinello, pela contribuição realizada na revisão do texto.

Resumo

O impacto da *Internet* está atingindo uma grande quantidade de usuários, e seu crescimento gera uma quantidade de informação muito grande, o que não significa que poderá ser encontrada com facilidade. Atualmente não é uma tarefa muito fácil encontrar a informação desejada na *Web*; tornando o ato da pesquisa uma tarefa árdua. Para minimizar as dificuldades em encontrar informações, algoritmos de classificação para os mecanismos de busca dos sistemas distribuídos precisam de melhores adaptações, no que tange a garantia de procura de informação correta, aplicações em Inteligência Artificial, etc. Neste sentido, o *Webdesigner* pode atuar de forma decisiva, proporcionando uma melhor resposta na classificação dos mecanismos de busca. Esse trabalho de investigação tem por objetivo descrever procedimentos que promovam a melhoraria da classificação do documento digital, e que estão ao alcance do responsável pela elaboração do *site*.

Palavras-chave: *Webdesign, Search Engines,* Sistemas Distribuídos, Arquitetura de Informação.

Abstract

The Internet impact is reaching a great number of users, consequently, it is generating a very big data information, however, with relation about this subject, there are much informations with no relevance and that these same informations, sometimes, are not so easy to find. To find an information web search engines classification algorithms need better adaptations with relation to the guarantee to find the right information, applications in artificial inteligence, etc. In this way Webdesigners can act in a decisive way, providing a better answer in the search engines classification. So this job has for objective to describe procedures that promote the form to make better the digital documents classification in which is the hands of the responsible person to create the site.

Keywords: Webdesign, Search Engines, Distributed Systems, information Architecture.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Sumário

Índice de figuras..... 191
Índice de gráficos..... 193
Índice de tabelas..... 194

Capítulo 1

Detalhamento da dissertação..... 12
1.1 Introdução..... 17
1.2 Redes de Computador e *Internet*..... 19
1.3 Sistemas Distribuídos como Suporte à Segurança de Informação..... 20
1.4 O problema: Otimizar as Possibilidades de Classificação de documentos digitais e encontrar informação segura..... 22

Capítulo 2

Características dos Sistemas Distribuídos de Informação

2.1 Introdução..... 26
2.2 Funções básicas de um sistema de busca..... 28
2.3 Classes de sistemas de busca..... 30
2.3.1 Partes de um mecanismo de busca..... 31
2.3.2 Organização dos mecanismos de busca..... 33
2.4 Diretórios..... 35
2.4.1 Riscos de não cadastramento em diretórios..... 36
2.4.2 Diferenças entre diretórios e mecanismos de busca..... 36
2.4.3 Mecanismos de Busca e Diretórios: características..... 38
Tamanho..... 38
Modos de cadastramento e de classificação..... 39
Atualização..... 40
2.5 Pesquisa Híbrida..... 41
2.6 Custos Operacionais..... 42

unesp

2.7 Funcionamento dos *Sites* de Sistemas de Busca..... 42

Capítulo 3

Visão do *Designer* sobre a qualidade e quantidade das informações disponibilizadas nos sistemas distribuídos

3.1 Introdução..... 45
3.2 Interface..... 48
3.2.1 Metáforas..... 50
3.3 Arquitetura de informação..... 51
3.4 *E-Commerce, Design* e Sistemas Distribuídos..... 54
3.4.1 Características do *E-commerce*..... 56
3.4.2 Logística do *E-commerce*..... 57
3.5 Usabilidade..... 58
3.6 Integração entre *E-commerce* e *Design*..... 61
3.7 Visibilidade para Sistemas Distribuídos..... 62

Capítulo 4

Metodologia de interfaces inteligentes para mecanismos de busca

4.1 Introdução..... 67
4.2 Características de Bases de Dados..... 71
4.3 Metáfora da Classificação..... 71
4.4 Mecanismos de Busca com Banco de Dados de Índice..... 72
4.5 Mecanismos de Busca sem Banco de Dados de Índice..... 73
4.6 Método de Classificação Profundo..... 75
4.7 Método de Classificação Rasa..... 76
4.8 Exemplos de Métodos de Classificação..... 78
4.8.1 Propriedades dos mecanismos de busca..... 78
4.8.1.1 AltaVista..... 78
4.8.1.2 *Web Crawler*..... 81
4.8.1.3 *Excite*..... 82

4.8.2 Conclusão sobre as propriedades dos mecanismos de busca.. 83
4.9 *Meta-tags*..... 84
4.9.1 Importância das *Meta-tags*..... 85
4.9.2 *Tags* para cadastramento em mecanismos de busca..... 86
4.9.3 Lista de palavras-chave..... 86
4.9.4 Descrição da página ou do *site*..... 87
4.9.5 Controle do *Spider*..... 87
4.9.6 Expiração da página..... 88
4.9.7 Revisitação do *Spider*..... 89
4.9.8 Carregamento da Página..... 89
4.9.9 Atualização da Página..... 90
4.9.10 Controle do *Cache*..... 90
4.9.11 Suporte a Idiomas..... 91
4.10 *Tag Title*..... 93
4.11 *Tag Alt*..... 93
4.12 Identificação de elementos não textuais para os mecanismos de busca..... 95
4.13 Características de uso das *meta-tags* e *tags alt*..... 98
4.14 *Spam*..... 99
4.15 Base de dados que buscam mais itens com procura única de uma palavra..... 100
4.16 Base de dados Patrocinados..... 101
4.17 Proposta de metodologia para classificação otimizada de documentos digitais nos mecanismos de busca..... 102

Capítulo 5
5.1 Processo de Criação..... 110
5.2 Desenvolvimento do Sistema *Web Goal*..... 114
5.3 Implementação e teste do Sistema *Web Goal*..... 115
5.3.1 Estrutura..... 115
5.3.2 Identificação do endereço do *site*..... 119

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

5.3.3 Utilização da *tag alt*..... 119

5.3.4 Navegabilidade..... 121

5.3.4.1 *Links* Internos..... 121

5.3.4.2 *Links* Externos..... 121

5.4 Períodos de análise..... 121

Capítulo 6

6.1 Conclusões e trabalhos futuros..... 133

Referencias Bibliográficas..... 135

Glossário..... 141

Anexos

Anexo1..... 147

Carvalho. R,F. Marar. J,F. **Arquitetura de informação para documentos digitais para *Web*** Artigo aprovado, apresentado e publicado no ISKM2003, *International Simposium on Knowledge Management* (Simpósio Internacional de Gestão do Conhecimento), PUC – Curitiba PR, 14/08/2003.

Anexo2..... 157

Carvalho. R,F. Marar. J,F. **Interfaces inteligentes para mecanismos de busca: *Design* de informação para *Web. Smart interfaces for searching engines: information design for the Web.*** Artigo aprovado, apresentado e publicado no SBDI, Sociedade Brasileira de *Design* de Informação, Recife PE, 11/09/2003.

Anexo3..... 179

Carvalho. R,F. Marar. J,F. ***WebDesign Goal*: Recuperação de Informações em Sistemas Distribuidos. *WebDesign Goal: Information Retrieval Systems*.** Artigo aprovado, apresentado e publicado no 2º Congresso Internacional de Pesquisa em *Design.* Rio de Janeiro RJ, 18/10/2003.

Detalhamento da dissertação

Capítulo 1

Este capítulo faz uma introdução aos processos de armazenamento e consulta, e às necessidades que causaram uma grande mudança na forma de se lidar com a informação distribuída através da *Internet*. Revela também a necessidade de aprimoramento quanto a forma de indexação e reconhecimento de informações contidas nos documentos digitais, para que se desenvolva uma maneira mais adequada de aproximar as informações contidas na *Internet* em relação às necessidades do usuário, tornando a experiência da pesquisa de fácil e rápido acesso.

Capítulo 2

Características dos Sistemas Distribuídos de Informação

O capítulo 2 abordará e detalhará a estrutura de funcionamento e características dos Sistemas Distribuídos de informação. Será possível entender como os sistemas de busca funcionam. E quais são as diferenças existentes entre mecanismos de busca e diretórios.

Será também discriminada a dimensão existente entre os sistemas de busca e sua relação com a eficiência direta na pesquisa realizada pelo usuário. Outras questões serão abordadas neste capítulo: qual a vantagem ou desvantagem entre mecanismos de busca e diretórios? São eles sistemas opostos, ou relacionados?



Quais as chances que o documento digital possui de ser cadastrado nas bases dos sistemas distribuídos de informação?

Capítulo 3

Visão do *Designer* sobre a qualidade e quantidade das informações disponibilizadas nos sistemas distribuídos

O capítulo 3 descreve uma visão ampliada do *designer* em relação às várias etapas da elaboração do projeto de um documento digital, que pode possuir vários direcionamentos, como pessoal, grupos específicos, corporativos e comerciais. Além disso, discute-se a importância da qualidade e quantidade da informação disponibilizada nos sistemas distribuídos de informação.

Este capítulo também revela os motivos da crescente dificuldade de se encontrar um documento digital e aborda formas para melhorar as possibilidades de classificação das informações para o usuário, seja a informação procurada pessoal, educacional, corporativa ou comercial utilizando os mecanismos de procura da *Internet*, sem que se cometam erros que possam causar a exclusão nas bases de dados.

unesp

Capítulo 4

Metodologia para interfaces inteligentes para mecanismos de busca

O capítulo 4 descreve como é realizada a pesquisa dentro dos mecanismos de busca e como deve ser o planejamento estratégico dos elementos visíveis e não visíveis dentro de um documento digital para que, através deles, se possa obter uma classificação relevante nas listagens de respostas. Na sequência descrita a proposta de metodologia para classificação otimizada de documentos digitais nos mecanismos de busca.

Capítulo 5

Conclusão

O capítulo 5 relata de maneira detalhada a utilização da palavra-chave em todos os lugares possíveis na estrutura interna e externa do *site* SACI, com o objetivo de ser um dos primeiros itens oferecidos nas respostas das listagens de busca. Além de reportar à data de envio do documento digital e sua submissão para alguns mecanismos de busca e diretórios. Com o envio do endereço do *site* SACI para grandes serviços de busca de informação, será possível comprovar a eficiência da utilização dos parâmetros informados no item 4.17. Será também possível comprovar o relacionamento existente entre os menores e os maiores mecanismos de busca, quando for pesquisada a existência do *site* SACI em um mecanismo no qual o mesmo não tenha sido cadastrado e, se, verifica uma relevante classificação.



Além disso, foi colocado em um provedor de acesso grátis uma cópia do mesmo *site* (SACI), mas sem que seu endereço fosse enviado para qualquer mecanismo de busca ou diretório. A função deste *site* "espelho" foi a de verificar a velocidade dos agentes em percorrer bases de dados com acesso livre de informações. Verifica-se, desta forma que apenas a composição do documento digital e posteriormente sua colocação em um provedor de acesso não é suficiente para a sua classificação de forma satisfatória, mesmo sabendo que os agentes (*spiders*) ficam percorrendo constantemente as bases de informações da *Web.*

Complementarmente, será divulgado neste capítulo o processo de criação em *Webdesign* desenvolvido para o *site* SACI, desde estudos de casos de *sites* com informações semelhantes. Utilizou-se um método adaptativo de design de *brainstorm,* a planejar o documento digital adequadamente para que seus elementos pudessem ser transmitidos em velocidade adequada à expectativa dos usuários. Realizou-se por fim a descrição da concepção visual, navegabilidade e usabilidade desenvolvida no documento digital.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Capítulo 1

Detalhamento da dissertação..... 12

1.1 Introdução..... 17

1.2 Redes de Computador e *Internet*..... 19

1.3 Sistemas Distribuídos como Suporte à Segurança de Informação..... 20

1.4 O problema: Otimizar as Possibilidades de Classificação de documentos digitais e encontrar informação segura..... 22

Capítulo 1

1.1 Introdução

A comunidade científica investe em desenvolvimento de máquinas inteligentes, que possam fazer com que o trabalho profissional da ciência, da arte e da tecnologia, torne-se mais eficiente. Muito antes da Revolução Industrial, a indagação tem sido uma das principais ferramentas para que novos produtos possam desempenhar atividades que permitam a evolução da relação entre o ser humano e a máquina, na qual, a máquina deva ser adaptada às necessidades do usuário e nunca o oposto.

No período compreendido entre a Segunda Guerra Mundial e o pós-guerra houve grandes avanços neste campo do conhecimento. Nessa época, Vannevar Bush[1], coordenava o trabalho de mais de seis mil cientistas e uma das questões enfrentadas por ele era o volume crescente de dados que deveriam ser armazenados e

unesp

organizados de tal forma que esse armazenamento permitisse a outros pesquisadores a utilização destas informações de maneira rápida e eficiente (Johnson, 2001).

O volume de publicações, contudo, cresceu tanto que tomar conhecimento das novas técnicas manter-se atualizado em relação aos novos avanços de maneira cada vez mais rápida e eficiente, abrangendo todos os tipos de suportes, tornou-se tarefa impossível de ser realizada. Isto gerou a necessidade de uma instituição mais dinâmica que se antecipasse às demandas dos usuários, que além de selecionar, processar e armazenar o acervo intermediasse também o fluxo da informação (Luz, 1997).

Isso porque, as formas de armazenamento de informações conhecidas até aquele período, por mais eficientes que fossem, acabavam oferecendo dificuldades em relação ao acesso e arquivamento. Grandes quantidades de papéis, relatórios, documentos, livros, poderiam estar bem ordenadas ou indexadas em estantes, mas a criação constante de novas informações exigia cada vez mais espaço. Para eliminar este problema seria necessária a criação de uma nova tecnologia para armazenar e acessar a informação. Comparativamente, o cérebro opera por associação, o que torna o processo de indexar a informação, tanto de forma alfabética como numérica ineficiente. O pensamento é mantido em uma teia de conhecimento no cérebro. Assim, seria ideal encontrar uma forma de se fazer algo análogo de forma automatizada. (Gardner, 1999)

A informação pode implicar em várias linguagens e diferentes suportes. Equivocadamente pensamos em informação apenas como texto impresso, mas é possível obter atualmente informação na forma de som e/ou de imagem em suportes eletrônicos. Quando estes sistemas se combinam, a informação tem uma chance maior de se tornar conhecimento, muito mais rapidamente que qualquer uma

(1) Vannevar Bush, foi presidente do *Massachusetts Institute of Technology* (MIT) e diretor do *Office of Scientific Research and Development* no período da IIª Guerra Mundial, nos Estados Unidos. Veja o historioco texto de 1945 em: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm ou http://www.unicamp.br/~hans/mh/memex.html

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

das formas já citadas individualmente, pois temos vários sentidos em pleno processo de captação de informação (Idem, Ibidem).

A informação se torna conhecimento quando é necessário utilizá-la de maneira adequada à resolução de um dado problema. Assim, o cérebro processa o conteúdo informacional e reage à dificuldade existente tentando resolvê-la e, em algumas situações, constata que não é possível apenas com o conhecimento adquirido. Desta forma, a partir de uma dificuldade inicial, vários procedimentos podem ser adotados. Se a informação for suficiente para que se possa solucionar a dificuldade, a resposta se concretizará da forma mais rápida possível. Caso contrário, será necessário adquirir outras informações. Assim, é possível optar por um processo linear de aprendizagem, no qual, é preciso passar por todos os pontos do processo, até que se encontre a resposta para a resolução da dificuldade. Entretanto, é possível adquirir apenas a informação necessária para a resolução da dificuldade sem, contudo, construir uma base de conhecimentos. Em algumas situações encontrar apenas a informação necessária pode ser suficiente, mas em outras, esse procedimento pode acarretar sérios problemas.

1.2 Redes de Computador e *Internet*

O lançamento do satélite *Sputnik*, da então União Soviética, em 1955, provocou nos Estados Unidos a desconfiança de que o inimigo russo já possuía meios eletrônicos de armazenamento e disseminação da Informação. Isto estimulou este país a se lançar em busca de métodos mais adequados de produção e uso de conhecimento (Luz, 1997, p60.).

Em 1969, foi realizada a primeira conexão entre computadores nos *campi* de quatro universidades situadas em posições geográficas distantes. A experiência aconteceu em *Stanford Researsh Institute* (SRI), Universidade da Califórnia



(UCLA), Universidade de Santa Bárbara e a Universidade de *Utah*, em *Salt Lake City*, formando quatro pontos em uma rede de computadores interligados. A experiência realizada não tinha precedentes. O Professor Dr. Leonard Keinrock[2], pioneiro em Ciência da Computação da UCLA, e seu pequeno grupo de estudantes esperavam se conectar com o computador de Stanford e tentar enviar alguns dados. O grupo começou a digitar algumas teclas e esperavam que essas letras aparecessem em um outro monitor de vídeo em outra universidade. Assim, para verificar se o que era digitado em um computador estava sendo transmitido para um outro computador em outra universidade, foi utilizado o telefone, para confirmar verbalmente a chegada dos caracteres. Neste teste foram digitados sucessivamente os caracteres "L", "O", "G". Em seguida o sistema interrompeu a conexão e, a partir deste momento, originou-se por assim dizer, a revolução da comunicação digital. Desta forma, acelerou-se o processo de "aldeia global", termo criado por Marshall McLuhan em 1964 para explicar a simultaneidade em que o tempo e o espaço desaparecem, quando os meios eletrônicos de comunicação começam a envolver o ser humano. (McLuhan, 1996)

Os quatro pontos formados por aquelas universidades deram início ao que conhecemos hoje por *Internet*, não com a mesma aparência atual, mas funcionando estruturalmente da mesma forma que no passado (estrutura descrita no item 1.3). Como conseqüência desta experiência, a informação tem rompido várias barreiras como tempo, idade, distâncias e custos, criando desta forma, a sociedade da informação, ou a sociedade da "velocidade elétrica", como descreve Johnson (Op. cit., p9).

1.3 Sistemas Distribuídos como Suporte à Segurança de Informação

A arquitetura desenvolvida para o funcionamento da transmissão de conteúdo através da *Internet* foi elaborada para que nenhuma das bases

(2) Informações complementares sobre o Professor Dr. Leonard Keinrock, podem ser obtidas em:
http://www.lk.cs.ucla.edu/
http://www.ccst.ucr.edu/ccst/home/speakers/LKleinrockbio.html

possuíssem a totalidade das informações, simplesmente para assegurar que os computadores conectados não parassem de funcionar se um deles, por algum motivo, sofresse algum dano, ou que o computador que armazenasse todos os dados pudesse ser atingido e, conseqüentemente, parar toda a comunicação realizada através da rede formada pelos computadores. É o que se chama de Sistema Distribuído em rede ou hipermídia "distribuída"[3].

Desta forma, era possível um computador acessar informações contidas em uma outra base de dados, que poderia estar a uma grande distância do ponto inicial de procura, sem, contudo, causar demora no acesso e transmissão das informações, desde que o usuário consultante possuísse acesso à base em que os dados fossem encontrados. Amplia-se assim, o alcance do ser humano e começa a deixar virtualmente a distância da informação a um clique do usuário.

Através do desenvolvimento dos Sistemas Distribuídos e com a informação descentralizada, qualquer base de dados que por algum motivo estivesse fora de funcionamento não alteraria os outros computadores que formam as outras ligações da *Internet*, permitindo a normalidade de suas operações, apenas não tendo acesso às informações da base com problemas. E assim, se por algum motivo, uma rota não estiver funcionando, o sistema automaticamente a desvia, utilizando um caminho alternativo disponível para que a transmissão seja realizada, evitando a paralisação de contatos entre bases (Anderson & Kubiatowics, 2002).

Além disso, os documentos digitais que trafegam nessas rotas nos sistemas distribuídos não funcionam apenas com a elaboração do *design*, do conteúdo e da programação. Há também a arquitetura de informação[4], responsável por permitir que o usuário encontre o que procura com o menor número de interações possíveis. Para conseguir isso é necessário organizar o conteúdo de maneira clara e específica. A arquitetura de informação permite que os visitantes saibam onde estão e para onde

(3) Ver Francis Heylighen (1994 p:55-68) que cria o conceito de hipermídia "distribuída".
(4) Arquitetura de informação, a estrutura e organização lógica de funcionamento de um sistema computacional.

unesp

podem ir. É a definição clara do caminho lógico para se encontrar a informação.

A Arquitetura da Informação visa à organização de grandes massas de dados, preparando rotas de acesso a eles o que, posteriormente, desenvolverá os sistemas distribuídos. Uma arquitetura eficiente torna a informação acessível e compreensível aos usuários, qualquer que seja seu nível de conhecimento.

O primeiro relato identificando a organização de informações ocorreu na Biblioteca de Alexandria, quando surgiu a figura do profissional bibliotecário. Calíacus (305-240 a.C), versátil e admirado poeta grego, reconhecido como um dos primeiros administradores daquele centro cultural, separou as obras por tipos de autores: poetas, jurisconsultos, filósofos, historiadores e estabeleceu o primeiro catálogo sistemático de que se tem notícia, demonstrando sua preocupação com o acesso ao documento e à informação. (Luz, 1997, p48.).

1.4 O problema: Otimizar as Possibilidades de Classificação de Documentos Digitais e Encontrar Informação Segura

O propósito da *Internet* sempre foi o armazenamento de informação através de um acesso rápido. Mas, com o passar do tempo, podemos notar que seu funcionamento não atingiu plenamente esse requisito, como foi planejado. Ao contrário, desperdiça-se muito tempo na pesquisa e, muitas vezes, não se encontra nela aquilo que se deseja.

Apesar de a *Internet* ter mais de 30 anos, sua abertura ao acesso público chega há quase 10 anos, e a quantidade de informação torna-se um grande problema (Bharat, 2000; Chang at. al., 2000; Gandal, 2001). Como encontrar a informação necessária em uma simples pesquisa que pode nos trazer muito mais de um milhão de alternativas? Segundo (Kwok at al., 2001, p. 242), a crescente base de

dados que amplia e dificulta o rastreamento de informações, tornando uma pesquisa simples na *Web* uma tarefa às vezes problemática, ou pela falta ou por encontrar uma enorme quantidade de informações que podem não estar bem classificadas em uma listagem de respostas, fazendo com que o usuário tenha que percorrer vários itens de uma classificação equívoca, antes de encontrar a informação que o satisfaça. Os mecanismos de busca não funcionam todos da mesma forma, alguns possuem mais informações e outros menos. Alguns mecanismos se relacionam, outros não. Como se pode avaliar e confiar na relevância do resultado oferecido pelo mecanismo de busca?

Alguns estudiosos[4] afirmam que apenas 20 por cento de todo material depositado na *Internet* têm chance de ser acessado, pois certos métodos de cadastramento do documento digital ou são desprezados ou são desconhecidos por quem disponibiliza a informação. Assim, o material publicado na *Internet* permanece oculto, sem acesso, pelo fato de que procedimentos de identificação foram ignorados. Por isso, mais um instrumento foi projetado para a *Internet*, o mecanismo de busca. Nos últimos anos a *Web* cresceu tanto que é impossível existir um único lugar que inclua todos os *sites*[5]. Gráfico 1.

O presente trabalho se enquadra neste contexto.

Dimensão dos Mecanismos de Busca
(Bilhões de páginas *Web*)

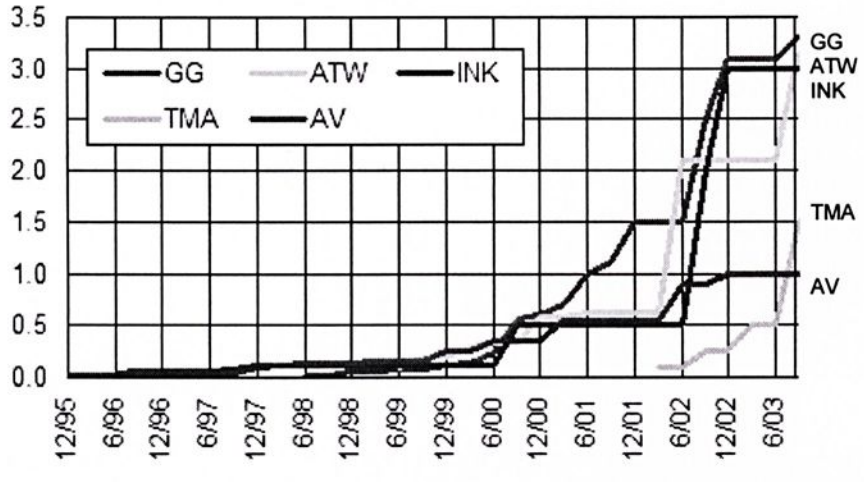


Gráfico 1. Comparativo de crescimento das bases de dados. GG-*Google*, ATW-*All the Web*, , INK-*Inktomi*, TMA- *Teoma*, AV-AltaVista 02/09/2003. Fonte: ttp://www.searchenginewatch.com/reports/article.php/2156481. acessado em 03/09/2003.

(4) Brin, (1998), Castro, (2000), Sullivan,(2000) consultor de *Internet* e estudioso do comportamento dos mecanismos de busca e como fazem a indexação dos *sites* desde 1995, também foi o fundador do *site Search Engine Watch,* http://www.searchenginewatch.com. Leavitt, co-fundador da empresa *Lycos* e CEO (*Chief Executive Officer*) em tecnologia *Web*, http://www.steampunk.com/jrrl/vita.html.
(5) *Site*, referência a uma posição virtual, no qual se pode encontrar um ou vários documentos digitais ou páginas de *internet*.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Capítulo 2

Características dos Sistemas Distribuídos de Informação

2.1 Introdução..... 26
2.2 Funções básicas de um sistema de busca..... 28
2.3 Classes de sistemas de busca..... 30
2.3.1 Partes de um mecanismo de busca..... 31
2.3.2 Organização dos mecanismos de busca..... 33
2.4 Diretórios..... 35
2.4.1 Riscos de não cadastramento em diretórios..... 36
2.4.2 Diferenças entre diretórios e mecanismos de busca..... 36
2.4.3 Mecanismos de Busca e Diretórios: características..... 38
Tamanho..... 38
Modos de cadastramento e de classificação..... 40
Atualização..... 41
2.5 Pesquisa Híbrida..... 41
2.6 Custos Operacionais..... 42
2.7 Funcionamento dos *Sites* de Sistemas de Busca..... 42

Capítulo 2

Características dos Sistemas Distribuídos de Informação

2.1 Introdução

Os sistemas de busca foram criados para facilitar o acesso à informação e foi justamente a partir desse ponto, devido à crescente dificuldade de se encontrar informações na *Web*, proporcionalmente ao aumento gradativo das informações nas bases de dados na *Internet* (Tu & Hsiang, 2000), que a ela foi associado um banco de dados para o guia de busca e assim facilitar o processo de localizar os *sites* e as informações dentro da *Web*. Dessa maneira, iniciou-se uma corrida para desenvolver um guia de busca que funcionasse dentro dos moldes da *Internet*. A primeira forma de tentar organizar a *Internet* foi o WAIS[1] (*Wide Area*

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Information Servers - Servidores de Informação de Área Ampla), introduzido em 1991 e continua em uso até hoje. Através do WAIS era e é possível criar bancos de dados indexados, incluindo textos, sons, gráficos e vídeos. No mesmo ano foi introduzido o *Gopher*[2], mais um mecanismo destinado a interligar computadores para facilitar a pesquisa de informações entre universidades. Entretanto, nenhum deles funcionava da mesma forma que os sistemas de busca de hoje. Desta forma, informações que podem ser valiosas e que podem estar dentro da base de dados WAIS ou *Gopher* não serão acessadas em procuradores de informação que são usados atualmente.

A WWW (*World Wild Web*) surgida aproximadamente em 1990, corresponde ao domínio da *Internet* construída a partir dos princípios do hipertexto. Embora o termo hipertexto tenha sido criado por Ted Nelson na década de 60, assim como a idéia de texto em estrutura multidimensional, é a Tim Berners Lee que, trabalhando no laboratório suíço CERN, se deve o desenvolvimento da WWW. (Rosenstiehl, 1988, p: 228-46)

Para que um usuário tenha sucesso em obter a resposta de sua pesquisa utilizando um guia de busca é necessário que os dados procurados já estejam indexados nas bases de informações. Esses dados precisam estar projetados adequadamente para que possam ser encontrados e não apenas visualizados e operacionalizados. Para que isso aconteça, uma grande combinação de recursos pode ser inserida no documento digital facilitando o rastreamento das informações através dos sistemas de busca sem, entretanto, desprezar as características de indexação de cada mecanismo que podem variar drasticamente de um para outro. Essa variação

---

(1) WAIS, ferramenta da *Internet*, para pesquisa de informações, foi um projeto criado inicialmente pela *Apple*, *Thinking Machines* e *Dow Jones* e tem como idéia fazer com que o usuário só receba o que desejar dentro de um vasta seleção de opções (Que pode ser servidores de listas, *NewsGroups*, *Gopher's*, FTP's, ARCHIE, *Telnet's* ou WWW).
O servidor WAIS é um recuperador de informações similar ao *Gopher*, que ajuda o usuário a achar e recuperar documentos e outras informações. A diferença é que, enquanto o *Gopher* é indicado para usuários iniciantes pois é baseado em uma interface amigável dirigida por menus, o WAIS se destina ao usuário mais avançado que sabe o que quer, embora possa não saber onde localizar. O WAIS tem a capacidade de localizar um documento baseando-se no conteúdo do mesmo, fazendo busca a partir de palavras-chave ou qualquer outra cadeia de palavras contidas em um documento. Fonte: www.w3c.org

(2) *Gopher* é o nome dado a uma ferramenta da *Internet* criada pela Universidade de Minnesota. É um sistema de computadores (servidores *Gopher*) interligados através da *Internet* contendo áreas de arquivos que podem ser acessados através de programas especiais de navegação *Gopher*. O nome *Gopher* foi dado em homenagem a um pequeno roedor norte-americano (marmota) que cava túneis subterrâneos formando uma rede. O Sistema *Gopher* é utilizado basicamente pelas Universidades para divulgar textos científicos, imagens, programação, serviços, cursos, etc. servindo como um veículo de intercâmbio universitário. Fonte: www.w3c.org



acontece devido a estratégias próprias dos mecanismos, os quais acabam oferecendo respostas diferentes para uma mesma pesquisa. Isso acontece porque as bases de dados dos sistemas de busca possuem algoritmos[3] de classificação, computadores, programas e também bases de informações diferentes, o que pode causar variações nas respostas obtidas.

Em 1994, a base de informações da *Web* continha aproximadamente 110.000 páginas. Em novembro de 1997, era possível encontrar dois milhões de páginas. A base de dados não apenas cresceu, como também cresceu a quantidade de procuras realizadas. Em março e abril de 1994, a quantidade de consultas realizadas aos procuradores existentes estava registrada em 1500 por dia. Em 1997, o procurador AltaVista declarava que havia um número equivalente a 20 milhões de procuras por dia (Brian & Page, 1998. p.108).

Vale destacar que, ao realizar uma pesquisa em qualquer sistema de busca, o usuário não está pesquisando diretamente a *Web*. Está sim pesquisando uma base de dados localizada em um *site* da *Web*, (Chang et al., 2000) que é o *site* do sistema de busca usado pelo usuário. Nenhum sistema de busca tem em sua base de dados toda a *Web* catalogada, todo o universo de mais de um bilhão de páginas. Segundo *Search Engine Watch*[4], o maior dos sistemas de busca é o *Google*[5], que possui 3.307.998.701 páginas *Web* indexadas em sua base de dados. O segundo maior sistema de busca, o *All the Web*[6], tem cerca de 3.000.000.000 páginas *Web* em sua base de dados (março de 2003).

2.2 Funções básicas de um sistema de busca

Os sistemas de busca possuem três funções básicas para que

(3) Algoritmo: uma seqüência lógica finita, que leve à resolução de um problema.
(4) *Search Engine Watch*, http://searchenginewatch.com/reports/sizes.html
(5) *Google*, http://www.google.com.br
(6) *Fast Search All the Web*, http://www.alltheweb.com

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

possam oferecer algum resultado ao usuário. São eles:

- Análise e a indexação (ou cópia) das páginas da *Web*;
- Armazenamento das "cópias" efetuadas e;
- A recuperação das páginas que preenchem os requisitos indicados pelo usuário por ocasião da consulta.

Um dos primeiros sistemas de busca a entrar em operação depois do WAIS e *Gopher* foi o *Lycos*[7]. Introduzido em 1994, seu principal diferencial em relação a seus primos "pré-históricos" foi a presença de um pequeno programa chamado *spider*, criado por John Leavitt[8]. Um *spider* é um programa cuja função é ficar rastreando informação pela *Internet*, visitando *sites* e catalogando o seu conteúdo, mantendo os bancos de dados da máquina de procura de páginas atualizados. Os *spiders* identificam páginas novas, atualizam páginas conhecidas, e apagam referências de dados obsoletos. Os mecanismos podem considerar uma informação obsoleta, por vários atributos, como por exemplo, quantidade de pessoas que acessam a página durante um certo intervalo de tempo, quais são os *sites* que possuem *links* para o material em questão, o uso de códigos que identificam tempo de validade (*meta tags* de tempo de validade) entre outros (Walker, 2001). Depois disso, o sistema de busca cria uma lista de palavras e a deixa disponível para que seja possível ao guia de busca localizar o que foi indexado na sua base de informações.

No mesmo ano em que *Lycos* criou seu sistema de busca, começou a funcionar o *Yahoo*[9] *(Yet Another Hierarchical Officious Oracle*, ou seja, Mais um Oráculo Oficioso Hierárquico), que na verdade começou como apenas um passatempo e hoje é um dos maiores sistemas de busca da *Web*. Atualmente os

(7) *Lycos*, http://www.lycos.com
(8) John Leavitt, co-fundador da empresa *Lycos* e CEO (*Chief Executive Officer*) em tecnologia *Web*.
(9) http://www.yahoo.com.br

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

sistemas de busca possuem Interfaces gráficas[10] bastante parecidas entre si. O motivo para tal semelhança na aparência dos sistemas é facilitar ao usuário, que pode trocar do atual sistema de busca por outro sem maiores conseqüências na operacionalidade e, em uma situação na qual está com dificuldades em encontrar um determinado assunto, o usuário pode usar um outro sistema de busca, sem ter que aprender como esse outro sistema funciona. Na verdade, essa é uma das regras de usabilidade de interface que deve ser sempre seguida: nunca faça o usuário aprender novamente um determinado procedimento para realizar uma função já por ele conhecida (Minasi, 1994). E mais, a maioria destes sistemas de busca possui recursos de customização, isto é, ajustes para uma busca mais apurada.

2.3 Classes de sistemas de busca

Os sistemas de busca dividem-se em duas classes: os diretórios e os mecanismos de busca e ambos têm a mesma finalidade;

- Do ponto de vista do usuário: possibilitar a localização de *sites* e páginas (*home pages*) que contém um determinado assunto ou abordar um determinado aspecto de um assunto;
- Do ponto de vista do proprietário-dono-autor de uma página: fazer com que o seu *site* seja localizado da maneira mais fácil possível pelo usuário, eventualmente um consumidor do produto divulgado no *site*.

  Para evitar confusão, será adotada aqui a seguinte terminologia:
- Sistema de busca é o termo genérico que engloba as duas categorias: os mecanismos de busca e os diretórios;

(10) Interface: Elemento material que assegura a ligação ou interação entre homem e o computador ou entre dois sistemas informáticos. Tipo de interface com o usuário, em que a interação está baseada no amplo emprego de imagens, e não restrita apenas a textos ou caracteres, e que faz uso de um conjunto de ferramentas que inclui janelas, ícones, botões, e um meio de apontamento e seleção, como o *mouse*.

unesp

- **Mecanismos de busca** são os sistemas de busca baseados no uso exclusivo de programas de computador para a indexação das páginas da *Web*;

- **Diretórios** são os sistemas de busca nos quais a indexação das páginas da *Web* é realizada por seres humanos.

Para criar a base de dados de um mecanismo de busca, o programa *Spider* sai visitando os *sites* da *Web*. Ao passar pelas páginas de cada *site*, o *Spider* anota os URLs[11] existentes nelas para depois ir visitar cada um desses URLs. Visitar as páginas, fazer as cópias e repetir a mesma operação cópia e armazenamento na base de dados do que ele encontrar nesses *sites,* essa é uma das formas de um mecanismo de busca encontrar os *sites* na *Web*.

Outra maneira do sistema de busca encontrar os *sites* na *Web* é o responsável pelo *site* informar ao sistema de busca qual o endereço, o URL do *site*. Todos os sistemas de busca têm um quadro reservado para o cadastramento, submissão ou inscrição de novas páginas. De forma geral, é um *hyperlink* que recebe diversas denominações, conforme o sistema de busca. Adicionalmente, enviar o endereço do documento digital para o sistema de busca é uma característica utilizada tanto para os mecanismos de busca quanto para os diretórios.

2.3.1 Partes de um mecanismo de busca

Vistos de uma forma simplificada, os mecanismos de busca têm três componentes principais:

- O primeiro componente é um programa de computador denominado "agente" (Johnson op. cit p.127) que aparece com nomes como *spider*, *robot*, *crawler*, *wanderer*, *knowbot*, *worm* ou *web-bot*. Aqui, em nosso trabalho, vamos chamá-los

(11) URL, *Uniform resource locator,* ou o endereço de acesso de um *site*.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

indistintamente de *spider*. Esse programa (*spider*) visita os *sites* ou páginas armazenadas em variados provedores na *Web*. Ao chegar em cada *site*, o programa *spider* pára em cada página e cria uma cópia ou réplica do texto contido na página visitada e guarda essa cópia para si. Essa cópia ou réplica irá compor a sua base de dados. O *spider* retorna ao local em uma base regular e, posteriormente, retorna sua visita para verificar se existe mudança de conteúdo para atualizar sua base de informações.

- O segundo componente é a base de dados constituída das cópias efetuadas pelo *spider*. Essa base de dados, às vezes denominada índice ou catálogo, fica armazenada no computador, também chamado de servidor do mecanismo de busca. Tudo o que o *spider* encontra vai para a segunda parte do motor de busca, o índex. O índex, algumas vezes chamado de catálogo, é como um livro gigante contendo uma cópia de todas as páginas que o *spider* encontrou e, se uma página muda, então este livro é atualizado com a nova informação. Logo após a visita e catalogação, os mecanismos de busca classificam as informações classificando dados semelhantes. Todos os dados que possuem a mesma identificação ou semelhança através de *keywords* (palavras-chave que podem identificar o assunto do documento digital), ficam armazenados em um mesmo agrupamento (*cluster*). Isso acaba trazendo nas listagens de resposta uma quantidade muito grande de informações e, muitas vezes, no meio dessas informações encontram-se os chamados "ruídos"[12], termo que descreve os dados que não possuem nenhum vínculo com a informação procurada (Tu & Hsiang, 2000).

  É certo que pode levar um tempo para que novas páginas ou mudanças que o *spider* encontrou sejam incorporadas ao índex. Assim, uma página pode ter sido visitada pelo *spider,* mas não indexada e, até que ela o seja e incorporada ao índex, ela não estará disponível para a procura nos mecanismos de busca.

- O terceiro e último componente é o programa de busca propriamente dito. Esse programa é acionado cada vez que alguém realiza uma pesquisa. Nesse instante, o

(12) Na teoria de Informação designa-se por "ruído" tudo o que é indesejável no processo comunicativo.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

programa sai percorrendo a base de dados do mecanismo em busca dos endereços - os URLs - das páginas que contêm as palavras, expressões ou frases informadas na consulta. Em seguida, os endereços encontrados são apresentados ao usuário.

Estas máquinas de busca são sistemas que têm por objetivo encontrar informação de interesse dos usuários na *World Wide Web*. Em termos gerais, elas coletam continuamente os dados disponíveis na *Web* e montam uma grande base de dados que é processada para aumentar a rapidez na recuperação de informação. Sem os sistemas de busca seria praticamente impossível encontrar informações na *Internet*, uma vez que a quantidade de documentos digitais aumenta matematicamente, de forma exponencial. E é por essa última razão que uma pesquisa é feita rapidamente. Os mecanismos de busca não saem procurando a informação no momento em que se digita a *string* (conjunto de letras e ou números que podem formar uma palavra ou frase), mas sim, a procuram em um banco de informações já criado.

2.3.2 Organização dos mecanismos de busca

Para organizar as informações os *spiders* varrem a *Web* visitando páginas, lendo-as e extraindo os itens para eles interessantes, movendo-se de um documento a outro por referências de *hyperlinks* embutidos nas páginas e utilizando o protocolo HTTP[13] para recuperar documentos dos servidores. Eles utilizam técnicas heurísticas[14] com o objetivo de encontrar e indexar os *sites* mais populares primeiro. Como por exemplo: verificando quais são as páginas mais procuradas sobre um determinado assunto, quais as páginas que apresentam as palavras chaves digitadas pelo usuário em seu conteúdo e também quantas vezes essa palavra chave se repete no *site*. (Brian, S & Page, L 1998, p.111)

(13) HTTP, *Hyper Text Transfer Protocol* ou Protocolo de Transferência de Hiper Texto.
(14) Heurística, (Ferreira, 1999), metodologia ou algoritmo, usado para resolver problemas por métodos que, embora não rigorosos, geralmente refletem o conhecimento humano e permitem obter uma solução satisfatória.
Heurísticas: do grego *Heuristikein* significa achar, são métodos de descoberta particularmente utilizados em Inteligência Artificial, quando se procura apelar aos Sistemas Simbólicos. (que partem de axiomas para construir teoremas possíveis). Permitem discernir num conjunto de derivações as que possam conduzir ao sucesso. (Ganascia; 1993, p:63-7;113)

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Quando os agentes (*spiders, robots* ou simplesmente *Bots*) recuperam uma URL, podem aprender a revisitar freqüentemente alguns *sites*, principalmente aqueles que mudam muito, ou que têm *links* para muitas páginas. A estratégia dos *spiders* para seguir as complexas ligações dos hipertextos dão a impressão de que eles se movem entre os *sites* como se fossem vírus, mas na verdade eles apenas os visitam com o objetivo de coletar dados para a indexação destes em suas bases de dados. Com relação aos sistemas de busca: AltaVista[15], *Excite/NetCenter*[16], *HotBot*[17], *Infoseek*[18], *LookSmart*[19], *Lycos* e *Yahoo*, apenas o AltaVista, atualiza suas informações em um período máximo de 24 horas, o que, para o comportamento da *Web* pode ser muito relevante, em se tratando de atualização de informação. Em comparação ao procedimento do AltaVista, todos os outros mecanismos citados atualizam suas bases de informações entre 7 a 10 dias. É interessante ressaltar que a base do AltaVista possui 1 bilhão de páginas e, que em relação aos outros sistemas de busca citados neste parágrafo, é a maior base. Além do que, a base do Altavista também é utilizada por outros mecanismos como o *Looksmart* (Walker, 2001).

Desta forma, os critérios utilizados pelos *spiders* para efeito de visitação – "*crawling*" – indexação de novas páginas e itens a serem indexados variam de mecanismo para mecanismo. Segundo (Sullivan, 2000), alguns *spiders* utilizam "*deep crawl*", um recurso que permite acessar muitas páginas de um *site*, mesmo quando estas não são submetidas a eles. Alguns *spiders* utilizam visitação em largura na qual, uma vez que os mesmos se encontram em uma página, o *spider* visita todos os *links* existentes na mesma.

(15) AltaVista, www.altavista.com.br
(16) *Excite*, www.excite.com
(17) *HotBot*, www.hotbot.com
(18) *InfoSeek*, www.infoseek.com
(19) *LookSmart*. www.looksmart.com

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

2.4 Diretórios

Um diretório tem dois componentes principais para que possa oferecer auxílio na pesquisa de informações aos usuários que são:

- Uma base de dados, também chamada de índice ou catálogo;
- Um programa de computador que faz a pesquisa na base de dados.

A montagem ou criação da base de dados de um diretório é realizada por seres humanos. São eles, os humanos, que fazem a análise e a indexação dos *sites* da *Web*. Nos diretórios, não existem *spiders* para a catalogação e a indexação da *Web*. Quem realiza todo o processo de análise e classificação são os humanos.

Enquanto os mecanismos de busca copiam todo o conteúdo das páginas que encontram pela frente e mantêm todas as informações em suas bases de dados, os diretórios mantêm em suas bases de dados apenas um resumo do conteúdo dos *sites* por ele catalogados. Muitas vezes, esse resumo que fica na base de dados do diretório contém apenas o título do *site* e mais duas ou três frases sobre o assunto nele contido. Esse resumo tanto pode ser elaborado pelo autor da página ou por quem a envia, como por um editor, dependendo do diretório. Assim, o diretório tem a mesma finalidade dos mecanismos de busca: a indexação e a recuperação de páginas da *Web*. Mas, ainda assim, existem algumas características fundamentais entre os diretórios e os mecanismos de busca. Vejamos a seguir:



2.4.1 Riscos de não cadastramento em Diretórios

Ninguém, nem mesmo o responsável pelo *site*, vai ter jamais a menor garantia de que este *site* enviado aparecerá indexado num diretório que tem editores para analisar as páginas do *site* enviadas.

Os seres humanos, ao contrário dos programas de computador, são cheios de preferências pessoais e usam aquilo a que chamam critérios de relevância. É segundo esses critérios de relevância que os editores vão avaliar se o *site* merece ou não freqüentar os índices. Trata-se de avaliação subjetiva. Nesse caso, se o *site* do usuário for rejeitado, não se pode fazer nada.

Mas o que pode parecer um defeito é, sem dúvida, uma das qualidades mais marcantes dos diretórios que possuem equipe de editores. A análise dos editores tem por finalidade evitar que se faça a indexação de coisas imprestáveis, lixo que jamais vai ser procurado por um usuário.

Se alguém visitar um diretório e procurar pela palavra gato, o felídeo, dificilmente encontrará uma página pessoal de um gato homenageado pelo(a) proprietário(a)... Nos "bons" diretórios, dificilmente se encontra uma página do tipo: Eu (uma foto); Eu passeando (outra foto); Eu estudando (outra foto)... Eu me chamo Fulano... e assim por diante.

2.4.2 Diferenças entre diretórios e mecanismos de busca

A primeira diferença existente entre os dois tipos de sistema de busca é o modo do diretório encontrar na *Web* os *sites* a serem por ele indexados. Enquanto o programa *spider* do mecanismo de busca[20] toma, ou pode tomar, a iniciativa de sair visitando os *sites* e suas páginas pelo mundo afora (isso, quando os

unesp

arquivos do *site* já foram enviados para um provedor), o diretório espera que o usuário, o dono do *site*, tome a iniciativa de apresentá-los a ele.

A segunda diferença é que o diretório classifica o conteúdo dos *sites* segundo categorias e subcategorias, setores de atividade econômica ou ramos do conhecimento. Quando o usuário for informar o URL de seu *site* para que o diretório faça a indexação é necessário também dizer ao mecanismo de busca qual dessas categorias o *site* e/ou página pertence. (É bom lembrar que alguns mecanismos de busca também solicitam que o usuário informe o assunto ou a categoria do *site*.)

Mas a diferença mais significativa fica por conta de quem faz a indexação das páginas da *Web*, enquanto que, nos mecanismos de busca, quem realiza essa atividade é uma máquina, um programa de computador. Nos diretórios quem faz a análise e a indexação dos *sites* são seres humanos. Deste modo, é claro que os humanos, que trabalham na construção e na montagem dos diretórios, também usam programas de computador para criar a base de dados. Mas são eles, os humanos, que fazem a análise dos *sites* apresentados e é aí que se encontra a diferença fundamental entre os mecanismos de busca e os diretórios: a participação dos seres humanos na construção da base de dados.

Enquanto a Inteligência Humana descreve a organização da base de dados dos diretórios, por outro lado o processo de aquisição da base é realizado mecanicamente (virtualmente) pelos agentes.

E com relação ao cadastro do *site* e/ou página nos diretórios, o usuário ou o responsável pelo documento digital é quem deve tomar a iniciativa, sempre Informando ao diretório o título, a URL, a descrição do conteúdo, a categoria a que o *site* pertence e mais algumas informações complementares, como no caso do *Yahoo*.

Caso o diretório disponha de uma equipe de editores, eles irão até o *site* e farão uma vistoria no conteúdo. Uma vez aprovada a inclusão, o *site* é inscrito

(20) O termo mecanismo de busca designa *softwares* que utilizam agentes externos (*spiders*, etc) que possuem a função de auxiliar o usuário. Exemplo: *Google*, AltaVista, etc.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

no índice. Entretanto, também existem os diretórios sem editores, que aceitam as informações do jeito que foram enviadas e as arquivam no índice. Entre os diretórios com editores destacam-se o *Yahoo* e o *LookSmart*. O *Yahoo*, dos Estados Unidos, dispõe de uma equipe de mais de 100 editores, já o *LookSmart* tem cerca de 200 editores. Assim, quando ocorre de o diretório só fazer a inclusão do *site* após a conferência dos editores, três coisas podem acontecer:

- Talvez o *site* do usuário demore muito tempo para aparecer na base de dados deles. Alguns diretórios prometem fazer a inclusão de um *site* no prazo de duas semanas. Outros diretórios dão um prazo de até seis semanas;
- Talvez o *site* do usuário jamais apareça por lá ou;
- Talvez o *site* do usuário seja logo indexado e possa ser encontrado nas buscas efetuadas.

2.4.3 Mecanismos de Busca e Diretórios: Características

Os mecanismos de busca e os diretórios diferem entre si em três aspectos:

- **Tamanho;**

O tamanho da base de dados varia de mecanismo de busca para mecanismo de busca e de diretório para diretório e, na maioria das vezes, o conteúdo das bases de dados diferem, em detrimento da estratégia de indexação individual, utilizada por cada mecanismo. O tamanho da base de dados é responsável pela amplitude da pesquisa e quanto mais *links* o serviço possuir, mais itens ele recuperará e assim se tornará mais popular. Os itens coletados pelo *spider* durante o processo de indexação são armazenados nas bases de dados. Sendo assim, o conteúdo destas também é completamente dependente da estratégia de indexação utilizada pelo

mecanismo de busca. Exemplos de itens que podem ser encontrados nas bases de dados são endereços ou URL's, títulos, cabeçalhos, resumos, palavras da primeira ou primeiras linhas dentre outras. (Brian, S & Page, L 1998, p. 112)

Os mecanismos de busca são, de uma maneira geral, enormes. Vejam-se alguns números de setembro de 2003:

MECANISMOS DE BUSCA	NÚMERO DE PÁGINAS
Google	3.300.000.000
Fast Search All the Web	3.200.000.000
Inktomi	3.000.000.000
Teoma	1.500.000.000
AltaVista	1.000.000.000

Tabela 1. Dimensão dos mecanismos de busca
Fonte: http://www.searchenginewatch.com/reports/article.php/2156481

Tamanhos de alguns diretórios

DIRETÓRIOS	NÚMERO DE *SITES* INDEXADOS	DATA DA ESTIMATIVA
Looksmart	2.500.000	Janeiro 2003
Open Directory	2.714.693	Janeiro 2003
Yahoo USA	1,8 milhão	Janeiro 2003

Tabela 2. Dimensão dos Diretórios. Fonte: http://www.searchenginewatch.com/reports/article.php/2156411

Os diretórios, geralmente, têm uma base de dados de menor tamanho comparado aos mecanismos de busca. Mas isso não significa que eles produzam, necessariamente, resultados inferiores. Muito pelo contrário. Devido ao modo de criação de sua base de dados, eles têm, geralmente, um índice de relevância bem maior.

- **Modo de cadastramento e de classificação;**

Nos mecanismos de busca, há duas possibilidades de a página ser cadastrada:

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

1. Um programa de computador se encarrega de encontrar a página;

2. O responsável pelo documento digital informa ao mecanismo de busca o URL da página, para que a indexação seja feita de maneira mais rápida.

Para que o *site* apareça cadastrado no diretório, torna-se necessário ir até a página de cadastramento (submissão ou inscrição) e apresentá-la. O criador do *site,* ou responsável, terá de informar além do URL, a categoria a que o seu *site* pertence e mais um resumo do conteúdo, além do seu *e-mail*. Cada diretório possui o seu próprio sistema de classificação ou de definição de categorias. Segundo *Search Engine Watch*, o *Open Directory*[21] tem 361 mil categorias, o *LookSmart* 200 mil. Veja um exemplo de classificação. Ao se procurar no *Yahoo* Brasil por Chico *Science*, criador do movimento Mangue *Beat* no Recife, será encontrado em:

Entretenimento > Música > Artistas > Por Gênero > *Rock* e Pop Nacional > Mangue *Beat* > Nação Zumbi.

Mais um aspecto a considerar: enquanto um mecanismo de busca pode cadastrar todas as páginas de um *site* uma a uma, um diretório cadastra o *site* por inteiro (um *site* pode conter várias páginas). Não é raro encontrar nos resultados de uma pesquisa em um mecanismo de busca, várias páginas de um mesmo *site* abordando um mesmo assunto. Isto não ocorre num diretório.

- **Atualização**

As bases de dados dos mecanismos de busca são ou deveriam ser, de uma maneira geral, mais atualizadas do que as bases de dados dos diretórios. A razão dessa maior atualidade dos mecanismos de busca é muito simples. Os seus *spiders* não dependem dos seres humanos para a atividade de catalogação da *Web*. Já os diretórios dependem totalmente dos humanos para essa catalogação (Walker, 2001).

O tempo de espera, desde a apresentação de um *site* ou URL a

(21) Open Directory, http://dmoz.org/

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

um mecanismo de busca até que ele venha a surgir nas pesquisas, pode ser de um dia a três semanas. Nos diretórios, esse tempo de espera geralmente é muito maior.

No que se refere ao modo de pesquisa, os diretórios permitem as duas formas de pesquisa: através de palavras chave e através de diretório ou índice hierárquico. Dos mecanismos de busca mais conhecidos, apenas o *All the Web*[22] e o AltaVista não dispõem de índice hierárquico disponível em sua página de abertura. Fica difícil, portanto, estabelecer uma diferença entre os sistemas de busca baseando-se exclusivamente no modo de pesquisa.

2.5 Pesquisa Híbrida

Não se deve pensar em mecanismos de busca e em diretórios como duas entidades que se contrapõem e se excluem. Eles devem se complementar. Veja o caso do *Google*, um mecanismo de busca. Ele apresenta na página de abertura uma opção denominada "diretório" que oferece uma coleção de categorias. Neste caso, o usuário tem duas alternativas de pesquisa:

- Pesquisar através das categorias, ou;
- Fazer a pesquisa através de palavras-chave.

Ao clicar numa das categorias, o programa transfere o controle para o diretório *Open Directory* onde a pesquisa é realizada. Caso análogo ocorre com o *Yahoo*. Se o assunto indicado no quadro de pesquisa não for encontrado em sua base de dados, o *Yahoo* vai fazer a pesquisa na base de dados do mecanismo de busca *Google*.

(22) *All the Web*, www.alltheweb.com

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

2.6 Custos Operacionais

A maioria dos sistemas de busca prestam um serviço gratuito, o de divulgação do *site* ou de ajuda numa pesquisa. Nos serviços gratuitos, a remuneração deles vem da publicidade inserida, por exemplo, no topo da página. Além disso, outra propaganda mais insidiosa é aquela apresentada quando da resposta a uma consulta. O resultado é um produto que, por acaso, tem a ver com a sua consulta.

Alguns dos sistemas de busca para manter seus custos operacionais, cobram pela inserção do documento digital ou pelo fornecimento de informações ou artigos. São eles: o *Galaxy*[23] e o *Excite*.

2.7 Funcionamento dos *Sites* de Sistemas de Busca

Os *sites* de busca ou sistemas de busca possuem diversas maneiras de funcionamento e esta diversidade de opções acaba sendo um dos pontos relacionados ao desempenho da pesquisa. Mesmo que o mecanismo atue diretamente no desempenho, o planejamento interno do código do *site* e/ou página, com o propósito de tornar o documento digital mais visível para os mecanismos de busca, é fundamental e deve ser proposto pelo seu responsável. A escolha do mecanismo de busca e a adequada codificação interna do *site* e/ou página, convergem diretamente em direção ao sucesso ou fracasso na busca de informações. A seguir, foram relacionadas algumas características básicas e sistemas de busca.

- **Catálogo**

Funciona como uma lista de *sites* sobre vários assuntos. Como exemplo temos o *Yahoo* que, além de funcionar como catálogo de outros *sites,* funciona também através de uma busca por palavra-chave.

(23) *Galaxy*, www.galaxy.com

- **Palavra-chave**

Funciona da seguinte maneira: em se digitando uma ou mais palavras, o *site* procura em seu banco de dados em quais *sites* da *Internet* há ocorrência desta(s) palavra(s). A busca por palavra-chave (*keywords*) é possível em todos os sistemas de busca. Como a *Internet* cresceu de mais, será preciso ter paciência para encontrar o que se procura.

- ***Web-robots* ou *Bots***

Funciona com palavras-chave, mas usa *spider* para alimentar seus índices.

- ***Metasearch***

Funciona como uma pesquisa dentro dos *sites* de mecanismos de busca. Ao invés de fazer uma consulta em cada um dos principais *sites*, o metabuscador analisa informações de todos os *sites* de mecanismo de busca, ou em alguns deles, ao mesmo tempo.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Capítulo 3

Visão do *Designer* sobre a qualidade e quantidade das informações disponibilizadas nos sistemas distribuídos

3.1 Introdução..... 45

3.2 Interface..... 48

3.2.1 Metáforas..... 50

3.3 Arquitetura de informação..... 51

3.4 *E-Commerce, Design* e Sistemas Distribuídos..... 54

3.4.1 Características do *E-commerce*..... 56

3.4.2 Logística do *E-commerce*..... 57

3.5 Usabilidade..... 58

3.6 Integração entre *E-commerce* e *Design*..... 61

3.7 Visibilidade para Sistemas Distribuídos..... 62

Capítulo 3

Visão do *Designer* sobre a qualidade e quantidade das informações disponibilizadas nos sistemas distribuídos

3.1 Introdução

A *Word Wide Web*, também conhecida como www, w3 ou *Web,* foi criada entre final da década de 80 e início de 90 do século XX, no CERN (Laboratório Europeu de Física de Partículas) pelo físico Tim Barnners-Lee e pode ser traduzida como “Teia de Alcance Mundial”.

Até então, a *Internet* não possuía aplicações que apresentassem uma interface, digamos "amigável”. Boa parte de seus recursos era acessada

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

puramente através de texto. Seu uso chegava a ser restritivo àqueles que dominassem os comandos do Unix, sistema operacional amplamente utilizado na *Internet.*

A *Web* veio para mudar isso. Aproximou a *Internet* do usuário "comum" que queria cores, ícones e imagens, geralmente em ambiente *Windows*. Para isso teve papel decisivo o programa *Mosaic*, criado por Marc Anderssen, então estudante do NCSA (*National Center for Super computing Applications*). O *Mosaic* acabava com a era da navegação em modo texto: funcionava em modo gráfico e era capaz de exibir imagens. Figura 1.

Era o que faltava para a *Internet* "explodir" no mundo todo. Foi a partir da invenção da *Web* que milhares de empresas passaram a colocar informações na rede, bem como a prestar serviços ou dar suporte técnico a eventuais problemas. A partir daí, o número de usuários passou a crescer aceleradamente e uma mudança drástica aconteceu. Era necessário atrair o usuário para a *Web* e mostrar que esse novo meio seria mais uma e grandiosa extensão do ser humano. Essa atração deveria ser mais agradável que simplesmente uma tela de texto apresentada de um lado a outro do vídeo. Muitos itens deveriam estar reunidos para chamar a atenção e o interesse do usuário. Uma boa aparência e ou apresentação se tornaria essencial, mas não era o bastante. Na verdade, conteúdo é o item mais importante, mas se for bem apresentado conquistaria o usuário (Siegal, 1998).

Assim, um documento digital bem elaborado com características direcionadas ao público específico se torna fundamental. E para tornar o material o mais adequado possível, são necessárias várias recomendações, entre elas, o conhecimento de programação das várias linguagens que podem ser utilizadas para a *Web*, como também conhecimento das funções dos aplicativos gráficos, *webwriting*, *design*, arquitetura de informação, banco de dados, segurança de informação, comportamento dos sistemas distributivos de informação, público-alvo, etc.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

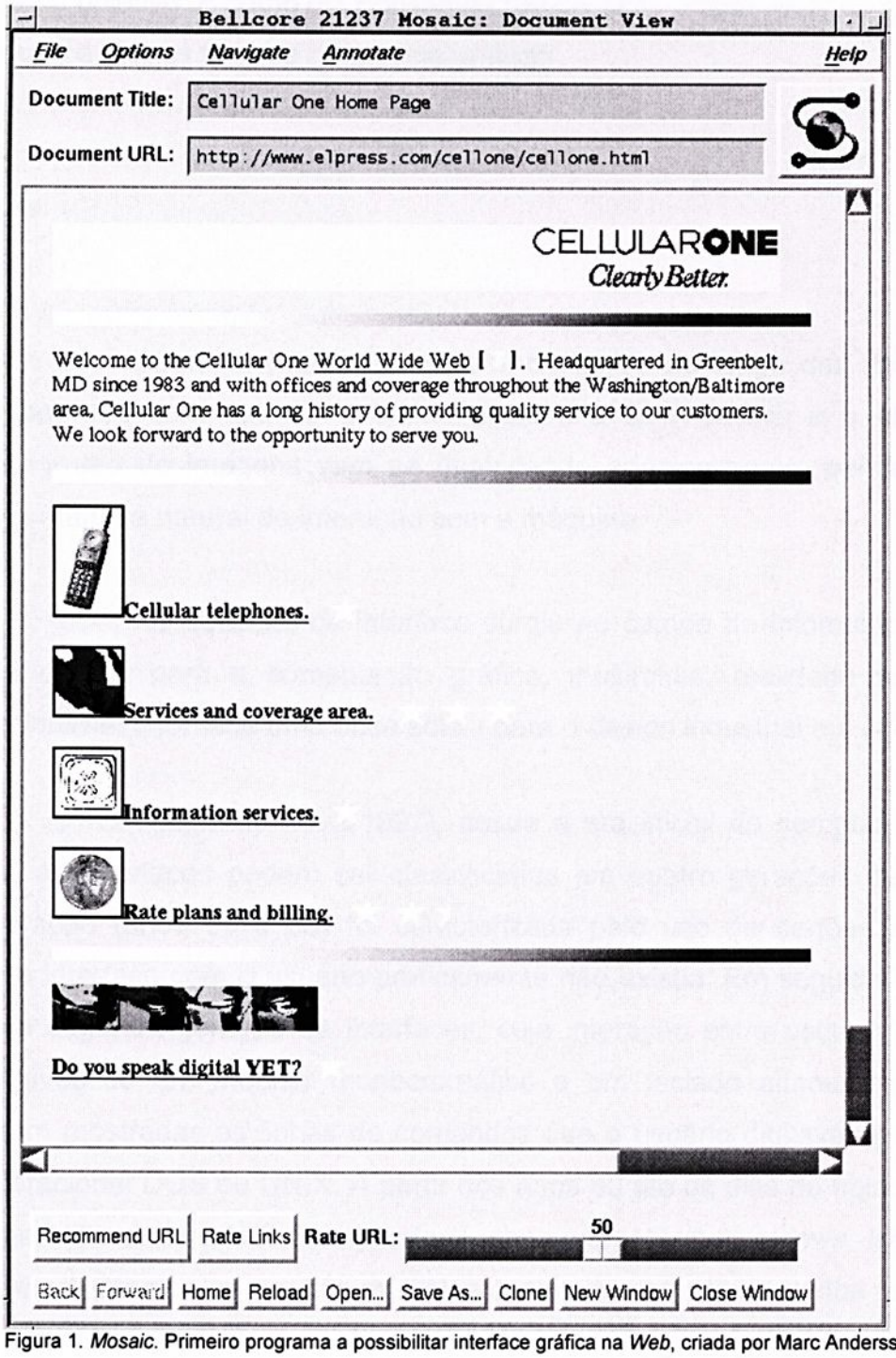


Figura 1. *Mosaic*. Primeiro programa a possibilitar interface gráfica na *Web*, criada por Marc Anderssen.

Reunir todos os itens descritos no parágrafo anterior e mais alguns outros em apenas um profissional pode acabar sendo muita pretensão. Desta forma, o mais adequado seria desenvolver o trabalho da elaboração do documento digital através de uma equipe de profissionais que possam unir seus conhecimentos em

prol do mesmo objetivo. E como diz Bonsiepe (1997 p.173), "Fazer *design* significa mais escutar o outro e menos fixar-se no próprio umbigo".

3.2 Interface

Com o desenvolvimento tecnológico ao longo dos anos surgem novas perspectivas com relação à interface entre o ser humano e o computador. Atualmente, o uso de imagens vem se destacando principalmente, por fornecer ao usuário uma maneira natural de interação com a máquina.

O conceito de interface surgiu no campo da informática. Ele tem importância central para a computação gráfica, multimídia, realidade virtual, tele-presença e *Internet* e fornece uma base sólida para o *design* industrial e o *infodesign*[1].

Segundo Silva(1997), desde a era inicial da computação até os dias atuais, as interfaces podem ser classificadas em quatro gerações de estilos. A primeira geração (anos 50 a 60) foi caracterizada pelo uso de cartões perfurados, sendo que a interface com o usuário praticamente não existia. Em seguida (anos 60 a 80), surge a segunda geração de interfaces, cuja interação entre usuário e máquina dava-se através de um monitor monocromático e um teclado alfanumérico. Nesse monitor, eram mostradas as linhas de comandos que o usuário digitava no *prompt* do sistema operacional DOS ou UNIX. A partir dos anos 80 até os dias de hoje predomina a terceira geração de interface. Tal interface, chamada WIMP (W*indows, Icons, Menu, Pointer devices*), oferece ao usuário múltiplos canais de entrada de dados para acesso e controle de múltiplas janelas, combinando texto com imagens gráficas, sons, vídeo e comunicação remota.

(1) Segundo Bonsiepe, (1997, p.146), possivelmente o termo *Infodesign* substituirá possivelmente, o termo *Design* gráfico. Ao invés de traduzir e transformar conceitos na dimensão do visível, o *designer* exercerá a função de um organizador autoral de informações. Esta reorientação separará o *design* gráfico do campo da publicidade, cujo peso econômico não se põe em dúvida.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Observando a evolução dessas três gerações de interfaces, nota-se que a tecnologia vem sendo empregada no sentido de tornar mais fácil e natural a interação entre usuário e computador. Como continuação dessa tendência, já é possível perceber o surgimento de uma quarta geração de interface, batizada por alguns autores como Van Dam(1997), como interfaces pós WIMP. Entre as principais características dessas interfaces, destacam-se a capacidade de reconhecimento de fala e, principalmente, o emprego de técnicas de Realidade Virtual. Tais técnicas permitem ao usuário interagir com as informações fornecidas pelo computador através de um ambiente gráfico tridimensional. Resalte-se, que este trabalho não se destina a se aprofundar sobre realidade virtual, a qual foi apenas citada para dar uma visão geral da evolução da interface.

Se as possibilidades abertas pela *Web* serão ou não utilizadas, dependerá em grande parte da qualidade da interface. A interface é um meio que:

- Pode frustar ou irritar;
- Pode facilitar ou dificultar a aprendizagem. Figura 2;
- Pode ser divertida ou chata;
- Pode revelar relações entre informações ou deixá-las confusas;
- Pode abrir ou excluir possibilidades de ação efetiva instrumental ou comunicativa (Bonsiepe, 1997).

Do ponto de vista do usuário, a interface é sinônimo de programa.



Figura 2: exemplo de como a elaboração de uma interface pode facilitar ou dificultar. Jornada 568 Personal Organizer.

3.2.1 Metáforas

A interface gráfica também foi responsável pela evolução dos *softwares* e contribuiu muito para a aproximação dos usuários em sua fase de reconhecimento e uso de funções, simplesmente porque fazia uso de metáforas que o usuário pudesse conhecer, como botões, controles deslizantes e ícones. Por sua vez, o usuário momentaneamente acabava esquecendo que estava usando uma máquina que internamente trabalha com impulsos elétricos, entendidos estes por zeros e uns, ou por uma linguagem binária, esta considerada de baixo nível.

É importante salientar que não é qualquer interface gráfica que fará o usuário navegar no mar de informações contidas internamente no computador, e também fora dele, como é o caso do acesso de informações realizada via *Internet*. Se a interface gráfica não for considerada como parte do projeto de interatividade e meio

facilitador no processo de transmissão de informação, o *software* corre o risco de não ser entendido pelo usuário. E, uma vez não entendido, todo o projeto funcional referente ao *Back end* (programação do *software*) será corrompido pelo *Front end* (visual final do *software* apresentado na tela do computador). Isso, no entanto, não significa que um é mais importante do que o outro. Na verdade, os dois devem trabalhar juntos em prol do melhor funcionamento do *software* para o usuário.

Podemos também vincular os estudos de interface com o usuário fazendo uma relação com a interface dos sistemas de busca. Segundo Dempsey at. al.,(2000, p.270), eles devem oferecer interfaces intuitivas para que facilitem a interação de usuários com mais idade ou com pouca experiência na utilização do sistema de busca. Um pequeno teste realizado pelos autores com uma amostra de 19 pessoas entre 18 a 29 anos mostrou que apenas nove usam sistemas digitais e mecanismos de busca. A pesquisa ainda mostra que outros 10 usuários entre 30 anos ou mais quase não têm contado com tecnologia digital.

3.3 Arquitetura de informação

Adicionalmente ao uso da interface e da metáfora, um outro item possui relevância fundamental: a estrutura hierárquica em que a informação será apresentada ao usuário. Esta estrutura pode ser classificada em três tipos diferentes, apresentadas a seguir:

- Arquitetura de informação linear.

É a informação que precisa seguir seqüência do início ao fim de forma integral. Pode ser exemplificada como alguns programas apresentados na TV, alguns programas de rádio ou um livro de estrutura narrativa, nos quais é necessário acompanhar toda a seqüência para que se possa compreender a informação. Figura 3.

unesp

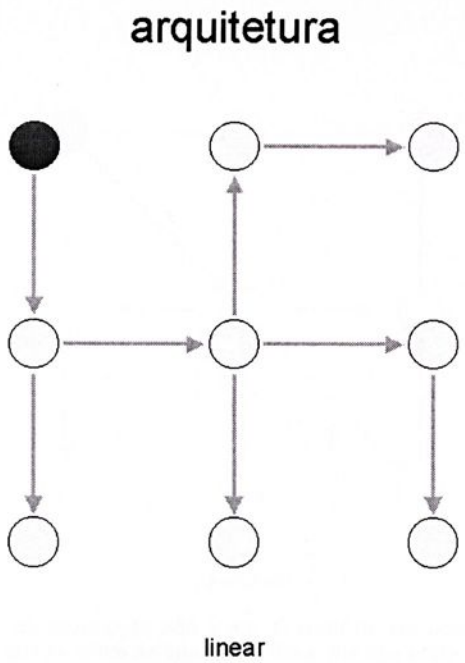


Figura 3. Demonstração de arquitetura de informação linear. Não é possível passar para outro item sem ter que passar pela seqüência previamente elaborada.

- Arquitetura de informação não linear.

Pode ser exemplificada como os vídeo *clips* dos canais de TV que transmitem conteúdo musical, tiras de quadrinhos em que não há necessidade de continuar lendo o próximo jornal para entender o acontecimento, a lista telefônica, um manual de referência ou a edição de imagens de certos comerciais, que usam seqüências de imagens que não possuem continuidade, mas como resultado final, se obtém um conjunto atraente. Figura 4.

- Arquitetura de Informação de Hipertexto[2]

É a possibilidade de o usuário seguir o seu próprio caminho dentro de conexões pré-definidas em um documento. O hipertexto não é um termo aplicado apenas na *Internet,* mas em muitos outros recursos de comunicação, como por exemplo, o editor de textos e o DVD (*digital vídeo Disk*), Figura 5.

(2) Segundo Pierre Lévy (1999 p.27), hipertexto é um formato digital, reconfiguravel e fluido. Ele é composto por blocos elementares ligados por *links* que podem ser explorados em tempo real na tela. A noção de hiperdocumento generaliza, para todas as categorias de signos (imagens, animações, sons etc.), o princípio da mensagem em rede móvel que caracteriza o hipertexto.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

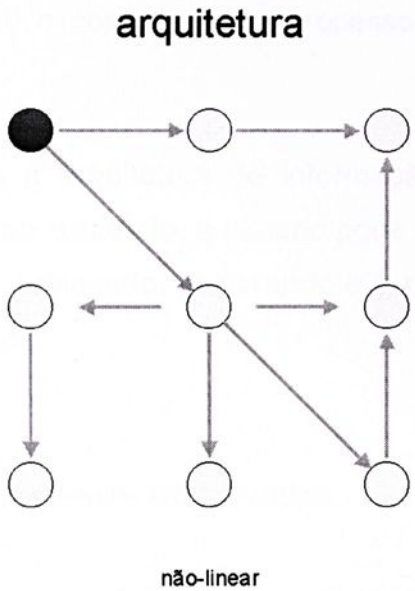


Figura 4. Demonstração de arquitetura de informação não linear. A partir de um ponto inicial, é possível optar por seqüências diferentes ao se realizar um percurso, o que se difere da arquitetura linear por não seguir uma ordem seqüencial.

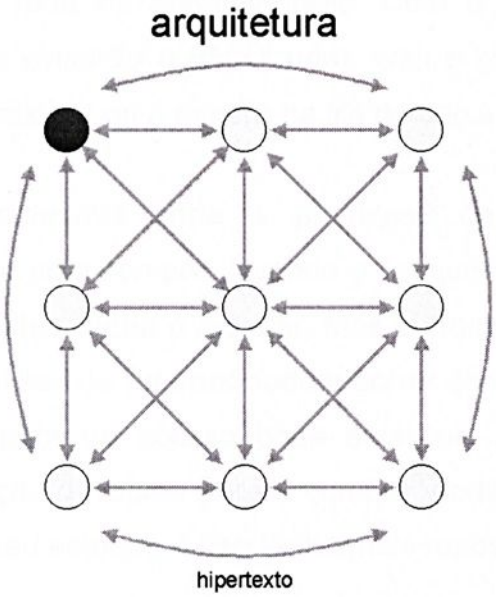


Figura 5. Demonstração de arquitetura de informação de Hipertexto. A partir de um ponto inicial, é possível realizar qualquer caminho, independente da ordem que a informação foi estruturada.

Se o conteúdo *Web* possuir uma arquitetura de informação linear, as chances desse material estar comprometido com a navegabilidade será muito grande (Radfharer, 2001). O ideal é que se use a arquitetura de hipertexto para dar ao usuário o controle geral do documento, pois quem está navegando na I*nternet* deve ter

a sensação de quem possui o controle ou o processo ativo de interação (Nielsen, 2000).

Se a arquitetura de informação, não for planejada com o objetivo de facilitar o acesso ao conteúdo, o usuário pode acabar desistindo do que está fazendo: lendo, pesquisando, brincando, comprando etc, desestimulado pela dificuldade na efetivação da ação proposta.

3.4 *E-Commerce*, *Design* e Sistemas Distribuídos

Com o aparecimento da *Internet* também surgiu o conceito de comercializar utilizando a rede de computadores. Com o *e-commerce*, poder-se-ia comprar sem sair de casa, evitando o trânsito das ruas e avenidas congestionadas à procura de vagas para estacionar ou à espera na fila do caixa.

O *e-commerce* tinha a promessa de acabar com todas as necessidades periféricas de uma compra, levando o consumidor a realizar eficasmente apenas o processo de escolher pagar e receber. Mas o processo nem sempre funciona como o desejado, as conexões de *Internet* podem sofrer com o congestionamento das linhas telefônicas, o provedor de acesso pode estar em manutenção e não estar disponibilizando a informação desejada para a compra, pode acontecer que o produto oferecido tenha esgotado seu estoque e não tenha mais reposição, etc.

Além disso, a segurança no envio do número do cartão de crédito é outro ponto muito importante. Quais são as garantias que a empresa oferece ao consumidor se, por algum motivo, o número do cartão de crédito for usado indevidamente? Essa e também outras questões relacionadas ao *e-commerce* devem ser explicadas no próprio *site*, para que o processo possa oferecer segurança, confiabilidade e transparência ao consumidor. Figura 6.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

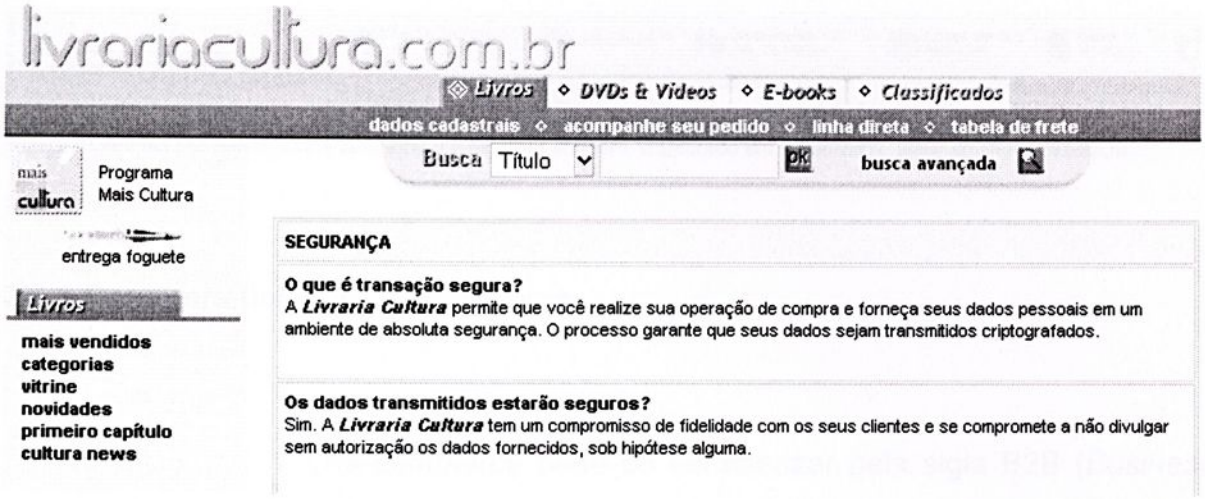


Figura 6. Política de privacidade. Capturado em 28/08/2002. www.livrariacultura.com.br.

O *e-commerce* deve também como opção oferecer vendas efetuadas através de boletos bancários Figura 7, ou mesmo vendas por telefone Figura 8, já que certas pessoas não confiam na segurança do envio do número do cartão de crédito e muitos outros consumidores não o possuem.

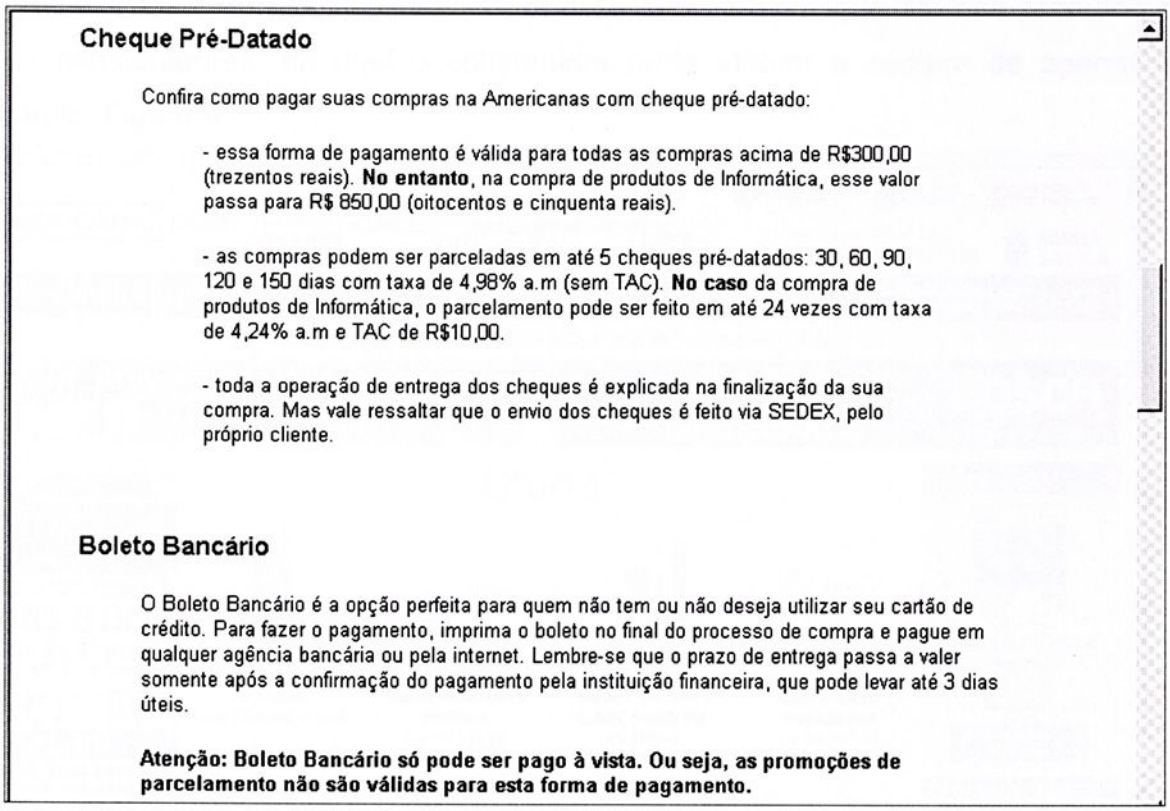


Figura 7. Formas de pagamento diferentes. Capturado em 28/08/2002. www.americanas.com.br.

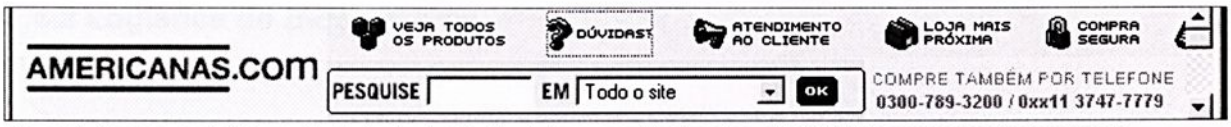


Figura 8. Vendas por telefone vide lado direito inferior da imagem. Capturado em 28/08/2002. www.americanas.com.br.

3.4.1 Características do *E-commerce*

O *e-commerce* pode se caracterizar pela sigla B2B (*Business to Business*) ou comércio de empresas para empresas. Por exemplo, um grande mercado que compra muitos produtos do próprio produtor, obtendo preços menores por causa do grande volume de mercadorias. O termo B2B pode também ser denominado como *extranet*.

O *e-commerce* também pode se caracterizar pela sigla B2C(*Business to Commerce*) ou comércio de empresas para consumidores, por exemplo, uma empresa que tem um *site* de *e-commerce* que oferece produtos para seus consumidores, no qual o consumidor pode efetuar a compra de apenas uma unidade. Figura 9.



Figura 9. Categoria B2C. www.americanas.com.br.

3.4.2 Logística do *E-commerce*

O *e-commerce* não funciona apenas no contexto digital. Existe a parte totalmente física do processo envolvendo várias pessoas e várias empresas. Por conseguinte, é a logística que envolve todo o processo do *e-commerce.* De nada adiantaria realizar as operações comerciais de forma digital se a entrega do produto não for rápida e se o produto não for de qualidade. O consumidor usa o *e-commerce* para, de alguma forma, ter vantagem em ralação ao meio convencional de compra. Além de objetivar um produto mais barato, pois, no *e-commerce* economizam-se custos operacionais existentes em lojas convencionais como, a comissão do funcionário, aluguel, luz, telefone, etc, o usuário tem ainda outras vantagens, como a opção de comprar na hora que achar melhor, usando o tempo necessário para a escolha do produto, obtendo garantias sobre o produto que está comprando, e até mesmo, não efetuando a compra sem nenhum constrangimento.

Quando se utiliza o *e-commerce* é necessário estabelecer processos que possam garantir a qualidade do produto a ser enviado, pois se o processo de escolha se restringir ao aspecto visual ou mesmo a uma descrição mínima de suas características, o produto pode não satisfazer as exigências reais do consumidor. Isso será o bastante para que esse consumidor nunca mais use ou compre qualquer produto da empresa, seja de forma digital ou realmente física, além de propagar o processo viral contra a empresa, verbalmente ou através de outros métodos.

Perspectivas do *e-commerce* para o Brasil apresentam claros sinais de evolução. É o que conclui Albertini(2001). Segundo o estudo, o setor movimentou US$ 2,1 bilhões em 2001. Deste valor - US$ 1,6 bilhão foi proveniente das transações entre empresas (*business-to-business* ou B-*to*-B), o que significa 1,18 por cento das cifras movimentadas nas transações tradicionais entre empresas. Os US$ 500 milhões restantes da venda para consumidores (*business-to-consumer* ou B-*to*-C) representam 0,35 por cento se comparado ao comércio convencional entre empresas e

consumidores. Estes índices podem parecer pequenos se analisados isoladamente. Considerando a evolução do setor, a tendência é de crescimento.

3.5 Usabilidade

No processo de *e-commerce*, a preocupação com a usabilidade[3] do documento digital (o *site*) é fundamental. Se as informações estiverem no *site*, mas estiverem enterradas em uma estrutura de camadas de *links* mal planejada, a venda não se tornará realidade. Assim, quem projeta o *site* deve fazê-lo não para si ou para o dono da empresa, mas para atender as características do público-alvo do produto e também para o *site* ser encontrado entre os primeiros itens de uma listagem de resposta de um sistema de busca.

Deve-se ter então como conceito que o *site* é um produto que, na grande maioria das vezes, é produzido para atender não apenas alguns poucos, mas uma grande parcela de usuários.

A má usabilidade de um *site* de *e-commerce*, pode acabar prejudicando o faturamento da empresa. Se o usuário não encontrar o que precisa em pouco tempo, ele simplesmente irá para outro *site*, pois a impaciência dos usuários é provada por estudiosos da área de usabilidade (Nielsen & Tahir 2002; Krug, 2001). A diferença entre uma grande e uma pequena empresa da *Web* estará justamente na facilidade que o material digital oferece ao seu usuário. A maior usabilidade oferecida por um documento digital pode se tornar o ponto chave de sucesso de uma empresa de *e-commerce* ou de qualquer outro foco de interesse. A má usabilidade é um dos fatores causadores da queda da bolsa de valores eletrônica NASDAQ. Muitas empresas contrataram profissionais especializados em tecnologia que dominavam todas ou várias

(3) segundo Nielsen & Tahir (2002, p. 52), o termo usabilidade refere-se a convenções que os usuários conhecem de outros *sites*. Portanto, quanto mais os *sites* executarem atividades de determinada maneira, tanto mais a usabilidade aumentará ao seguir essa convenção específica.



técnicas relacionadas aos *softwares* para a composição de um *site* de *e-commerce*. Tudo funcionava perfeitamente bem com relação ao foco máquina. Mas esses mesmos projetos que estavam funcionando perfeitamente não estavam adequados aos usuários dos produtos. Assim a equação fica fácil: se o consumidor não consegue manipular o sistema, encontrar a informação e efetuar a compra, a empresa não vende e fecha.

Marchiori (1997) revela que para adequar a informação mais importante o mais próximo possível da menor quantidade de *links* é importante usar um dos recursos de *webwriting*, que exige que se coloque o conteúdo da resposta na frente de qualquer outro conteúdo secundário e, logo depois da informação principal, se for o caso, colocar o restante da informação. Utilizando este formato, que possui o nome de pirâmide invertida, (Dotta, 2000) é possível colocar os dados mais importantes o mais próximo do usuário, ou seja: quanto maior a importância, menor deve ser a profundidade do *link*. Figura 10.

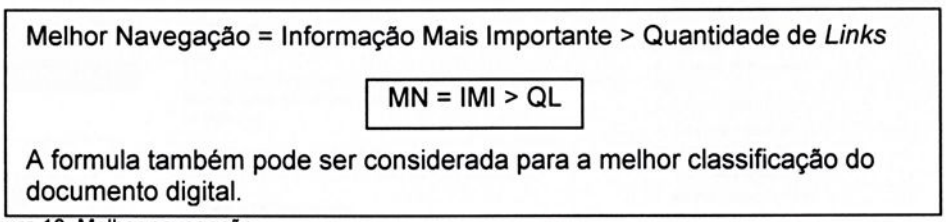


Figura 10. Melhor navegação

Um caso real pode ser averiguado em qualquer agência dos correios em que é possível passar um telegrama para alguma localidade, preenchendo no balcão da agência uma ficha com os dados necessários e escrevendo uma mensagem. Se o mesmo processo for feito pelo *site* da agência dos correios, o usuário deverá procurar o item para o telegrama além da profundidade inicial de 430 *pixels*, pois não existe como item principal da interface um *link* para telegrama, como ilustra a figura 11.

Deve ser notado que, nos primeiros 400 *pixels*[5] de profundidade, não existe nenhum item discriminando o serviço de telegrama. O item existe abaixo de 400 *pixels*. Segundo estudos de usabilidade realizados por especialistas como Nielsen[6] e Black[7], este procedimento é desaconselhado. Além do que, encontrar o item para o telegrama abaixo dos 400 *pixels* depois de visualisar vários outros itens não é uma

tarefa das mais faceis de se realizar. Deve-se levar em consideração que o item não tem um destaque apropriado tornando-o um elemento comum dentro da estrutura da informação, podendo, desta forma, deixar o usuário perdido e sem respostas rápidas.

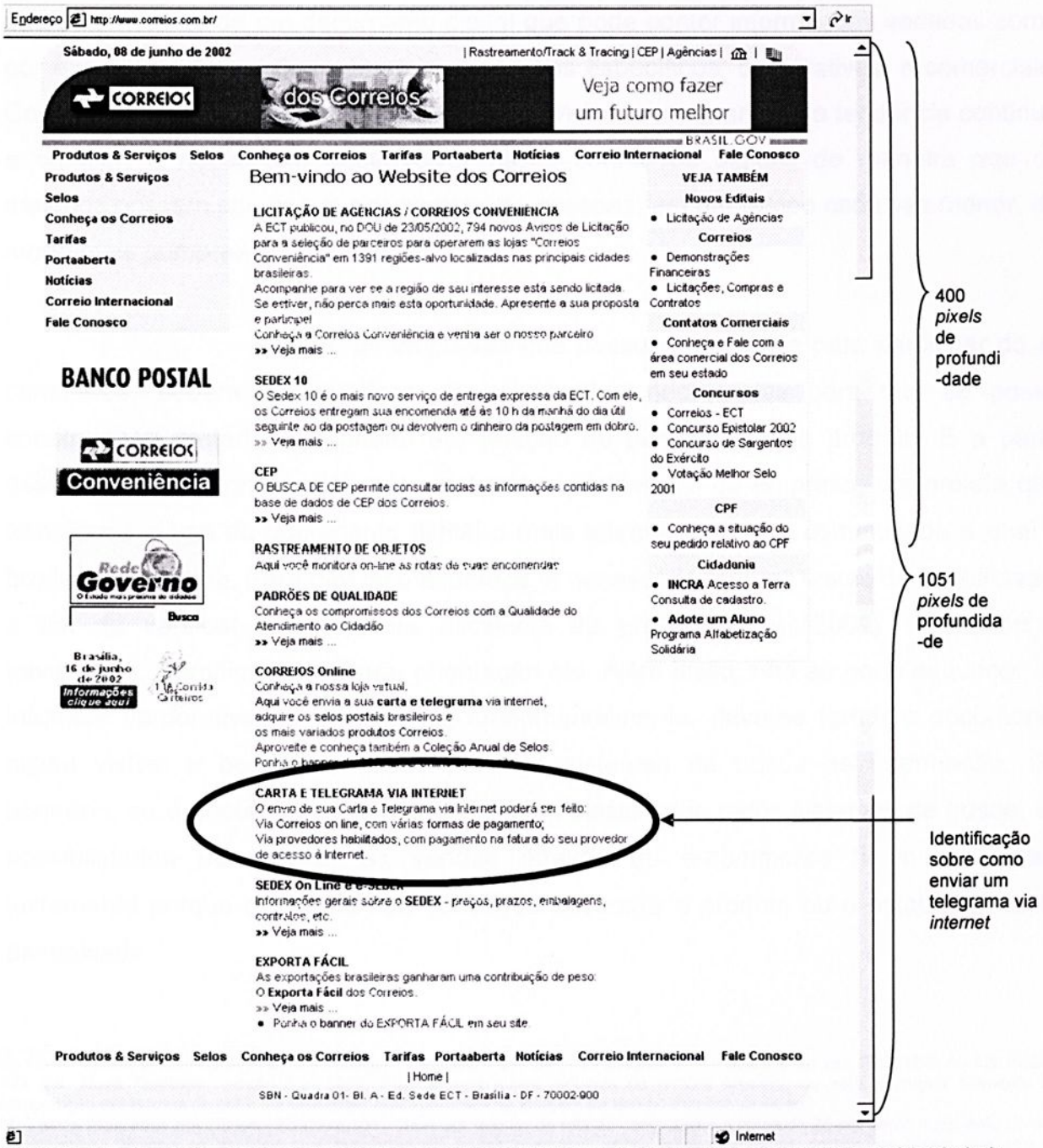


figura 11. *Site* do correio . www.correios.com.br capturado em 08/06/2002. O serviço de telegrama não é um dos principais itens relacionados na página inicial, além de estar localizado depois dos 400 *pixels* de profundidade, referencialmente é um item sem nenhum destaque.

3.6 Integração entre *E-commerce* e *Design*

Como foi descrito nos itens anteriores, há vários fatores para o desenvolvimento de um documento digital que pode conter informações variadas como por exemplo: informações pessoais, de grupos específicos, corporativos, e comerciais. Como o volume de informações contidas na *Web* é muito grande e a tendência continua a crescer, é necessário desenvolver esses conteúdos digitais de maneira que os mesmos possam atingir o maior número de pessoas, em um tempo cada vez menor, do modo mais compreensível possível.

Assim, as empresas que possuem estrutura para participar do *e-commerce*, podem disponibilizar as informações necessárias para que se possa construir um material apropriado em relação ao público-alvo do produto. E a parte destinada ao *design* irá elaborar, dentro dos parâmetros da empresa, um projeto que transforme o uso do documento digital o mais adequado para a comunidade a qual o produto se destina. Para que isso aconteça, é necessário realizar testes de usabilidade, a fim de verificar a verdadeira eficiência do projeto (Krug, 2001), facilitando a navegação, identificação, leitura, orientação etc. Além disso, não se pode esquecer da interface corporativa da empresa e, fundamentalmente, deve-se tornar o documento digital visível e bem classificado para os sistemas de busca de informação. Do contrário, se o documento digital não for bem classificado pelos sistemas de busca, as possibilidades de ampliar as vendas através do *e-commerce* ficam reduzidas, justamente porque o usuário não consegue encontrar o produto ou o estabelecimento pesquisado.

---

(5) *Pixel*, menor unidade gráfica de uma imagem matricial, e que só pode assumir uma única cor por vez. Dicionário Aurélio Século XXI. Ed. Nova Fronteira, versão 3.0. 1999. *Pixel* é uma forma reduzida de *picture element,* ou seja, o menor elemento de composição em termos de cor e luz uniformes que aparecem na imagem da tela no monitor de vídeo.

(6) Testes realizados por Nielsen, (2000, p.101). Revelam que ler da tela do computador é cerca de 25 por cento mais lento do que ler do papel. Mesmo os usuários que desconhecem essas pesquisas sobre fatores humanos geralmente dizem que sentem desconforto ao lerem texto *on-line*. Como resultado, as pessoas não querem ler muito texto das telas de computador. Portanto, deve-se escrever 50 por cento menos texto pois não se trata apenas de velocidade de leitura, mas de uma questão de conforto. As pesquisas comprovam que os usuários não gostam de rolar a tela: mais uma razão para manter a brevidade das páginas.

(7) testes realizados por Black, (1997, p.52). Indicam que 75 por cento dos usuários irão fazer a visualização até 600 *pixels* de profundidade, 20 por cento dos usuários irão fazer a visualização até o dobro de *pixels* (1200 *pixels*), e se a página possuir mais de 1200 *pixels* provavelmente o conteúdo colocado ficará sem leitores.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Por exemplo: se procurarmos a ocorrência da palavra Nokia nos sistemas de busca *Google*, Altavista, *FastSearch*, *AskJeevs*, *AOLsearch*, RadarUOL, Achei e *Yahoo*, encontramos em primeiro lugar o mesmo *site* e no Radix e *HotBot* em 2º lugar novamente o mesmo *site* (dados comprovados em 07/10/2003). Se formos contar internamente a ocorrência da palavra Nokia na página inicial, encontraremos 154 ocorrências sem, contudo, caracterizar um *spam.* O termo *spam* é utilizado para classificar e penalizar *sites* que apenas colocam de forma aleatória palavras sobre o assunto do *site,* simplesmente repetindo-as no código interno da página, com o objetivo de adicionar valor à pontuação dos mecanismos de busca na listagem de resposta de uma consulta. Desta forma, o *site* consegue, de forma artificial e sem a devida importância de conteúdo, ficar bem posicionado. Adicionalmente, se os mecanismos de busca identificarem o uso de uma ou mais, das várias técnicas de *spam* na codificação interna do *site*, o mesmo é penalizado e retirado da base de dados.

Assim, o *design* aliado ao *e-commerce* torna-se detalhe fundamental para que a informação seja transmitida da maneira mais apropriada ao público-alvo, transformando o conteúdo em uma experiência agradável. E, mesmo que não exista a efetivação da compra, deseja-se que o usuário possa voltar outras vezes para novas consultas ou até mesmo efetivar a compra. E mais: uma boa experiência navegacional, pode possibilitar bons comentários, a outros usuários numa importante contribuição na divulgação do documento digital, sem custos adicionais.

3.7 Visibilidade para Sistemas Distribuídos

Mais um fator de alta importância dentro de todo processo da criação de um documento digital é a visibilidade para os sistemas distribuídos de informação. Caso o documento desenvolvido tenha passado por todas as etapas de desenvolvimento, possua a segurança necessária, ótima navegabilidade e arquitetura de informação perfeita, mas não tenha visibilidade adequada para aparecer bem

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

classificado nos sistemas distribuídos de informação, o documento digital está perdido na *Internet* sem que seu conteúdo seja encontrado e utilizado.

Mesmo que os mecanismos de busca enviem seus agentes, os *spiders* à procura de novos documentos e os cadastrem em suas bases de dados um documento que não foi planejado adequadamente para se tornar bem visível e classificado, pode ficar escondido na enorme quantidade de informações que já estão nas bases de dados.

Certos recursos para melhorar a visibilidade dos documentos são itens internos das bases de dados, que não são revelados integralmente e também possuem suas diferenças quanto à classificação. O motivo pelo qual as bases de dados não revelam na íntegra todos os seus procedimentos para classificação das informações vem da necessidade do sistema de busca poder oferecer ou tentar oferecer uma melhor classificação da informação, sem induzir o usuário para um determinado conteúdo, o que inicialmente era o objetivo maior da *Internet*. Entretanto, é possível notar que certos sistemas de busca oferecem resultados patrocinados, induzindo o usuário para um determinado assunto, ou produto.

Outros ainda, em tentativas mais insidiosas ao usuário, apresentam produtos antes das listagens de resposta que possuem ligação com a consulta realizada. Do ponto de vista da empresa, o sistema é bem interessante. Do ponto de vista do usuário, nem tanto, pois dependendo do sistema de busca e da quantidade de informação que o usuário possui, o mesmo pode levar em consideração a informação melhor classificada, que pode estar patrocinada no sistema de busca como a sua verdade absoluta e, desta forma, esse usuário pode estar sendo prejudicado. Neste ponto, pode ser levantada a questão da confiabilidade que o usuário tem sobre a informação encontrada: qual o critério de avaliação foi adotado para que um determinado documento digital aparecesse bem classificado? O documento digital classificado com uma boa posição patrocina o sistema de busca? O que torna um documento digital bem classificado um instrumento de informação confiável? Alguns



mecanismos de busca utilizam, entre outras coisas, uma avaliação heurística, verificando quantos usuários já fizeram o uso do documento digital, como fator que contribui para a melhor classificação de um documento e, assim, melhorando a posição na listagem de classificação oferecida. Mesmo usando a avaliação heurística, os dados ainda podem não ser confiáveis, o que gera uma insegurança nas informações trazidas pelos documentos digitais.

Meghabghab(2000) revela que em uma pesquisa realizada nos sistemas de busca AltaVista, *Google*, *Yahoo* e *Northern Light* com várias entradas de *strings* (seqüências de combinações de letras números e palavras) todos os sistemas possuem alto índice de informações equívocas. O que mostra que a eficiência dos mecanismos de busca, não possui um grau de confiabilidade muito grande. Mas também não se pode atribuir toda a culpa pela falta da eficiência aos procuradores. Existe uma alta parcela de responsabilidade por parte dos desenvolvedores do *site,* quanto ao não preenchimento de informações que possam ser relevantes e que possam ser utilizadas pelos agentes, os *spiders* dos mecanismos de busca.

Contudo, existe uma alternativa com relação à visibilidade do documento digital que está nas mãos da equipe de criação. Que são alguns comandos do código HTML. Destinados a serem lidos pelos *spiders*, se usados adequadamente, podem ser uma alternativa para tornar os documentos visíveis para os mecanismos de busca. Além do mais, a maioria dos mecanismos utilizam para classificação o texto visível. O texto visível é o texto que é apresentado no corpo inicial da página no navegador. Outro elemento de classificação é a profundidade do documento. Quanto mais rolagem o documento possuir, menos pontuação será oferecida para a classificação. Contudo, isso não impede que um documento tenha rolagem, pois a menor classificação ocorre em relação ao texto visível, que não é a única forma de classificação.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Capítulo 4

Metodologia de interfaces inteligentes para mecanismos de busca

4.1 Introdução..... 67
4.2 Características de Bases de Dados..... 71
4.3 Metáfora da Classificação..... 71
4.4 Mecanismos de Busca com Banco de Dados de Índice..... 72
4.5 Mecanismos de Busca sem Banco de Dados de Índice..... 73
4.6 Método de Classificação Profundo..... 75
4.7 Método de Classificação Rasa..... 76
4.8 Exemplos de Métodos de Classificação..... 78
4.8.1 Propriedades dos mecanismos de busca..... 78
4.8.1.1 AltaVista..... 78
4.8.1.2 *Web Crawler*..... 81
4.8.1.3 *Excite*..... 82
4.8.2 Conclusão sobre as propriedades dos mecanismos de busca.. 83
4.9 *Meta-tags*..... 84
4.9.1 Importância das *Meta-tags*..... 85
4.9.2 *Tags* para cadastramento em mecanismos de busca..... 86
4.9.3 Lista de palavras-chave..... 86
4.9.4 Descrição da página ou do *site*..... 87
4.9.5 Controle do *Spider*..... 87
4.9.6 Expiração da página..... 88
4.9.7 Revisitação do *Spider*..... 89
4.9.8 Carregamento da Página..... 89
4.9.9 Atualização da Página..... 90
4.9.10 Controle do *Cache*..... 90
4.9.11 Suporte a Idiomas..... 91
4.10 *Tag Title*..... 93
4.11 *Tag Alt*..... 93

4.12 Identificação de elementos não textuais para os mecanismos de busca 95
4.13 Características de uso das *meta-tags* e *tags alt* 98
4.14 *Spam* 99
4.15 Base de dados que buscam mais itens com procura única de uma palavra 100
4.16 Base de dados Patrocinados 101
4.17 Proposta de metodologia para classificação otimizada de documentos digitais nos mecanismos de busca 102

Capítulo 4

Metodologia de interfaces inteligentes para mecanismos de busca

4.1 Introdução

Segundo Bergman(2001), há pesquisas revelando que do total de informações existentes na *Web* em média 44% são referentes a conteúdo *Web* com base em HTML[1]. O restante é atribuído, por exemplo, a linguagem XML[2], ou *Javascript*[3] e também a conteúdo multimídia como filmes, animações, músicas, além de outras formas de conteúdo, como PDF[4], dados dinâmicos, programas executáveis, planilhas de cálculos, arquivos textos de diversos formatos, etc.

unesp

Desta forma, quando os atributos de identificação são utilizados incorretamente, ou não são utilizados, as chances de uma boa classificação é eliminada e o documento digital fica escondido no provedor de acesso, sem servir ao propósito de ser encontrado para utilização e transferência de informação. Isso pode ser preocupante se o documento digital for elaborado para divulgação pessoal, corporativa ou comercial, pois não será encontrado com muita facilidade, prejudicando, assim, o usuário que pesquisa uma dada informação.

Além do mais, é importante deixar claro que seja qual for o mecanismo de busca utilizado, a classificação é realizada através da análise de texto (Silveira, 2002, p.30). Assim, qualquer elemento que não seja texto oferece dificuldade para ser rastreado e classificado nas bases de dados dos mecanismos de busca. Por esse motivo, elementos como, por exemplo, imagens, filmes, animações, sons, programas executáveis, etc, acabam sendo prejudicados em relação ao seu formato para que possam ser identificados e classificados nos mecanismos de busca. Isso porque, em sua essência não podem ser classificados simplesmente pelo material oferecido, justamente porque os métodos de classificação utilizam padrões de analise semântica, léxica e, em alguns casos, heurística e que, pela própria natureza dos outros arquivos que não possuem base textual, não podem ser analisados para classificação nas bases de dados (Kwok et al., 2001).

Assim, qualquer elemento que não seja texto não poderá ser classificado se não estiver vinculado com informações textuais referentes ao assunto do *site* e/ou do próprio elemento em questão. Figura 12.

---

(1) HTML *Hyper Text Markup Language* (Linguagem de marcação de hipertexto)
(2) XML *Extensible Markup Language*. Versão compacta da SGML (*Standard Generalized Markup Language*). Permite que os programadores e projetistas da *Web* criem *tags* personalizadas que porporcionem maior flexibilidade na organização e apresentação das informações do que é possível obter com o antigo sistema de codificação de documentos HTML.
(3) *Javascript*. Linguagem "interpretada" de criação de *scripts* desenvolvida pela *Netscape* e pela *Sun Microsystems*. É uma linguagem de programação feita para complementar as capacidades do HTML.
(4) PDF. *Portable Document Format*) - Formato em que as especificações do arquivo (negrito, itálico, tipo e tamanho de letra) são armazenadas identicamente em qualquer plataforma, sistema operacional e aplicativo.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

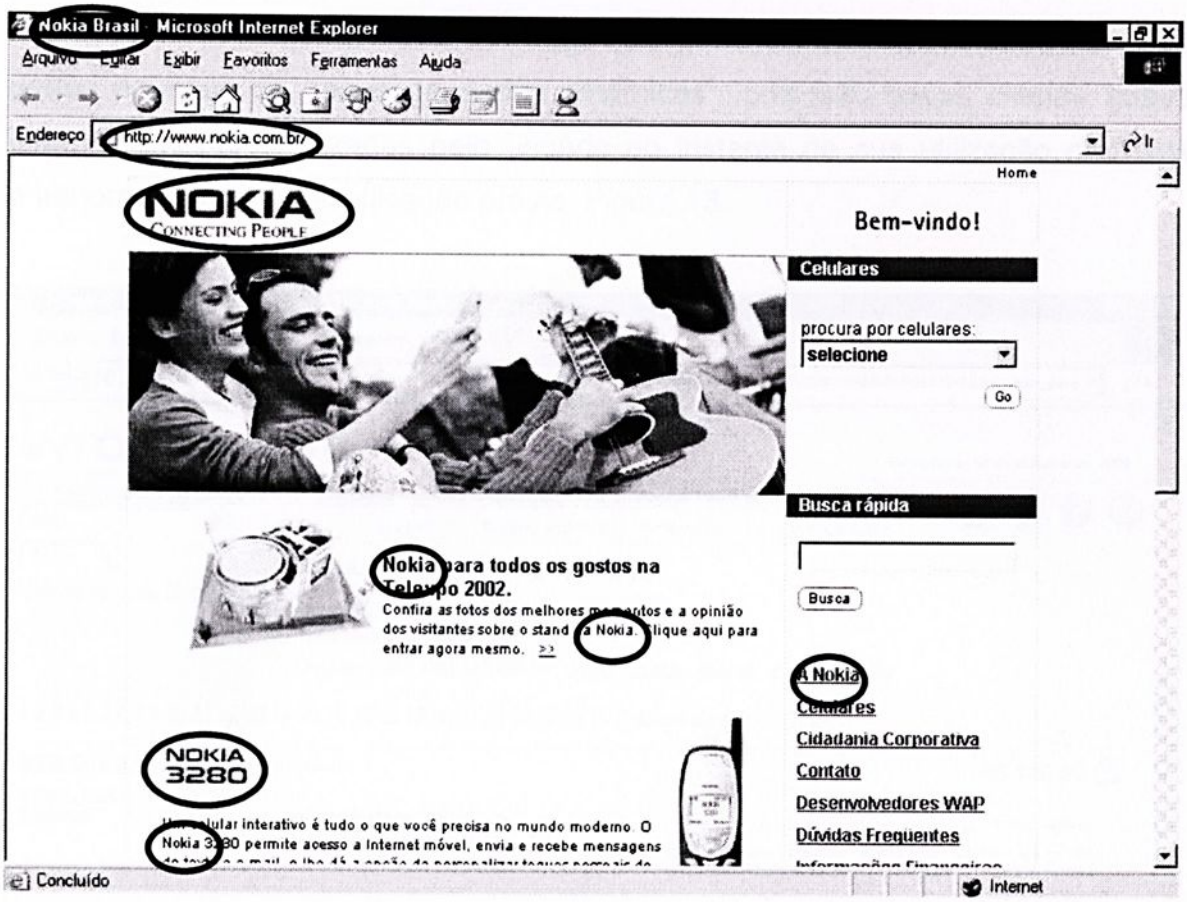


Figura 12. Exemplo de como os *spiders* podem vasculhar os documentos digitais à procura de informações para catalogação dos *sites* em seu banco de dados. www.nokia.com.br.

Na verdade, é necessário exclarecer que se o elemento não textual deixar de possuir identificação apropriada, não será motivo para que o documento digital deixe de funcionar,não sendo classificado nas bases de dados. Entretanto, será um elemento a menos que não colaborará para melhor classificação do material como um todo. E em alguns casos, esse elemento não textual pode ser a chave de acesso ou a entrada para o documento digital quando o usuário que esteja procurando alguma informação, não pelo nome do *site*, mas pelo nome de um dado objeto. Realiza-se, desta forma, a entrada no documento digital não pela página inicial, mas por um acesso paralelo, o que deve ser uma possibilidade para favorecer o usuário em relação à informação procurada (Sullivan, 2000. p8).

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Um outro item que não pode ser facilmente classificado são os dados oriundos de bases de dados dinâmicas[5], pois são bases criadas através de informações personalizadas pelo usuário no instante de sua utilização não existindo anteriormente para classificação prévia. Figura 13.

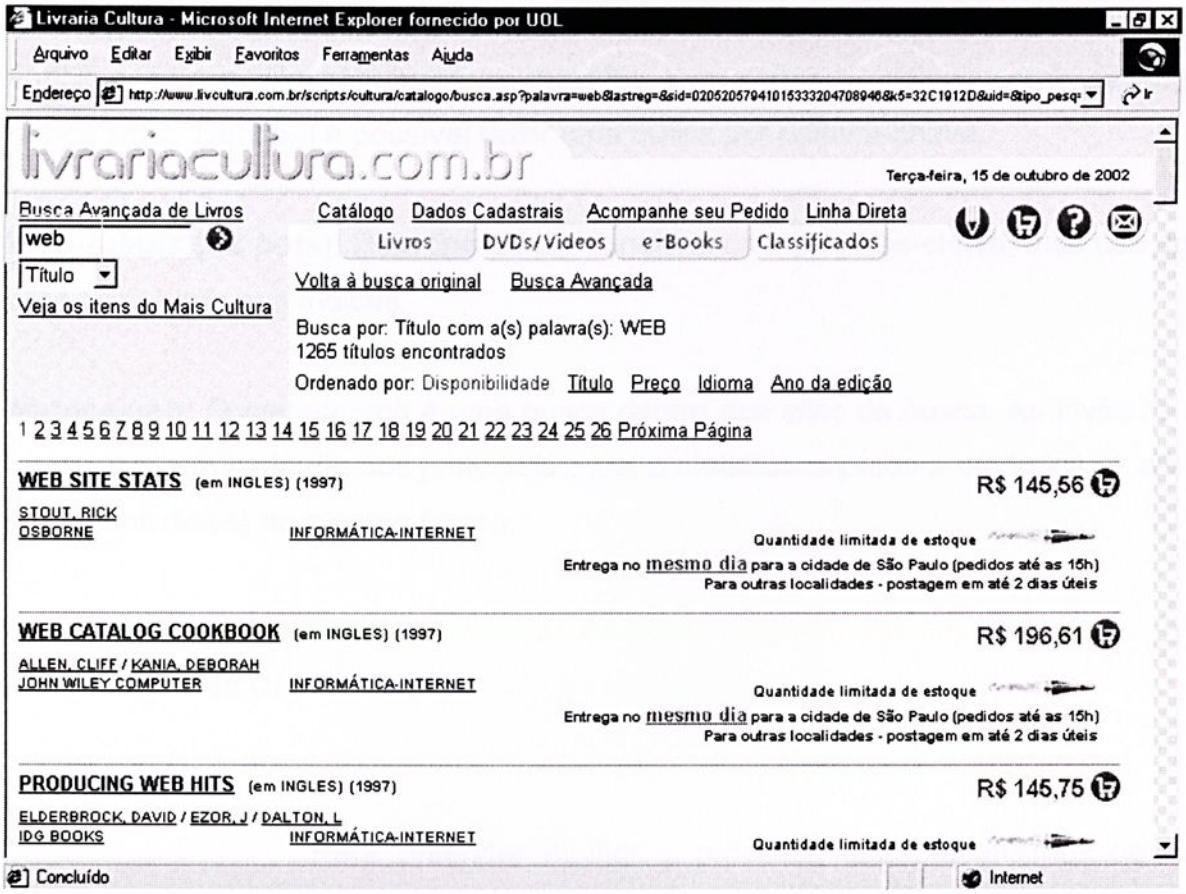


Figura 13. Exemplo de página criada no instante da consulta, extraída do *site* da Livraria Cultura em 15/10/2002. www.livrariacultura.com.br.. Esta página não poderia ser previamente encontrada, porque não existia antes da consulta do usuário. Foi criada dinamicamente utilizando um banco de dados.

Com relação às bases de dados, elas podem ser gerais e normalmente grandes, como por exemplo o *Google* e o AltaVista, ou específicas, direcionadas a determinados assuntos e em certos casos com acesso restrito, como por exemplo: o Probe (www.probe.br) e *Web of Knowledge* (http://isiknowledge.com), anteriormente conhecido com *Web of Science.*

(5) A Locução Bases de dados dinâmicas refere-se aquelas bases que são consultadas no momento de pesquisa e refletem os dados disponíveis naquele espaço de tempo da consulta.

4.2 Características de Bases de Dados

As bases de dados podem oferecer três tipos de serviços ao usuário:

Catálogo: Este tipo de base de dados funciona como uma lista de *sites* sobre vários assuntos. Um dos mais conhecidos é o *Yahoo*. Funciona como catálogo de outros *sites*. Também é possível fazer uma busca por palavra-chave.

***Web-robots* (ou *bots*):** Este tipo de *site* funciona com palavras-chave, mas usa *spiders* para alimentar seus índices.

***Metasearch*:** O *metasearch* é uma busca dentro dos *sites* de busca. Ao invés de fazer uma busca em cada um dos principais *sites,* o metabusca procura em todos os *sites* (ou em alguns deles) ao mesmo tempo.

4.3 Metáfora da Classificação

Para entender melhor o processo dos mecanismos de busca, imaginemos muitas pessoas especialistas em uma sala (a base de dados). Estas pessoas ficam esperando o mecanismo de busca fazer uma pergunta (no caso, o usuário digitando uma palavra ou frase). Tão logo o mecanismo receba a pergunta (a palavra ou frase digitada), ele a transmite para as pessoas que estão na sala. Neste momento, as pessoas na sala começam a conversar entre si para saber quem é que possui a informação que melhor possa satisfazer a pergunta. A pessoa que possuir conhecimento com mais relevância sobre o assunto, é que será a primeira alternativa na listagem oferecida. Caso não haja pessoas experientes na sala, elas farão uma votação para ver quais dentre elas possuem algum conhecimento sobre o assunto.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Nesse procedimento, a resposta à pergunta formulada pode acabar desapontando o usuário.

4.4 Mecanismos de Busca com Banco de Dados de Índice

Os elementos classificados pelos *spiders* dos mecanismos de busca são armazenados em um banco de dados de índice com o objetivo de facilitar a consulta do usuário através de palavras-chave.

Este tipo de mecanismo de busca utiliza um *software* conhecido como *spider,* responsável pela visita e indexação das páginas da *Web*. O *spider* visita cada página e retira de cada documento o conteúdo a ser indexado, armazenando-o no banco de dados de índice de forma adequada para sua eficiente recuperação. É importante salientar que no banco de dados de índice não estão armazenados os documentos na sua íntegra, (mesmo porque a capacidade de armazenamento teria que ser muito grande), mas sim algumas palavras-chave e outros dados que sejam julgados necessários como a posição da palavra no documento, sua URL (*Universal Resource Locator*) de origem, etc.

A consulta ao índice do mecanismo de busca é efetuada através do navegador do usuário, no *site* do mecanismo de busca. O mecanismo de busca analisa a consulta e retorna com as referências aos documentos, de forma ordenada, indexados por seu *spider*, de modo a satisfazer a consulta. Normalmente, além das referências, é apresentada uma pequena descrição do documento (retirada das primeiras linhas do mesmo), seu título e tamanho.

A arquitetura convencional deste tipo de mecanismo de busca está apresentada na figura 14.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

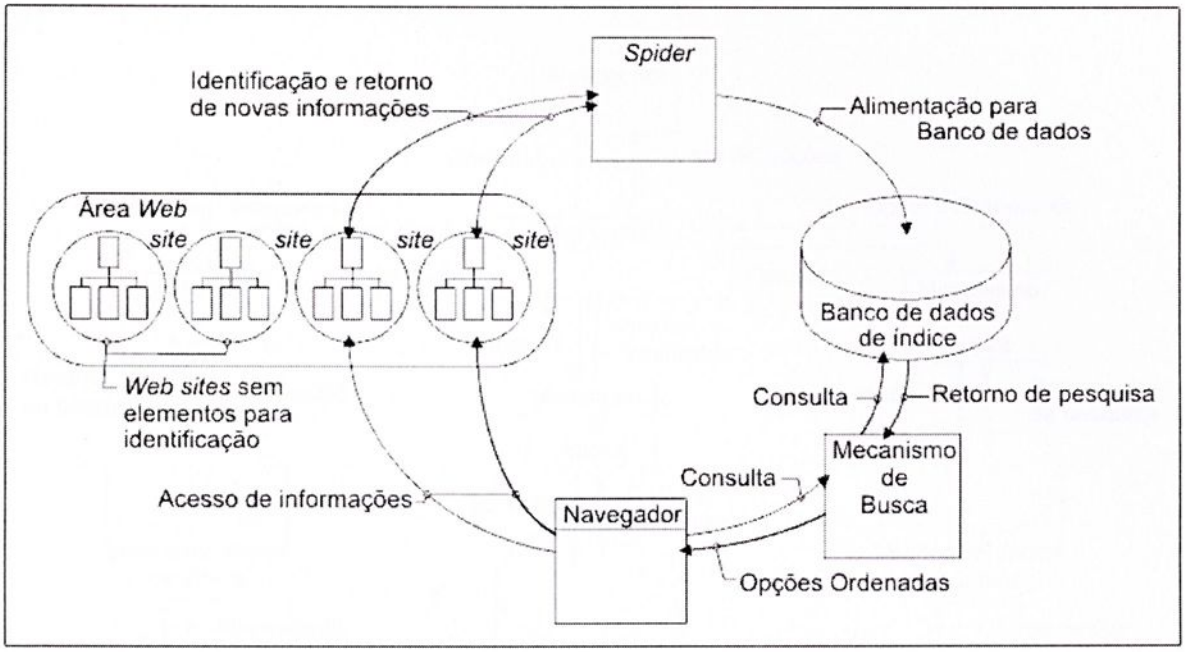


Figura 14. Arquitetura convencional de um mecanismo de busca baseado em *Spider*

4.5 Mecanismos de Busca sem Banco de Dados de Índice

Uma outra alternativa de mecanismo de busca são os meta-buscadores ou *meta-searches.* A diferença fundamental entre os mecanismos de busca com banco de dados de índice e o meta-buscador é que o serviço oferecido pelo meta-buscador não possui um banco de dados próprio. Assim, o meta-buscador acessa o banco de dados de outros mecanismos de busca, reunindo informações de vários mecanismos e selecionando, dentro de critérios próprios, opções para formar sua listagem de elementos que satisfaçam a requisição efetuada pelo usuário. Desta forma, ao invés de pesquisar em apenas um mecanismo de busca, ou individualmente fazer a mesma pesquisa usando vários mecanismos, um de cada vez, o meta-buscador faz esse trabalho, realizando a consulta do usuário em vários mecanismos de busca e retornando dados mais relevantes dentro do critério por ele estipulado. A estrutura básica do meta-buscador pode ser vista na figura 15.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

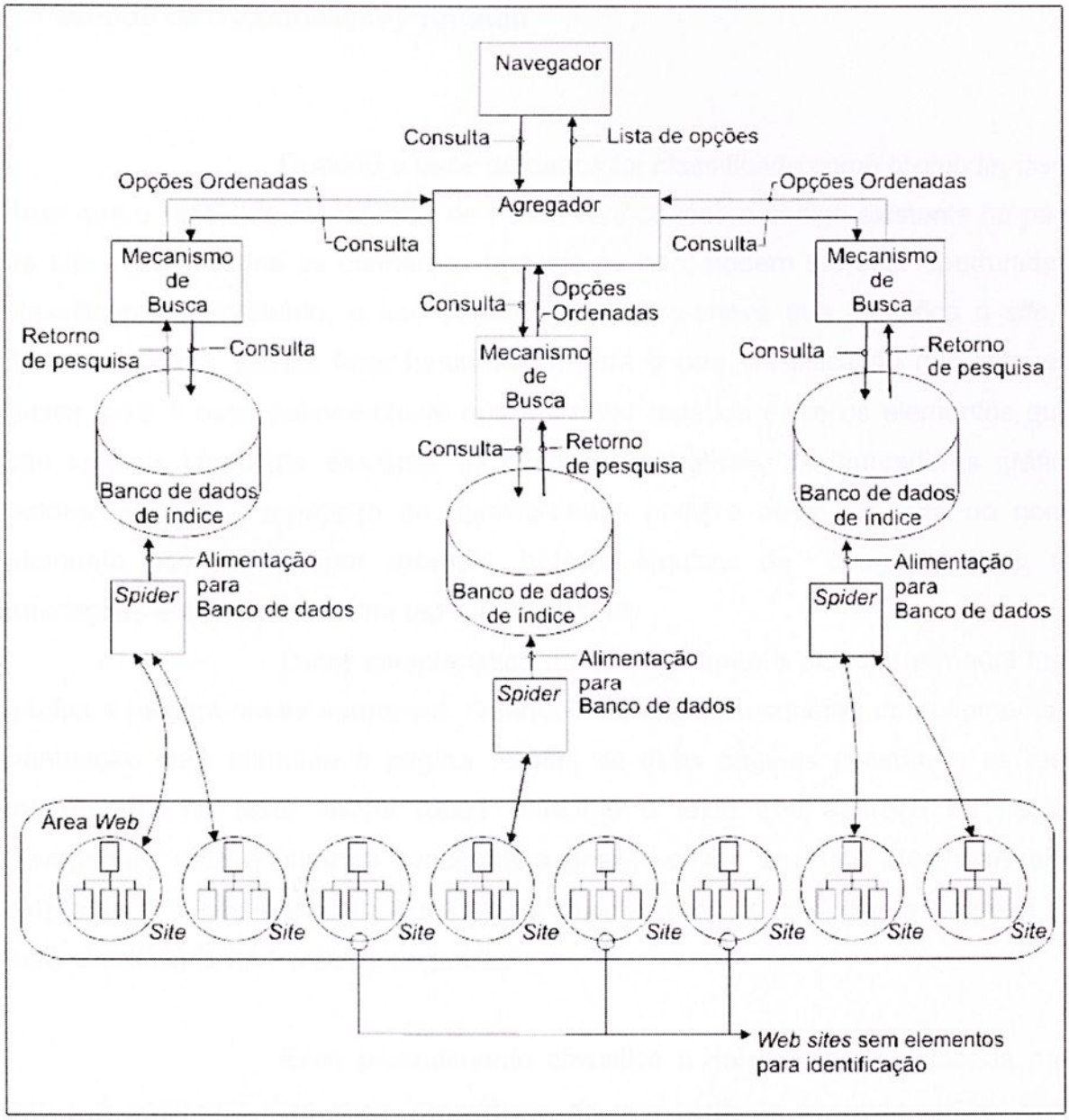


Figura 15. Arquitetura convencional de um meta-buscador.

Entre outras situações, o meta-buscador pode ser usado para fazer, de forma mais rápida do que usando um ou vários mecanismos de busca, uma consulta de preços de livros. O meta-buscador pode listar várias opções relacionadas aos *sites* que estejam vendendo um determinado livro, agilizando assim, o processo de pesquisa.

4.6 Método de Classificação Profundo

Quando a base de dados for classificada como profunda, isso quer dizer que o *spider* do mecanismo de busca verifica todo o código existente na página e no *site*. Assim, todos os elementos textuais ou não, podem oferecer oportunidade de classificação. Entretanto, o uso correto da palavra-chave que identifica o *site,* ou o material do *site,* possui fator fundamental para a boa classificação nas listagens de busca, pois é essa palavra-chave que deve ser repetida entre os elementos que não são textuais como por exemplo, as imagens ilustrativas, os marcadores gráficos,os botões, etc. Essa repetição da palavra-chave pode e deve ser feita no nome do elemento não textual (por exemplo: botões, arquivos de imagens, sons, filmes, animações etc), e também na *tag Alt.* Figura 16.

Outra característica muito importante é planejar em qual lugar da página a palavra-chave aparecerá. Quanto mais alto e à esquerda do documento, mais pontuação será atribuída à página. Assim, se duas páginas possuírem as mesmas informações no texto visível (texto visível é o texto que aparece na página no navegador), mas a primeira concentra a palavra-chave em uma área mais alta e à esquerda e a segunda coloca a palavra-chave em qualquer lugar, a primeira página será classificada na frente da segunda.

Esse procedimento classifica a palavra-chave colocada mais ao alto e à esquerda com mais importância do que texto da segunda opção. Isso não impede, por exemplo, que um documento possa oferecer rolagem de página, ou seja, um conteúdo mais extenso. No entanto, o desenvolvedor do documento digital deve estar ciente do processo de classificação (Mcluhan, 2000).

unesp

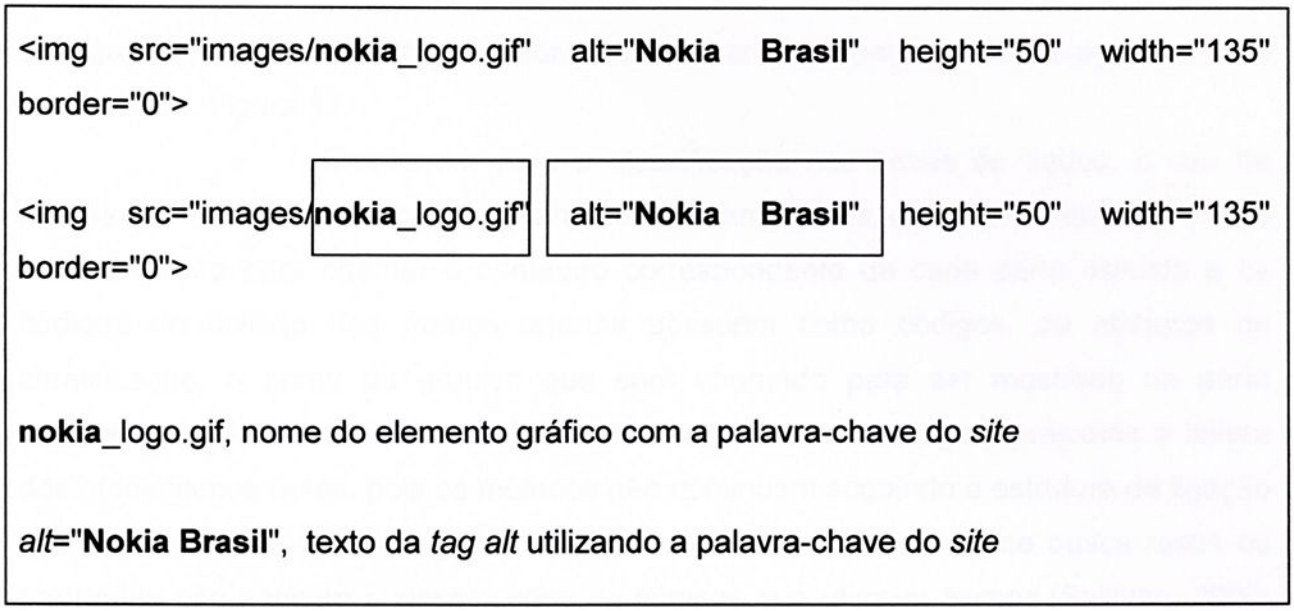


Figura 16. Exemplo de uso da palavra-chave como nome de elemento e uso na *tag alt*, extraído do *site* www.nokia.com.br

4.7 Método de Classificação Rasa

Quando a base de dados for classificada como rasa, significa que o mecanismo de busca rastreia apenas algumas partes do código da página ou do *site*. Entre outros detalhes, se uma página utiliza muitos códigos adicionais como por exemplo, *Javascript*, e se for uma quantidade muito grande, essas linhas de código adicionais podem acabar atrapalhando o rastreamento de classificação, pois os mecanismos rasos irão ler apenas uma quantidade pequena de linhas para fazer sua classificação.

Outras considerações ao método de classificação rasa podem ser feitas em relação ao uso da estrutura de *Frames*. Quando se utiliza *frames*[6] é possível visualizar várias páginas ao mesmo tempo, utilizando para isso, o mesmo espaço de tela oferecida pelo navegador. Esse procedimento pode ser útil em alguns projetos, quando por exemplo, certos *frames* podem ficar inalterados para a exibição de *links* de

(6) *Frame*. Superfície de tela eletrônica do monitor, formando um quadro de referência com unidade de características de texto e/ou imagem(s).

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

navegação, e uma outra área maior pode alterar seu conteúdo em relação ao item selecionado. Figura 17.

Entretanto, para a classificação nas bases de dados, o uso de *frames* não é recomendado, pois quando se utilizam *frames*, os mesmos dividem a área de navegação para chamar o conteúdo correspondente de cada parte dividida e os códigos de divisão dos *frames* apenas possuem como códigos, ou atributos de identificação, o nome do arquivo que será chamado para ser mostrado na parte destinada ao *frame* em questão. Essa característica de montagem prejudica a leitura dos mecanismos rasos, pois os mesmos não continuam seguindo a estrutura de ligação que os *frames* indicam, e algumas características de mecanismos de busca rasos ou profundos não aceitam a classificação de páginas que utilizam *frames* (Sullivan, 2000; Nielsen, 2000; Radfaher, 2001).



Figura 17. Exemplo de uso da estrutura de *frames*. Cada uma das partes de cor diferente é um arquivo carregado em uma estrutura principal de controle chamada *frame*. As opções do lado esquerdo podem ser escolhidas e o assunto será mostrado no *frame* principal de fundo branco.

4.8 Exemplos de Métodos de Classificação

Cada mecanismo de busca tem sua própria forma de classificação baseada em critérios individuais, objetivando oferecer um serviço de classificação de melhor qualidade ao usuário. A seguir, como exemplo, veremos algumas características referentes às bases de dados AltaVista, *WebCrawler* e *Excite* que devem ser utilizadas para obter um melhor resultado na classificação, juntamente com outras técnicas de classificação inseridas no documento digital, que serão comentadas neste capítulo.

4.8.1 Propriedades dos mecanismos de busca

4.8.1.1 AltaVista

Tamanho	1 Bilhão de URLs (aprox)
Classe de *Spider*	Profundo
Suporte a *meta-tag*	sim
suporte para *frame*	sim
suporte a mapa de imagens	sim
suporte a *tag alt*	sim
comentários HTML	não
pesquisa por URL	sim
Diretórios internos	sim
envio de URL	sim

Tabela 3. Propriedades de classificação do AltaVista

O mecanismo de busca do AltaVista é um mecanismo de busca profundo. Isto significa que o AltaVista indexa todos os dados existentes na página. Desta forma, se o usuário enviar o URL da página para o AltaVista, os *spiders* providenciarão a indexação de todo o restante do *site: links*, imagens, *meta-tags* e

outros, com exceção das páginas criadas dinamicamente. O AltaVista indexará a maioria do código HTML, possibilitando a inclusão dos comentários da *Tag Alt* em textos que identificam imagens, títulos, URLs, nomes de diretórios, textos visíveis, mapas de imagens e *meta-tags*. Os comentários feitos em HTML[7] são ignorados pelo serviço do AltaVista.

O mecanismo de busca do AltaVista utiliza as *meta-tags Keywords* e *description*. Os *Keywords* (palavras-chave) ajudam a aumentar a relevância da página em uma procura de um assunto particular, enquanto que a *meta-tag description* oferece, como resultado de uma pesquisa, um resumo na tela de resposta dos mecanismos de busca, além de contribuir para a relevância da pesquisa. Há ainda várias coisas importantes que precisam ser descritas. O título da lista de resposta do procurador é obtido diretamente da *tag title* do código HTML. Figura 18. Se for omitido o título, o AltaVista substitui por um "sem título" no lugar.

A descrição para aparecer na listagem de classificação vem diretamente da *meta- tag* de descrição. (*meta name description,* Figura 18) O AltaVista data as entradas, tornando a localização mais eficiente para a procura de atualizações de documentos digitais nos índices.

Se for omitida, a *meta-tag-description*, o AltaVista usará as primeiras 150 palavras do texto visível como resumo para a lista (texto visível é o texto que aparece na página no navegador).

Característica de ordenação do AltaVista

1- Conteúdo próximo do topo da página conta mais que conteúdo no final. Em particular, o título HTML e as primeiras linhas do texto são as mais importantes partes da página. Se as palavras e as frases que forem digitadas no mecanismo de busca forem idênticas às do título da página e/ou idênticas às primeiras linhas do texto das páginas, as chances desta página ser mostrada na frente de outras é muito maior que de outras nos resultados de uma procura.

(7) Comentário HTML, é uma descrição colocada no código interno usando uma *tag* de comentário. Por exemplo:
<!-- comentário HTML -->.

unesp

<html>
<head>

<title>**Nokia** Brasil</title>

<META NAME = "description" CONTENT ="A **Nokia** é lider mundial no fornecimento de aparelhos celulares e também no fronecimnto de infra-estrutura de redes fixas e celulares incluido serviços a cliente.">

<META NAME ="keywords" CONTENT="**nokia**, **Nokia**, **NOKIA**, telefones móveis, celular, celulares, Celulares, Celular, telecomunicação, telecomunicações, redes sem fios, rede sem fio, rede fixa, redes fixas, datacom, GSM, términais de multimedia, terminal de multimedia, monitor, monitores, aparelho de mão, apararelhos de mãos, aparelhos de mão, atendimentos ao consumidor, atendimento ao consumidor, atendimento aos consumidores, boletins de imprensa, boletim de imprensa, informação financeira, informações financeiras, intercâmbio de estudante, intercambio de estudantes, intercambios de estudantes, posição aberta, posições abertas, oportunidades de emprego, oportunidade de emprego, oportunidades de empregos, oportunidade de carreira na **Nokia**, oportunidades de carreira na **Nokia** " >

<!-- *end meta-information* -->

</head>

Figura18. Código interno de um documento digital, note a *tag tilte* na terceira linha.

2- O conteúdo é o que se pode ler na página. É por isso que as páginas, em sua maioria, possuem textos ao invés de imagens, ou pelo menos deveriam possuí-los. Na verdade, o texto é o que alimenta os mecanismos de busca, e, quando bem empregado na estrutura da página, pode fazer a diferença na colocação da resposta. A falta de conteúdo, provavelmente, é a razão principal de certos *sites* não encontrarem uma posição razoável. Deve ficar claro que esse não é o único método utilizado pelo AltaVista para ordenar as informações, mas as condições de procura e a inclusão de tais condições dentro de um *web site* são consideradas de importância vital no AltaVista.

As características de classificação do AltaVista não são as mesmas para outros sistemas de busca. Se o *site* for projetado com os procuradores

unesp

em mente, usando a combinação de *spider* e HTML amigavelmente e com conteúdo razoável, certamente tendem a fazer as páginas aparecer em posições mais altas e privilegiadas. A inclusão de *meta-tags* só ajuda ligeiramente na classificação do mecanismo de busca do AltaVista.

4.8.1.2 *WebCrawler*

Tamanho	Menos de 10 Milhões URLs (aprox)
Classe de *Spider*	raso
Suporte a *meta-tag*	sim
suporte para *frame*	não
suporte a mapa de imagens	não
suporte a *tag alt*	Não
comentários HTML	Não
pesquisa por URL	sim
Diretórios internos	sim
envio de URL	sim

Tabela 4. Propriedades de classificação do *WebCrawler*

O *WebCrawler* é uma das máquinas de procura mais antigas e uma das menores. Embora *WebCrawler* e o *Excite* tenham-se tornado uma só empresa, o *WebCrawler* conseguiu manter um sistema próprio.

O *spider* do *WebCrawler* é um mecanismo de procura rasa. Indexa todo o texto em uma página, não provê apoio para *frames* ou *imagemaps*[8]. Além disso, ignora comentários e texto em *tag alt*.

O *WebCrawler* foi o primeiro sistema a implementar uma rotina de inteligência artificial para gerar um resumo para a entrada de palavras-chave. Notaram-se porém alguns problemas ao se usar este método. Decidiram, então oferecer apoio

(8) *Imagemaps.* Imagem com vários *hyperlinks* que levam a destinos diferentes se clicados.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

pela *meta-tag* de descrição. Assim, se o *site* omitir as *meta-tags* de descrição, o *WebCrawler* invocará a rotina de IA (Inteligência Artificial[9]) para determinar um resumo para o *site*. Desta forma, para tirar melhor proveito na classificação do mecanismo de busca, os itens abaixo devem ser seguidos:

1- Para o título deve ser usado uma palavra ou frase exclusivamente descritiva da página ou *site*. O algoritmo de *indexing/relevance* do *WebCrawler* dá ligeiramente mais peso para títulos do que para texto visível.

2- A página principal deve descrever a extensão mais exata possível sobre o *site*. Essa descrição não deve ser longa e exaustiva, mas que componha um texto com palavras importantes, sem sacrificar o *design/layout* do *site*. Isso ajudará no posicionamento do *site* à frente de outros documentos que não se preocuparam com esses detalhes.

4.8.1.3 *Excite*

Tamanho	250 Milhões de URLs (aprox)
Classe de *Spider*	raso
Suporte a *meta-tag*	parcial
suporte para *frame*	não
suporte a mapa de imagens	não
suporte a *tag alt*	não
comentários html	não
pesquisa por url	não
Diretórios internos	sim
envio de URL	sim

Tabela 5. Propriedades de classificação do *Excite*

O *Excite* indexa todos os textos de página visível, mas ignora texto de *tag alt*, comentários e a *meta-tag-keyword*. Adicionalmente, o *Excite* não trabalha com *frames* nem com *imagemaps*. Em um curto período, o *Excite* fez uma mudança

(9) A rotina de I.A. constitui um conjunto de técnicas inteligentes próximo do pensamento humano, que são inseridas em autômatos (ou agentes, como os *spiders*).



radical na sua política para uso de *meta-tags*. Em sua versão original, o *Excite* ignorava a *meta-tag-keyword* e a *meta-tag-description*. Mas em vez de confiar no texto de abertura dentro de uma página, o *Excite* usava uma rotina de inteligência artificial para entender o assunto da página, e desta análise, o mecanismo de busca preenchia o resumo de resultados da procura.

Desnecessário é dizer que isto conduziu a muitas reclamações e muitos *webmasters* ficaram insatisfeitos com o serviço de classificação do *Excite.* Por isso, o método de classificação mudou. O *Excite* agora indexa por *meta-tag-description*. Se a página inclui as *tags* de identificação necessárias, o mecanismo de busca do *Excite* as usará para resumir o *web site*. Se a página omitir as *tags* de descrição, o *Excite* acionará sua rotina de inteligência artificial para analisar e criar um resumo para o *site.* Isso não significa que o *site* possa ter classificação relevante no resultado de uma pesquisa.

4.8.2 Conclusão sobre as propriedades dos mecanismos de busca

Nos itens anteriores (4.8.1.1, 4.8.1.2 e 4.8.1.3) foram descritas algumas particularidades de apenas três mecanismos de busca dentro de um grande universo. Cada um dos mecanismos apresenta suas características próprias com relação à forma de classificação de sua base de dados. Entretanto, todos utilizam elementos semelhantes entre eles para realizarem a sua classificação. Entre os elementos comuns podem ser usados a *tag title*, as *meta-tags-description* e *keyword* o texto visível e a denominação da *tag alt* para elementos não textuais. Assim, a diferença entre os mecanismos de busca estará na forma como seu programa de identificação trabalha: qual é a quantidade de caracteres lidos em cada *tag,* quais são as *tags* de classificação do mecanismo e se o mecanismo é de rastreamento profundo ou raso. Apesar de os itens de classificação de cada mecanismo serem muito variados,



a grande parte dos mecanismos de busca classifica basicamente com os mesmos atributos, diferenciando apenas o algoritmo usado para classificação.

Desta forma, é possível conseguir uma boa classificação do documento digital em relação ao mecanismo de busca através do conhecimento da estrutura da base que se pretende enviar o documento. Se os atributos de classificação do documento digital forem usados de forma correta, poderão acarretar boa classificação em outros mecanismos, levando-se em consideração que os atributos de classificação que se pode inserir nos documentos digitais são sempre os mesmos.

4.9 *Meta-tags*

Resultado de um esforço para criar uma linguagem para definição da estrutura de um documento que pudesse ser conectado a outros a partir dos conceitos de hipertexto e hipermídia[10], foi desenvolvida no CERN[11], no início da década de 90, por Tim Berners-Lee[12] juntamente com Robert Cailliau, a linguagem HTML e o conceito de *World Wide Web*. Esse feito possibilitou a catalogação das páginas que compunham um *site* para busca posterior (Furgeri, 2001, p33).

Para isso, foi especificada uma série de *tags*[13] chamadas de *meta-tags*, pois são *tags* de informação sobre a própria página onde estão inseridas. Essas *meta tags* formam um dos tópicos talvez mais incompreendidos e interessantes dentre os recursos da linguagem HTML.

---

(10) Hipermídia. Conjunto de informações apresentadas na forma de textos, gráficos, sons, vídeos e outros tipos de dados, e organizadas segundo o modelo associativo e de remissões, próprio do hipertexto. Hipertexto: conjunto de dados textuais, computadorizados num suporte eletrônico que podem ser lidos de diferentes maneiras através de conexões (*links*). Hipermídia é um hipertexto multimídia (texto, imagem e sons) (Laufer, R. & Scavelta, D.1992, p: 6-9). Pierre Levy (1997, p28) define hipertexto como um conjunto de nós ligados por conexões. Os nós podem ser palavras, páginas imagens, gráficos, seqüências sonoras ou mesmo outros hipertextos. Os itens de informação não são ligados linearmente mas são estruturas de conexões em estrela.
(11) Organização Européia para Pesquisas Nucleares, localizada entre a França e a Suíça.
(12) O inventor do termo hipertexto foi Ted Nelson, que exprime o sonho de manter os pensamentos em estrutura multidimensional, mas foi Tim Bernes-Lee o criador da WWW.
(13) *Tag*, termo usado para identificar os comandos da linguagem HTML.(*Hyper Text Markup Language*)



De acordo com a especificação oficial da HTML 4.01 (a versão mais recente) que pode ser obtida no *site* do W3 *Consortium* (http://www.w3c.org), as *meta-tags* devem ser inseridas na seção iniciada pela *tag* <*head*>. Não faz diferença se serão escritas antes ou depois do título da página. As *meta-tags* não são *tags* de container e, por isso, não possuem o respectivo fechamento.

Exemplo de *tag* com, e sem container:

Tag com container	*Tag* sem container
** ** Música popular ** **	**<META** name="distribution" content="global"**>**

4.9.1 Importância das *Meta-tags*

Existem vários recursos disponíveis com o uso das *meta-tags*, mas os principais relacionam-se com os *sites* de busca baseados em *search engines* (ou mecanismos de busca) como o AltaVista, o RadarUOL, o *Google* o *Lycos* e vários outros. Basicamente, o que Berners Lee pretendeu foi definir uma forma de criar um banco de dados com endereços (URLs) de páginas que pudesse ser acessado através de alguma espécie de *software* de busca.

No entanto, algumas *meta-tags* prestam-se quase que somente para fins documentacionais, como aquelas onde se declara o nome do autor da página ou o *software* usado na criação dessa página. Existe também uma *meta-tag* para declarar informações de *copyright* (direitos autorais).

Algumas outras *meta-tags* influenciam na forma como a página será lida, se será atualizada e com que periodicidade, também influindo na forma como a página é ou não armazenada no *cache*[14] do navegador.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

4.9.2 *Tags* para cadastramento em mecanismos de busca

As *meta-tags* a seguir têm efeito somente no cadastramento de seus *sites* em *sites* de busca que se utilizam de mecanismos de busca, como o AltaVista, o *Google*, o *WebCrawler* e o RadarUOL, entre outros.

Sites como o *Yahoo* não se beneficiam dessas *meta-tags,* pois são diretórios nos quais o cadastramento das informações usadas na busca é feito manualmente.

Nos mecanismos de busca existem *softwares* chamados *spiders*, que fazem a busca e a catalogação das informações das páginas que compõem um *site,* desde que a URL deste *site* tenha sido submetida ao cadastramento no sistema de busca.

Cada mecanismo de busca possui métodos próprios para cadastrar as informações das páginas do *site*, mas a maioria deles costuma, catalogar, pelo menos, o conteúdo da *tag* <*title*> e as *meta-tags-Description* e *Keywords*. Alguns mecanismos de busca podem também catalogar os textos alternativos de imagens (atributo *alt* da *tag* <*img*>).

4.9.3 Lista de palavras-chave

<*META NAME="Keywords" CONTENT*="palavras_separadas_por_vírgula">

Essa *meta* permitirá que o *spider* catalogue no banco de dados do *site* de busca uma lista de palavras que o criador do *site* julga serem relevantes em uma busca que seja efetuada por um usuário da *Web*. Não há, a princípio, um limite para o

(14) *Cache*, Dispositivo de memória, de capacidade reduzida e alta velocidade, que funciona associado a um dispositivo de armazenamento de grande capacidade, porém mais lento, mantendo cópia temporária de dados acessados com mais freqüência ou mais recentemente, com o objetivo de agilizar o processamento de tais dados.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

número de palavras, mas cada mecanismo de busca estabelece um tamanho máximo para a lista de palavras-chave, de modo que o excedente é descartado.

4.9.4 Descrição da página ou do *site*

<*META NAME="Description" CONTENT*="descrição_da_página_ou_site">

Permitirá que o *spider* catalogue, juntamente com as palavras-chave, uma breve descrição da página ou do *site*, que também será levada em conta no momento de uma busca efetuada por um usuário. Assim como existe um limite definido pelo mecanismo de busca para o tamanho da lista de palavras-chave, existe também para o tamanho da descrição e o que exceder a esse limite é descartado.

4.9.5 Controle do *Spider*

Na maior parte das vezes, todas as páginas, a partir da indicada durante a submissão do *site* ao mecanismo de busca, serão indexadas. Isso é feito a partir dos *links* localizados na página submetida. No entanto, pode-se querer evitar que determinadas páginas ou até mesmas seções inteiras de um *site* sejam catalogadas como páginas de acesso restrito ou páginas temporárias. Para isso existem os *meta-Robots*.

<META NAME="Robots" CONTENT="all | index | noindex | follow | nofollow">

all - é o padrão que faz com que a página onde a *meta-tag* está inserida seja indexada, bem como todos os *links* sejam seguidos pelo *spider*;



index - faz com que a página onde a *meta-tag* está inserida seja indexada (é o comportamento *default*);

noindex - faz com que a página onde a *meta-tag* está inserida não seja indexada;

follow - faz com que os *links,* a partir da página onde a *meta-tag* está inserida, sejam pesquisados para indexação pelo *spider* (é o comportamento *default*);

nofollow - faz com que os *links,* a partir da página onde a *meta-tag* está inserida, não sejam pesquisados para indexação pelo *spiders*;

none - faz com que a página não seja indexada, bem como seus *links* não sejam seguidos pelo *spider* do mecanismo de busca.

4.9.6 Expiração da página

<META HTTP-EQUIV="Expires" CONTENT="data_no_formato_RFC850">

Define uma data para expiração da página, após a qual, a mesma pode ser removida do banco de dados do mecanismo de busca. Dependendo do navegador, essa *meta-tag* pode também definir uma data para expiração da página no *cache*.

Por exemplo, se a data de expiração for 31/05/2003, a *meta-Expires* deve ser assim especificada:

<ME*TA HTTP-EQUIV="Expires" CONTENT="Wed, 31 May 2003 3:00:00 GMT*">

4.9.7 Revisitação do *Spider*

<META NAME="Revisit-After" CONTENT="30 Days">

Essa *meta-tag* programa o *spider* do mecanismo de busca para que refaça uma visita a essa página dentro de 30 dias (podem ser especificados outros períodos de tempo). É muito útil caso o conteúdo da página seja modificado dentro desse período, fazendo com que o *spider* catalogue novamente as informações das *meta-tags*.

4.9.8 Carregamento da Página

Sempre que um arquivo qualquer (como uma página da *Web*, por exemplo) é transportado através do uso do protocolo HTTP (*Hyper Text Transfer Protocol*), no início de cada pacote aparece o que é chamado de "cabeçalho HTTP". Esse possui informações que definem como esse pacote será tratado pelo servidor e pelo cliente HTTP (conhecido também como navegador). É portanto o protocolo utilizado para transferência de páginas de hipertexto ou outros documentos na *Web*.

A RFC2616[15] define todas as informações (diretivas) que podem ser inseridas no cabeçalho HTTP, muitas delas, podendo ser inseridas na *meta-tag* "*http-equiv*", que será mostrada a seguir.

(15)Informações sobre a RFC2616 podem ser obtidas no endereço: http://www.ietf.org/rfc/rfc2616.txt, acessado em 11/07/2003.

unesp

4.9.9 Atualização da Página

<META HTTP-EQUIV="Refresh" CONTENT="120">

Essa diretiva faz com que a página seja recarregada automaticamente, caso o usuário fique nessa página durante 120 segundos. Esse recurso pode ser usado, por exemplo, em páginas que requerem atualização constante, como nos *sites* de notícias.

<META HTTP-EQUIV="Refresh" CONTENT="30;URL='endereco'">

Esse é outro uso de *Refresh*: provocar o que se chama de redirecionamento do navegador. Após um determinado tempo em segundos, o navegador automaticamente é levado a carregar a página especificada, cujo endereço está especificado no parâmetro URL. Isso é muito usado em páginas, cujos endereços foram alterados e o usuário vê uma mensagem explicativa antes que a nova página seja carregada automaticamente.

4.9.10 Controle do *Cache*

Pelo menos, no que diz respeito aos principais navegadores utilizados atualmente (*Internet Explorer*, *Netscape Navigator*, Opera, etc), todos os arquivos carregados são armazenados em algum diretório com o propósito de *cache*. A finalidade é permitir um recarregamento mais rápido das páginas, reaproveitando itens que não tenham sido alterados, como imagens por exemplo.

Normalmente, o navegador está configurado para, uma vez tendo carregado uma página, verificar automaticamente, a cada nova tentativa de acesso, se



houve alterações e somente fazer a requisição ao servidor em caso positivo. No entanto, essa configuração pode ser alterada pelo usuário e, assim, corre-se o risco do navegador nunca verificar se a página foi alterada, o que pode causar prejuízos, já que o usuário pode pensar que a página nunca é atualizada. Por isso, pode-se controlar o armazenamento das páginas no *cache* do navegador.

<META HTTP-EQUIV="Cache-Control" CONTENT="no-cache">

A diretiva *Cache-Control* com o valor "*no-cache*" faz com que o controle de *cache* do navegador não armazene a página no diretório de *cache*. Mas, como essa é uma implementação na versão 1.1 do protocolo HTTP, alguns navegadores mais antigos poderão não interpretar essa diretiva. Assim, outra *meta-tag* pode (e deve) ser usada.

<META HTTP-EQUIV="Pragma" CONTENT="no-cache">

A diretiva *pragma* ainda existe na versão 1.1 do protocolo HTTP para manter compatibilidade retroativa com a versão 1.0 desse mesmo protocolo, com a qual alguns navegadores mais antigos ainda trabalham.

4.9.11 Suporte a Idiomas

<META HTTP-EQUIV="Content-Language" CONTENT="br">

Define para o navegador (e em muitos casos para o *spider* do mecanismo de busca) qual a língua utilizada para produção do conteúdo da página. A importância disso é permitir que o navegador acione o suporte à língua se for necessário (quando se cria conteúdo em línguas orientais, por exemplo). O código da língua está definido na norma ISO[16] 3166.

Também pode ser usada a diretiva "*Content-Type*", que define o tipo do conteúdo carregado pelo navegador, bem como o conjunto de caracteres utilizado. Caso o navegador ainda não tenha suporte para esse idioma, o mesmo pode ser instalado no momento da utilização.

<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=Shift_JIS">

O exemplo acima informa ao navegador que deve ser ativado o suporte a um dos conjuntos de caracteres do idioma japonês. Caso o navegador não tenha esse suporte, ou ele será instalado no momento do carregamento da página, ou a mesma será ilegível.

<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=ISO-8859-8">

O exemplo anterior informa ao navegador que deve ser ativado o suporte ao conjunto de caracteres do idioma hebraico.
O código equivalente para páginas com conteúdo ocidental é:

<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=ISO-8859-1">

Pode-se notar que há uma série de controles e recursos disponíveis com o uso das *meta-tags*. Calcula-se que apenas 20% das páginas na *Web* as usem e esse pode ser um fator diferencial, principalmente, no que diz respeito à catalogação em mecanismos de busca, já que alguns *spiders* podem, simplesmente, ignorar o conteúdo da página que não tem essas *meta-tags* definidas.

Também é possível controlar a relação do *cache* do navegador com a página, o que é muito útil em páginas atualizadas com muita freqüência e o suporte a idiomas também deve ser considerado, já que a criação de *sites* multilinguais é uma possibilidade.

(16) ISO. (*International Standards Organization*), Organização Internacional definidora de normas para as mais variadas áreas de produtos e serviços.



4.10 *Tag Title*

A *tag title* deve ser utilizada para identificar sem sombra de dúvida o *site* ou o material que será encontrado no documento digital. No código HTML, a *tag title* é uma das primeiras *tags* a serem lidas pelos mecanismo de busca. Para o título da página ou *site* pode ser usado até 60 caracteres. Para os mecanismos de busca, a *tag title* tem grande importância, pois será analisada e comparada com a palavra procurada em uma pesquisa. Alguns títulos usam caracteres especiais com seqüências de sinais de maior ou outros símbolos antes da palavra ou frase de título. Essa técnica de decoração parece inofensiva, mas pode comprometer a classificação do documento digital na pontuação da listagem de resposta do mecanismo de busca.

Suponhamos que o título de um documento digital seja "Mundo", e de outro seja ">>Mundo", e que as duas alternativas possuíssem o mesmo conteúdo. A segunda alternativa de título seria classificada depois da primeira alternativa. Isso acontece porque a comparação dos mecanismos de busca se realiza com a ocorrência mais idêntica e de maior repetição no *site*. Assim, se o usuário digitar "mundo" em um mecanismo de busca, o mesmo irá identificar a primeira alternativa como o resultado mais próximo.

4.11 *Tag Alt*

A *tag alt* é outro recurso que possui alto poder de aumentar a classificação de um documento, desde que seja utilizada de forma adequada. E a forma adequada de usar a *tag alt* é primeiramente identificar qual é a palavra-chave do *site.* O segundo passo é identificar o que faz ou o que é o item, e reunir os dois itens no texto da *tag alt.* É possível também identificar apenas o que é ou o que faz o elemento digital. Mas desta forma, a classificação não terá um vínculo com a palavra-chave selecionada.



Entretanto, a *tag alt* não se destina a ser utilizada apenas para tornar um documento melhor classificado. Assim, é possível operacionalizar com a *tag alt* o conceito de usabilidade voltado ao usuário, no qual é possível transmitir para o usuário o que será encontrado do outro lado do *link*, sem que o mesmo tenha que efetuá-lo. Para que esse recurso funcione adequadamente, é necessário descrever com poucas e adequadas palavras o que se poderá encontrar do outro lado do *link* utilizando a *tag alt*. Figura 19.



Figura 19. Exemplo fictício de como a *tag alt* apresenta seu conteúdo para o usuário. O pequeno retângulo claro com o texto "Projeto Tamar – Produtos" aparece quando o *mouse* fica mais de um segundo sob o botão produtos.

Outro fator de usabilidade é a possibilidade do conteúdo ser lido pelo navegador transformando o conteúdo da *tag alt* em som e, assim, permitir o uso de usuários com deficiência visual.

E por último, em alguns casos, determinados usuários desconsideram a navegação com imagens, para ganhar tempo no descarregamento dos elementos do documento digital. Desse modo, com o uso da *tag alt*, os lugares que usam imagens apresentam a área da imagem e o texto explicativo para que, mesmo sem a imagem, o usuário consiga navegar no *site*. Além disso, alguns usuários por preferência utilizam preferentemente navegadores totalmente textuais e que não apresentam nenhum tipo de elemento gráfico como por exemplo, o navegador *Lynxs*. Desta forma, para que se possa utilizar este tipo de navegador usando elementos

gráficos com itens navegacionais, é de extrema importância o uso da *tag alt* como identificador. Figura 20.

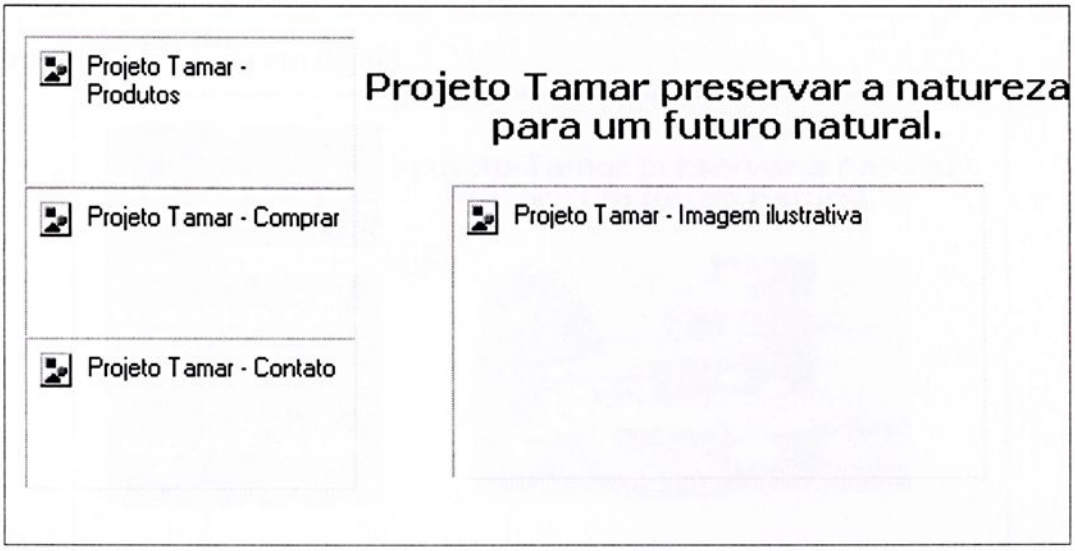


Figura 20. Visão de como a *tag alt* pode facilitar a navegação para usuários que configuram o navegador para não apresentar imagens.

4.12 Identificação de elementos não textuais para os mecanismos de busca

Os mecanismos de busca classificam o conteúdo dos documentos digitais através e exclusivamente por texto. Assim, qualquer outro item da composição do *site* que não estiver devidamente nomeado e/ou rotulado, acaba desperdiçando a oportunidade de ser um elemento contribuinte para a melhor classificação do documento digital.

Desse modo, todos os elementos não textuais deveriam possuir um nome vinculado à palavra-chave do *site* e/ou um nome significativo que identificasse esse elemento da forma mais clara possível para o usuário. Esta forma contribui para uma melhor classificação nas listagens de resposta dos mecanismos de busca, como exemplo, os botões "produtos", "comprar" e "contato" da figura 21, inicialmente nomeados como respectivamente, b1.gif, b2.gif e b3.gif. Para otimizar o processo de classificação dos mecanismos de busca levando em consideração a palavra-chave

"Projeto Tamar", os nomes dos botões deveriam ser trocados para "projeto_tamar_produtos.gif", "projeto_tamar_comprar.gif" e "projeto_tamar_contato.gif". Deste modo, cada um dos botões estaria contribuindo para a melhor pontuação e classificação do documento digital.



Figura 21. Elementos não textuais como botões e imagens se preparados corretamente podem contribuir para a classificação dos mecanismos de busca.

Mais um elemento não textual que pode prejudicar a classificação são os documentos digitais construídos, utilizando o formato SWF (*Shock Wave Flash*), conhecidos também como *sites* em *flash*. Com o formato SWF é possível criar documentos digitais inteiros, parciais ou mesclados com o HTML, dinâmicos ou estáticos, com ou sem acesso a banco de dados, mas que podem comprometer a classificação do documento, pois sua arquitetura interna vetorial impede qualquer utilização pelos mecanismos de busca que utilizam apenas elementos textuais.

Todos os *sites* que utilizam o formato SWF precisam de uma base HTML que chama o arquivo SWF. Essa base HTML precisa oferecer uma descrição bem elaborada, utilizando a *meta-tag-description* e a *meta-tag-keyword* para que os mecanismos de busca possam classificar alguns elementos. Adicionalmente, o arquivo SWF deve usar um nome vinculado à palavra-chave do *site* ou o material desenvolvido.

Como o arquivo SWF não pode ser identificado pelos mecanismos de busca, uma outra técnica é construir uma página inicial que possa descrever o que o

site pode oferecer e nessa página inicial, deve ser inserida um *link* para o *site* desenvolvido em formato SWF.

Outra possibilidade de uso para fazer o documento digital ser encontrado pelos mecanismos de busca é desenvolver novamente uma página inicial com uma boa descrição do *site.* É necessário que existam nesta página dois caminhos: um para a versão do documento digital com base HTML e outra com base SWF. Como o mecanismo de busca classifica o material feito em HTML, também classificará a versão SWF.

Os arquivos SWF não possuem apenas características negativas. Eles oferecem vantagens, como por exemplo, maior liberdade no desenvolvimento da interface, possibilidade de criação de conteúdo mais atrativo, visualização no navegador do usuário da forma como foi projetado, sem perder características de tipos de letras ou cores utilizadas. Oferecem também, de acordo com o planejamento, adaptação ao tamanho do monitor de forma automática, pois o formato SWF possui características vetoriais[17] e podem se adaptar às dimensões da tela. Além disso, imagens, sons e filmes podem ser vinculados a um único arquivo SWF.

Outros elementos que podem ser inseridos na *Web* como músicas, imagens, animações, filmes, programas aplicativos, arquivos PDFs, planilhas, arquivos de texto, etc, precisam ser identificados para que possam ser encontrados em uma situação de pesquisa. E para que os mecanismos possam encontrá-los, é necessário uma descrição textual adequada. Essa descrição pode ser feita no nome do arquivo, na *tag alt*, ou vinculando um texto visível que seja descritivo e que esteja ao redor do elemento disponibilizado(Kwok et al., 2001). É importante ressaltar que alguns mecanismos de busca já possuem a propriedade de classificar e indexar arquivos no formato PDF em suas bases de dados.

(17) Conjunto de n quantidades que dependem de um sistema de coordenadas n-dimensionais e que se transforma segundo leis bem determinadas quando se muda o sistema.

unesp

4.13 Características de uso das *meta-tags* e *tags alt*

Entre as diversas *meta-tags* existentes, as *meta-tags-description* e *keyword* são essenciais. A *meta-tag-description* deve ser usada para descrever em, até no máximo 255 caracteres, quais são as características do material divulgado do documento digital. Alguns mecanismos classificam menos que 255 caracteres. Mas isso não impede o funcionamento do *site, a*penas limitará a quantidade de caracteres lida pelo mecanismo de busca. Abaixo, o código da *meta-tag-description* do *site* www.nokia.com.br

<META NAME = "description" *CONTENT* ="A Nokia é lider mundial no fornecimento de aparelhos celulares e também no fornecimento de infra-estrutura de redes fixas e celulares incluindo serviços a clientes.">.

Ao contrário da *meta-tag-description,* que é visualizada na listagem de resposta do mecanismo de busca, a *meta-tag-keyword* não aparece em nenhum lugar, mas é utilizada para classificação na base do mecanismo de busca. É recomendado não ultrapassar os 255 caracteres, pois o que passar desse limite pode ser ignorado por alguns mecanismos. Se for de interesse do responsável pela identificação do documento digital, inserir mais *keywords*, é possível repetir a *tag* quantas vezes forem necessárias. Mas dependerá da política de classificação do mecanismo de busca classificar ou não mais de uma.

Entretanto, os mecanismos de busca podem detectar a repetição seqüencial de palavras destinadas a melhorar a classificação do documento digital de forma artificial. Isso é classificado como *Spam.* O *Spam* é radicalmente evitado, pois o propósito dos mecanismos de busca é não classificar um documento digital porque seu responsável aumentou artificialmente a classificação do *site,* mas tentar trazer a informação procurada mais próxima possível do interesse do usuário. Quando os



mecanismos de busca detectam a utilização de *Spam,* os *sites* que os utilizam são excluídos das bases de dados.
Abaixo exemplo do código da *meta-tag-keyword* do *site* www.nokia.com.br

<*META NAME* ="*keywords*" *CONTENT*="nokia, Nokia, NOKIA, telefones móveis, celular, celulares, Celulares, Celular, telecomunicação, telecomunicações, redes sem fios, rede sem fio, rede fixa, redes fixas, datacom, GSM, términais de multimedia, terminal de multimedia, monitor, monitores, aparelho de mão, apararelhos de mãos, aparelhos de mão, atendimentos ao consumidor, atendimento ao consumidor, atendimento aos consumidores, boletins de imprensa, boletim de imprensa, informação financeira, informações financeiras, intercâmbio de estudante, intercâmbio de estudantes, intercâmbios de estudantes, posição aberta, posições abertas, oportunidades de emprego, oportunidade de emprego, oportunidades de empregos, oportunidade de carreira na Nokia, oportunidades de carreira na Nokia " >

Uma técnica muito importante que deve ser utilizada e que é comprovada na primeira linha de código acima descrita, é prever a possibilidade de digitação que o usuário poderia realizar. Por esse motivo é que a palavra nokia é escrita de três formas diferentes, sem ser considerado um *Spam.*

Adicionalmente, as palavras inseridas na *meta-tag-keyword* devem ser obrigatoriamente separadas por vírgula para que sejam interpretadas como várias palavras, pois se não forem colocadas, o mecanismo interpretará como sendo um conjunto de uma só palavra.

4.14 *Spam*

Alguns responsáveis em divulgar o documento digital acreditam que podem, através de técnicas artificiais, aumentar a classificação nas listagens de

busca. Entretanto, se o mecanismo de busca detectar a utilização o *site* é excluído da base. Abaixo seguem-se alguns métodos de *Spam:*

- Seqüências idênticas de palavras ou frases;
- Preenchimento do texto visível ou não visível com a palavra-chave de forma aleatória no corpo do documento HTML, geralmente sem nenhum sentido textual;
- Preenchimento de palavras-chave com a mesma cor de fundo, tornando-as invisíveis para a visualização do usuário, mas visível para o mecanismo de busca;
- Outros *sites* criados exclusivamente com vários *links* idênticos para o *site* principal.

Alguns mecanismos de busca não revelam o que consideram como técnica de *Spam,* outros revelam parcialmente. Desta forma, não se deve julgar que, porque um mecanismo de busca não forneça dados referentes a técnicas de inibição de *Spam,* que ele não adote nenhuma.

4.15 Base de dados que buscam mais itens com procura única de uma palavra

A atitude de impedir a classificação do documento digital de forma artificial está ligada ao procedimento de facilitar o usuário na tentativa de encontrar o que está procurando. Assim, os mecanismos de busca permitem a utilização de combinações booleanas entre palavras (and, or, not, etc) uns com mais, outros com menos operadores. Essas características são utilizadas para filtrar o assunto procurado para minimizar a quantidade de opções das listagens de busca. Deve-se ressaltar que a utilização dos operadores booleanos é feita apenas por usuários experientes. Desta forma, uma preocupação dos mecanismos de busca também é atender, da melhor forma possível, aos usuários iniciantes. Assim, métodos de análise semântica, léxica e heurística contribuem para facilitar o uso de mecanismos de busca. De acordo com

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

pesquisa desenvolvida por Holscher & Strube, 2000, o mecanismo *Google* é o que rastreia maior quantidade de documentos digitais com procura realizada com uma só palavra. Selecionadas 25 palavras de maneira aleatória, as 23 primeiras posições pertenciam ao *Google,* batendo qualquer outro mecanismo de busca. Em uma outra pesquisa, Notess, 2002, comprova que em outro teste com mais 25 palavras aleatórias, que o *Google* encontra melhor que qualquer outro mecanismo de busca todas as 25 palavras. Contudo, isso não significa que as respostas sejam eficientes com relação ao seu conteúdo, pois, a palavra-chave no documento digital tem função fundamental e pode determinar maior ou menor índice neste tipo de procura efetuada por apenas uma palavra.

4.16 Base de dados Patrocinados

Muitas bases de dados existentes na *Web* oferecem dados patrocinados ou seja, em certas pesquisas os primeiros itens podem ter comprado sua posição na listagem de resposta. Essas bases que oferecem a possibilidade de comprar a posição de classificação estão descaracterizando o objetivo inicial da *internet* que era um meio rápido de se encontrar informação. Entretanto, por serem patrocinados, podem oferecer um serviço de acesso mais rápido e mais extenso em comparação a bases que não são patrocinadas. Mas isso não significa que o serviço oferecido pelos não patrocinados possa ser de menor qualidade. Muitas vezes, por ser uma base de menor amplitude e de assunto restrito, as informações contidas nessas bases podem ter maior relevância que em bases convencionais.

Outra característica de documentos digitais patrocinados é a inclusão, no topo da listagem de busca, de algum elemento gráfico que divulga algum material que, em alguns casos, pode ter ligação direta com a pesquisa realizada.

Adicionalmente, essa atitude de vender a posição nas listagens de pesquisa não é interessante para o usuário, pois o mesmo pode levar em consideração



uma informação patrocinada obtida nas listagens de busca como verdade e transformar essa informação em verdade absoluta, prejudicando a forma de seu julgamento em relação à informação encontrada.

Além disso, mais alguns problemas podem ser citados. Algumas bases de dados que evitam a utilização de técnicas de *Spam* vendem lugar na classificação de suas listagens de respostas. Outros mecanismos de busca oferecem dificuldade de acesso a documentos digitais que não tenham em sua extensão de URL a complementação .com. Isso significa que o conteúdo tem alguma relação com comércio, deixando, assim, de classificar outros documentos digitais, que podem trazer em sua extensão de URL o complemento .org, .gov, entre outros.

4.17 Proposta de metodologia para classificação otimizada de documentos digitais nos mecanismos de busca

Este estudo tem como objetivo apresentar uma metodologia para tornar um documento digital melhor classificado, utilizando como recursos seu código, adaptando, adequadamente, itens internos que podem contribuir para a melhor classificação nos mecanismos de busca.

Desta forma, alguns itens abaixo podem ser seguidos para se obter uma classificação otimizada:

1º Escolher qual ou quais serão os mecanismos de busca que serão usados para indexação do conteúdo digital;

2º Extrair dos mecanismos de busca escolhidos quais são os itens de maior importância para que os mesmos, realizem o processo de classificação de seus bancos de dados.



3º Com o material de desenvolvimento do *site* em mãos, identificar qual é a palavra-chave de maior importância (ou palavras-chave). De preferência, que o *site* ainda esteja em processo de planejamento pois, caso contrário, será necessário alterar muitos itens, como por exemplo, todos os elementos não textuais existentes no documento digital.

4º Usando os mecanismos de busca que indexarão o documento digital ainda a ser criado, é necessário realizar uma pesquisa com a palavra-chave identificada no item anterior. É preciso ainda verificar nos primeiros *sites* encontrados na listagem de busca qual é o número de vezes que a palavra-chave é referenciada para poder classificar os *sites* encontrados. Assim, se o objetivo for ser o primeiro na listagem de busca, o número de repetições da palavra-chave deve ser maior que a usada no *site,* que foi classificado como primeiro, sem a utilização de técnicas de *spam.* Além do mais, alguns *sites* podem ser classificados em posições privilegiadas sem terem utilizado recursos de palavras-chave. Quando for encontrada esse tipo de situação, significa que o documento digital comprou sua posição no mecanismo de busca.

5º Os nomes internos dos possíveis diretórios para armazenar determinados conteúdos referentes ao documento digital devem também ser nomeados, utilizando a palavra-chave do *site* e mais um complemento que possa caracterizar os elementos que serão guardados neste diretório. Por exemplo: projeto_tamar_imagens. Esses pequenos detalhes podem diferenciar um documento digital de outro na classificação.

6º O URL (*Uniform Resource Locator*) ou simplesmente o endereço do *site* deve, dentro do possível, possuir uma referência à palavra-chave, sempre com o objetivo de pontuação. Exemplo de URL com a palavra-chave inserida: http://www.projeto_tamar.org.br ou http://www.projeto_tamar.hpg.com.br.

7º Utilizar a *tag title* referenciando o nome do *site,* o assunto que trata o *site*, ou produto que se encontra no *site.* Não se deve usar artigo antes do substantivo que identifica a *tag title.* Não se deve colocar nenhum outro caractere antes ou depois do conteúdo da *tag title,* pois a ocorrência mais idêntica à palavra pesquisada pelo usuário será



classificada em uma posição mais elevada que uma ocorrência similar. De preferência, o conteúdo da *tag title* deve ser o mais próximo possível da palavra-chave escolhida. Não se deve colocar mais de 60 caracteres na *tag title.*

8º A identificação do documento digital é essencial para o rastreamento dos mecanismos de busca. Assim, as *meta-tags* não podem ser esquecidas. Dependendo do conteúdo do *site*, não existe a necessidade da utilização de todas elas. Mas todo *site* deve usar pelo menos as seguintes:

- ***<META NAME="Keywords" CONTENT="palavras_separadas_por_vírgula">***
- ***<META NAME="Description" CONTENT="descrição_da_página_ou_site">***
- ***<META NAME="Robots" CONTENT="all \| index \| noindex \| follow">***
- ***<META HTTP-EQUIV="Content-Language" CONTENT="br">***

As características dessas e outras *meta-tags* podem se consultadas no item 4.9.

9º Utilizar a *tag alt* corretamente para que se possa somar pontos na classificação do mecanismo de busca. Para se fazer uso correto da *tag alt*, é necessário vincular a palavra-chave do *site* com o nome do item ou uma pequena descrição do que será encontrado na outra ponta do *link*, isso se o elemento for um *link.*

10º Todos os elementos não textuais como botões, marcadores, arquivos de imagens fotográficas, arquivos de imagens de desenhos ou logomarcas, arquivos de música ou efeitos sonoros, animações, apresentações, arquivos PDFs (alguns mecanismos de busca classificam e indexam arquivos no formato PDF), arquivos executáveis, planilhas, arquivos de textos, etc, devem estar vinculados a uma *tag alt* ou envolvidos por texto descritivo, para que possam ser detectados e classificados.

11º A posição do texto visível vinculado com a posição da palavra-chave é crucial para a boa pontuação do documento digital. Assim, respeitando o *layout* da página, quanto mais próximo do início desta e à esquerda, mais forte é a pontuação. Essa



característica não é a única forma de pontuação, pois é possível compor um *layout* fora desses padrões estabelecidos, pontuando menos em relação ao texto, mas pontuando mais em relação a outros elementos.

12º Quanto mais próximo o *link* da informação desejada, mais pontos é oferecido para a classificação do *site*. Para isso é preciso evitar posicionar a informação em camadas de *links* internos. Isso obrigaria o usuário a clicar inúmeras vezes. Como o objetivo é tornar a informação a mais próxima do usuário, deve-se estabelecer como regra geral sobre *links* uma distância máxima de três cliques.

13º Os *Links* que estão no documento digital que apontem para outros *sites* bem acessados e que possuam ligações com o material oferecido, pode aumentar a probabilidade de classificação. Outros *sites* que possam apontar para o documento digital também oferecem pontuação na classificação, se o *site* que aponta for de uma grande empresa, ou de um *site* que tenha um grande índice de acessos, pode acarretar uma classificação melhor. Desde que exista um elo de ligação entre o *site* que aponta e o *site* apontado. Nem todos os mecanismos possuem, no entanto, essa característica.

Se o responsável ou a equipe responsável pelo desenvolvimento do *site* conseguir reunir partes desses elementos ou todos eles, as possibilidades de boa classificação do documento digital serão ampliadas nos mecanismos de busca.

Paralelamente aos itens descritos, foram desenvolvidos em equipes de alunos do 3º ano do Colegial Técnico Industrial na Unesp de Bauru em 2002, três *sites* corporativos que seguindo os itens anteriormente descritos estão classificados em posições relevantes nas listagens de busca do AltaVista, *Google* , *Yahoo* com procura de palavra chave respectivamente "Zôo Bauru" , "Cot" e "Sihop", figura 22, 23 e 24.



Figura 22. *Site* corporativo do Zoológico de Bauru. Primeira posição usando a palavra-chave "zôo bauru" no mecanismo de busca AltaVista. www.zoobauru.kit.net.

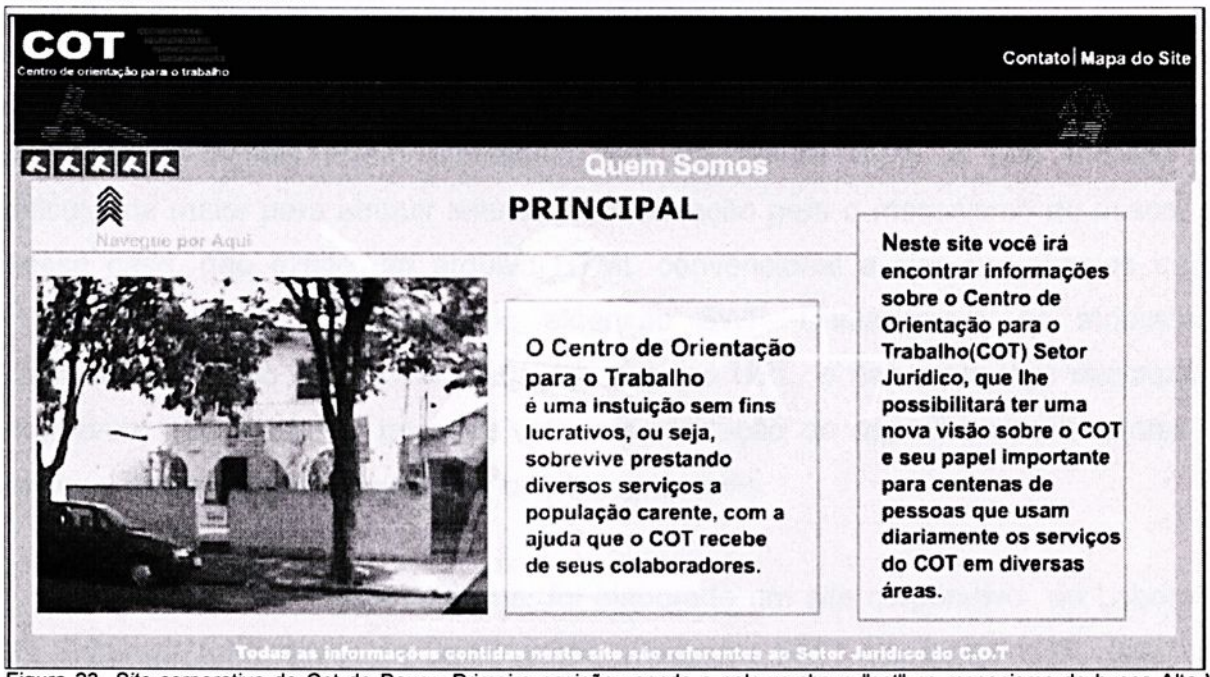


Figura 23. *Site* corporativo do Cot de Bauru. Primeira posição usando a palavra-chave "cot" no mecanismo de busca Alta Vista. www.cot-bauru.cjb.net.



Figura 24. *Site* do Sistema de Horário de Professores do Colégio Técnico Industrial, classificado no *Yahoo* em primeira posição com a palavra-chave "Sihop", também classificado entre os 10 primeiros no Google e AltaVista www.sihop.kit.net.

É importante detalhar que nos dois primeiros casos, os documentos digitais foram planejados em arquitetura SWF, o que oferece uma dificuldade maior para atribuir leitura e classificação para o mecanismo de busca, pois nesse caso, não existe um arquivo HTML convencional e sim, um simples arquivo HTML que chama o arquivo de extensão SWF. Desta forma, os atributos de classificação estão concentrados da *tag title*, no URL, e nas *meta-tags description* e *keywords,* não existindo texto visível ou a utilização da *tag alt* para elementos não textuais como imagens ilustrativas ou botões de *links*.

Posteriormente, foi elaborado um *site* corporativo do Laboratório de Sistemas Adaptativos de Computação Inteligente SACI que também foi classificado nos mecanismos de busca, comprovando a necessidade de planejar, adequadamente, o código interno dos documentos digitais para se obter uma classificação otimizada em relação a outros documentos não planejados. A figura 25 demonstra o comportamento esperado do *site* SACI entre os mecanismos de busca. O capítulo 5 abordará o processo de criação e os resultados obtidos na classificação dos mecanismos de busca.

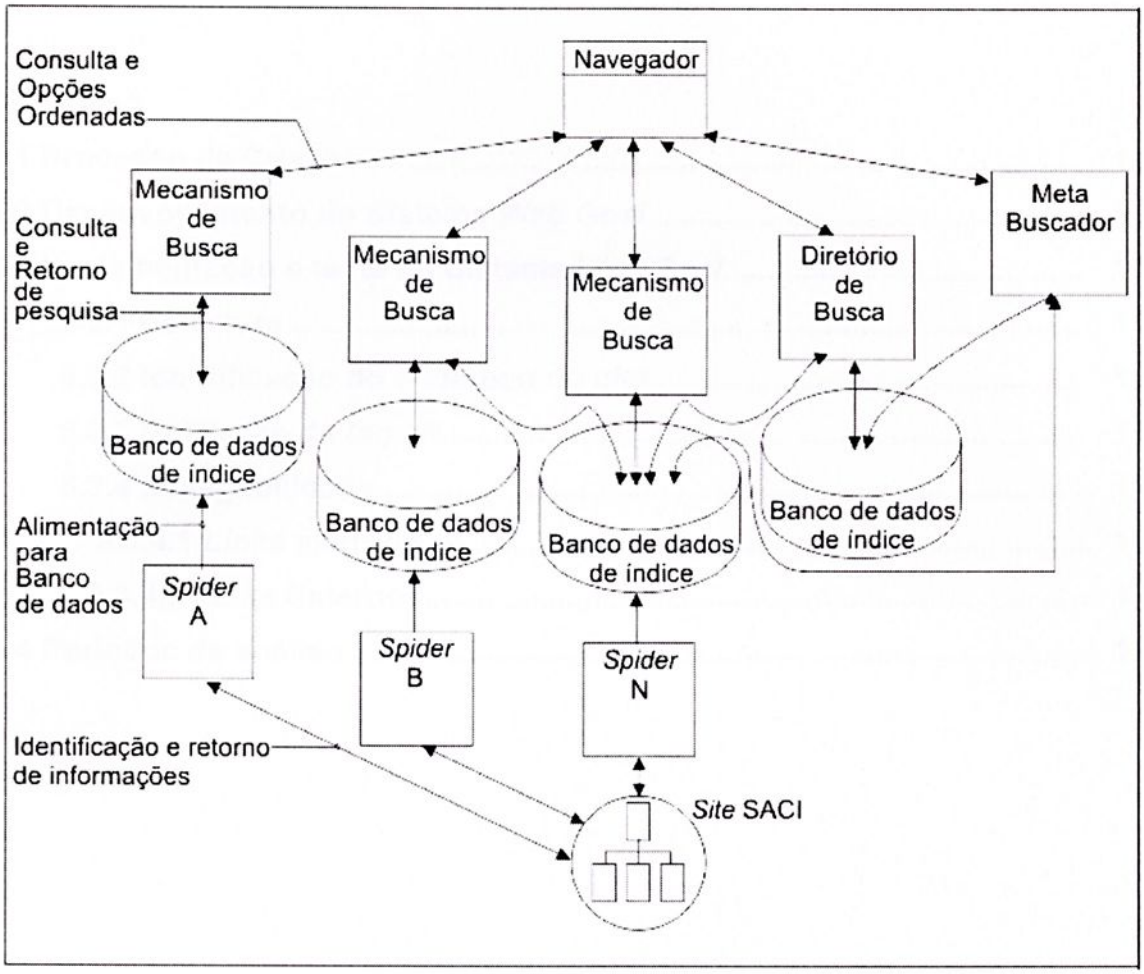


Figura 25. Diagrama da proposta de identificação do *site* SACI entre os mecanismos de busca, meta-buscadores e diretórios.

Capítulo 5

5.1 Processo de Criação..... 110
5.2 Desenvolvimento do Sistema *Web Goal*..... 114
5.3 Implementação e teste do Sistema *Web Goal*..... 115
5.3.1 Estrutura..... 115
5.3.2 Identificação do endereço do *site*..... 119
5.3.3 Utilização da *tag alt*..... 119
5.3.4 Navegabilidade..... 121
5.3.4.1 *Links* Internos..... 121
5.3.4.2 *Links* Externos..... 121
5.4 Períodos de análise..... 121

Capítulo 5

5.1 Processo de Criação

Para o desenvolvimento do processo de criação do *site* SACI foi realizado um estudo de caso. Neste estudo, foram consultados alguns *sites* que traziam algum vínculo informacional em relação ao tema "Sistemas Adaptativos e Computação Inteligente". Desta forma, verificamos através de consultas aos mecanismos de busca e diretórios, quais os *sites* classificados em boas posições (de primeira a décima posição), e verificamos também a concepção visual de cada um deles.

Essa pesquisa, feita com os materiais já disponibilizados na *Web,* foi de grande valia, pois entre outros fatores, foi possível estabelecer critérios de como não oferecer conteúdo textual nas páginas do *site* de forma a não prejudicar o fator de

unesp

usabilidade na leitura das informações apresentadas. Após a pesquisa, foram elaboradas várias propostas de modelos de interface.

O próximo passo do processo foi o de apurar as idéias sempre com a intenção de estabelecer uma composição agradável e de pouco peso para descarregamento (*download*). O desenvolvimento da segunda etapa pode ser vista a seguir. Figura 26 a 33.

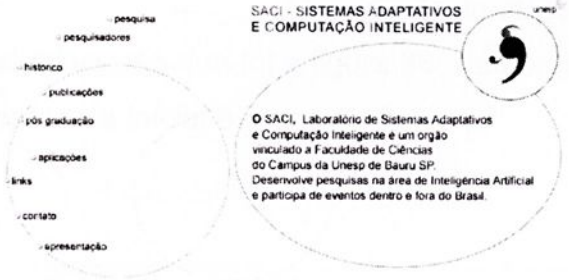


Figura 26. Opção de interface 1.



Figura 27. Opção de interface 2.



Figura 28. Opção de interface 3.

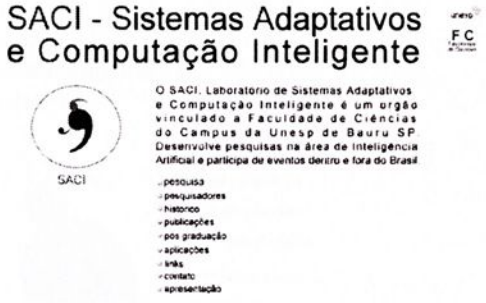


Figura 29. Opção de interface 4.

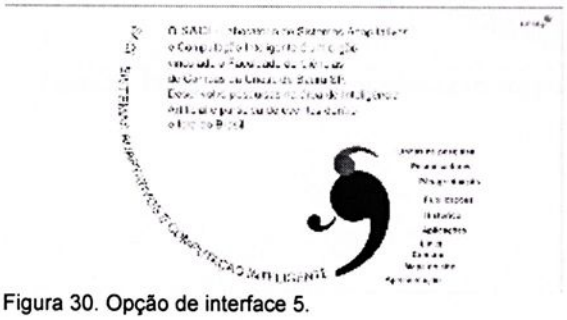


Figura 30. Opção de interface 5.



Figura 31. Opção de interface 6.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18



Figura 32. Opção de interface 7.



Figura 33. Opção de interface 8.

A partir desta etapa, foi escolhida a interface que deu origem a atual versão que foi à figura 30. Nas próximas figuras, 34 e 35 podemos verificar a proposta inicial e a versão final.



Figura 34. Proposta de interface escolhida para desenvolvimento da *home page* do Lab. SACI.



Figura 35. Interface atual do *site* do Lab. SACI. O logo do Laboratório foi desenvolvido por alunos do curso de Desenho Industrial.

Em relação à imagem da figura 34, algumas modificações foram realizadas até chegarmos à imagem da figura 35. O texto central de identificação foi retirado e foram colocadas imagens ilustrativas referentes aos estudos e pesquisas desenvolvidos no laboratório SACI. Também pode ser notado na parte intermediária da página um item de usabilidade não muito aplicado em documentos digitais, que é o endereço físico do laboratório.

5.2 Desenvolvimento do Sistema *Web Goal*

Com o objetivo de comprovar nossa metodologia para classificação de documentos digitais descritos no capítulo 4, foi elaborado o documento digital referente ao *site* SACI – Laboratório de Sistemas Adaptativos e Computação Inteligente. Assim, foram eleitos alguns mecanismos de busca como o *Google*, o AltaVista, o Radix, o Acheiaqui e alguns diretórios de busca como o *Yahoo,* o *Open Directory* e o Cadê para submeter o endereço do *site* SACI nas respectivas bases de dados.

Assim, definidos os mecanismos de busca e os diretórios escolhidos, iniciamos a verificação das informações referentes aos itens que podem ser utilizados para a classificação do *site* SACI. Verificamos que todos os mecanismos de busca e diretórios consultados possuem algumas particularidades. Utilizaremos como exemplo os mecanismos de busca *Google* e AltaVista e o diretório *Yahoo.*

- ***Google***

  Atribui grande importância a apontamentos externos e/ou internos a que o documento digital se refere.

- **AltaVista**

  Não indica em nenhum momento se apontamentos externos ou internos podem favorecer a classificação de um documento digital.

- ***Yahoo***

  Os *sites* enviados são selecionados por editores humanos como já foi informado e, desta forma, passam por avaliação subjetiva podendo resultar em um não cadastramento nas bases dos diretórios.



Além disso, algumas bases de dados, sejam mecanismos ou diretórios, podem cobrar a hospedagem do documento digital, oferecendo ou não uma melhor classificação no resultado de busca de acordo com o interesse e a disposição financeira do interessado. Para o desenvolvimento deste estudo optamos por mecanismos de busca e diretórios livres de taxas de hospedagem.

5.3 Implementação e teste do Sistema *Web Goal*

5.3.1 Estrutura

Logo depois de terem sido escolhidos os mecanismos de busca e diretórios e com o material de desenvolvimento em mãos, foi planejada a estrutura de navegação do *site*. De acordo com esse planejamento criamos uma estrutura de hipertexto que estabelece contato direto para qualquer ponto do documento digital, eliminando uma possível navegação linear. Outra característica de planejamento foi o direcionamento externo para outros *web sites* com informações relativas às áreas de interesse que o *site* SACI tem como objetivo. Esses *links* externos têm sua importância na pontuação em alguns mecanismos de busca. Desta forma, a escolha e colocação de um *link* externo com boa audiência poderá acarretar melhor pontuação na classificação dos mecanismos de busca. E mesmo que alguns mecanismos não classifiquem *links* externos, a colocação dos mesmos favorece a usabilidade do *site* SACI em relação às opções oferecidas pelo documento digital ao usuário, formando assim uma rede de informações interligadas entre si.

Outro item fundamental foi a escolha da(s) palavra(s)-chave, que no caso do *site* do Laboratório SACI foram definidos como:

- redes_neurais_sistemas_adaptativos;
- redes_neurais;
- neurais.



Essas palavras-chave foram amplamente utilizadas em todos os elementos internos (no código html) e externos (objetos gráficos e texto visível), sendo utilizadas da seguinte forma: “redes_neurais_sistemas_adaptativos”, “redes_neurais” e “neurais” juntamente com mais uma palavra que complementasse e identificasse o objeto em questão. Como exemplo real podemos citar o logotipo do Laboratório SACI que se encontra em todas as páginas internas no canto superior esquerdo. Todos os logotipos possuem o mesmo nome “redes_neurais_saci.gif” e apenas com a referência à imagem do logotipo se somam nove citações a “redes_neurais” em uma ótima posição, que é o lado superior esquerdo. Figura 36. Outro exemplo real pode ser visto na figura 37, onde todos os arquivos e pastas foram referenciados com uma das palavras-chave já citadas.

Um importante procedimento realizado com a palavra-chave dentro da estrutura do documento digital é superar o número de ocorrências existentes em *sites* concorrentes na classificação. Isso deve ser realizado para assegurar que o material que está sendo desenvolvido possa ter uma referência à palavra-chave maior que o número de vezes que apresenta(m) o(s) *site(s)* concorrente(s). Desde que os mesmos não sejam *sites* patrocinados.



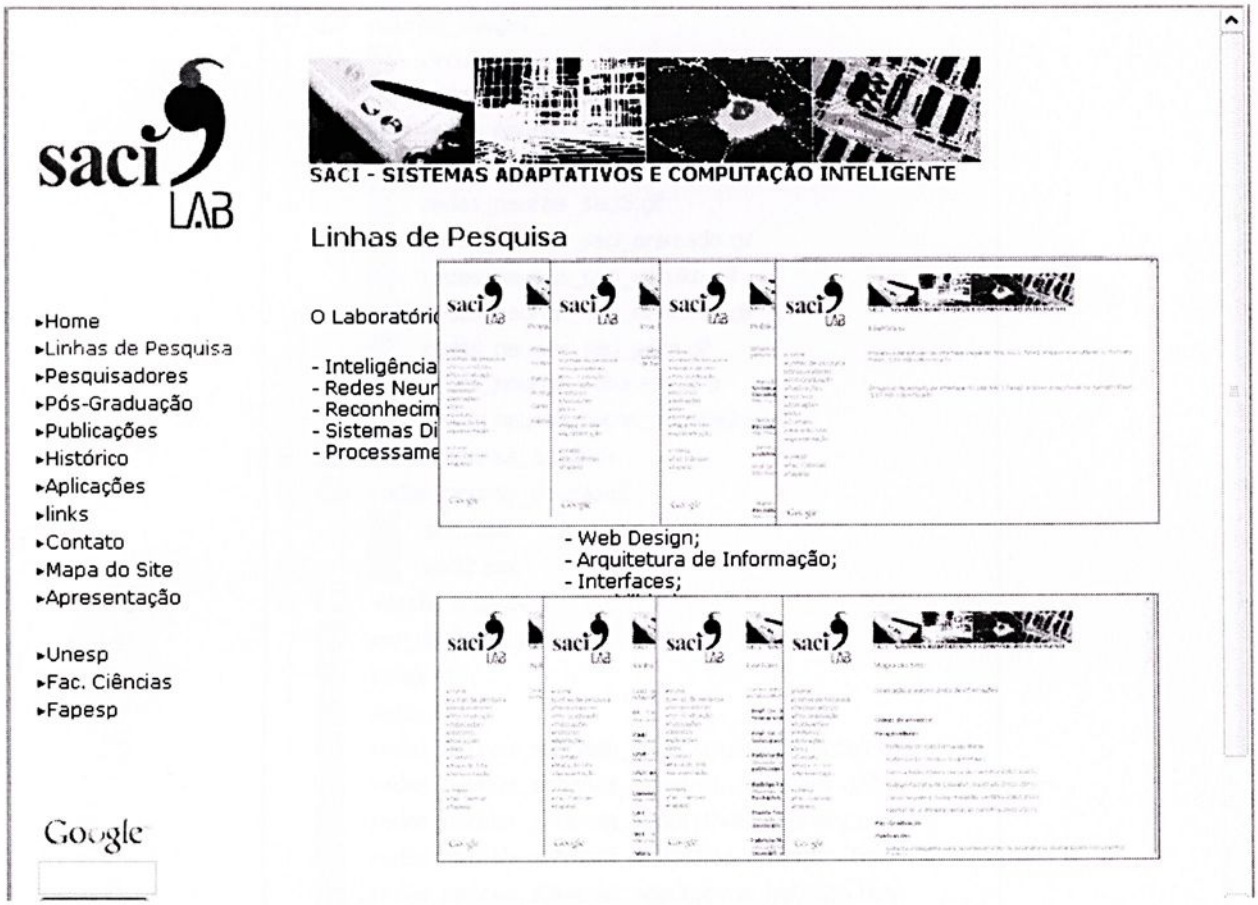


Figura 36. Em segundo plano, página interna do *site* SACI e a logomarca, em primeiro plano, outras páginas internas. Todas as imagens possuem o nome "redes_neurais_saci.gif", como nome do logo.

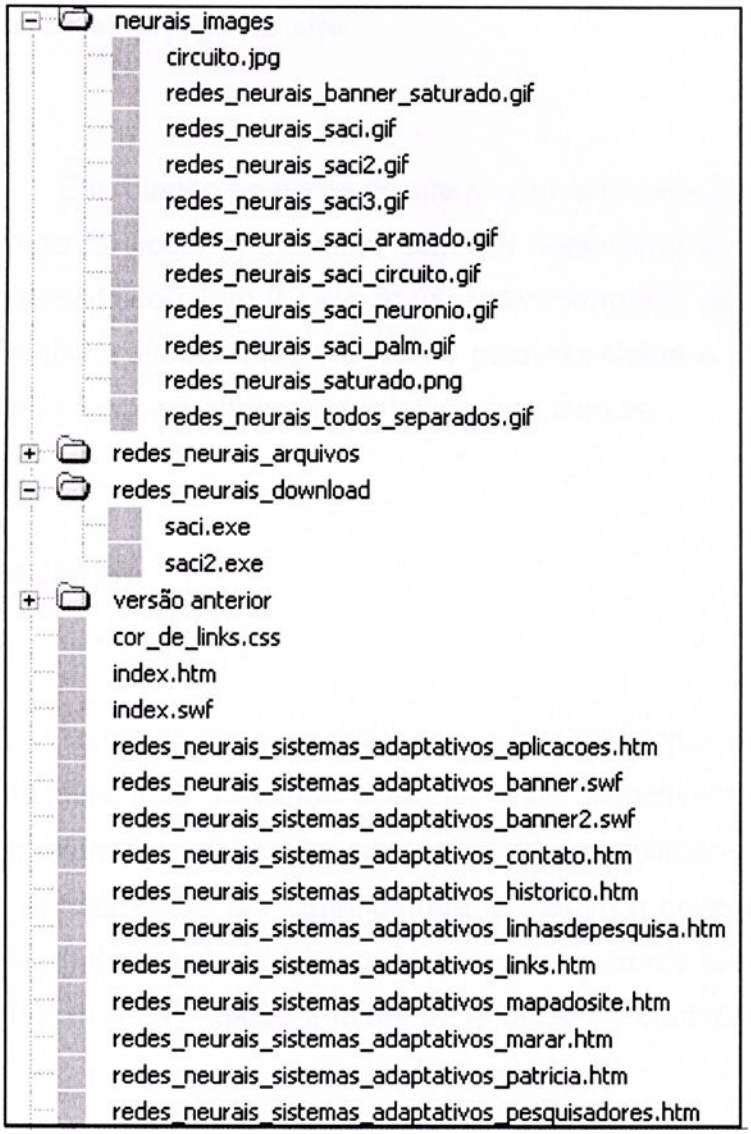


Figura 37. Visualização dos nomes das pastas e arquivos com a utilização das palavras-chave.

5.3.2 Identificação do endereço do *site*

Em relação ao nome do *site* no provedor de acesso do Campus da Unesp de Bauru, não foi possível trabalhar com um nome próximo da palavra chave. Desta forma, o endereço ficou: http://wwwp.fc.unesp.br/~fermarar. Já o endereço do *site* espelho está utilizando mais adequadamente as palavras-chave e pode ser acessado pelo endereço: http://www.sacisistemasadaptativos.hpg.com.br.

5.3.3 Utilização da *tag alt*

Todas as imagens utilizadas no *site* SACI que são do tipo JPG[1] ou GIF[2] possuem uma descrição utilizando a *tag alt.* Este procedimento favorece vários aspectos como por exemplo: pontuação, navegabilidade, usabilidade (para o caso de o documento digital ser acessado por usuários que formatam o navegador para mostrar apenas conteúdo textual, ou por usuários que usam navegadores textuais) e por último, por usuários que fazem a navegação através de audição do conteúdo textual lido pelo navegador.

Outra *tag* de identificação de grande valor é a *tag title* que atribui nome para o conteúdo da página. Para o *site* SACI foram elaborados os seguintes títulos. Tabela 6.

(1) JPG. Sigla para *Joint Photographic Experts Group*, o nome original do comitê que escreveu o padrão desse formato de compressão de imagens. Funciona bem com fotos e desenhos naturalísticos, mas não é tão eficiente com desenhos de letras, linhas e *cartoons*.
(2) GIF. Sigla para *Graphics Interchange Format*. Formato de arquivos de imagens mais utilizado na *Web*. O formato GIF cria arquivos de imagens de tamanho relativamente pequeno em relação aos demais formatos. O tipo de compactação utilizada no formato GIF funciona melhor quando a imagem tem áreas contínuas da mesma cor e, principalmente, poucas cores.

Identificação da página	Título da página
Página inicial	SACI, Sistemas Adaptativos e Computação Inteligente – FC – Faculdade de Ciências, Unesp, Bauru
Linhas de Pesquisa	SACI, Linhas de pesquisa
Pesquisadores	SACI, Pesquisadores
Pós-Graduação	SACI, Pós Graduação
Publicações	SACI, Publicações
Histórico	SACI, Histórico
Aplicações	SACI, Aplicações
Links	SACI, Links
Contato	SACI, Contato
Mapa do Site	SACI, Mapa do Site
Apresentação	SACI, Apresentação

Tabela 6. Identificação dos títulos da *tag alt* de cada página do *site* SACI.

As identificações das *Meta tags* foram colocadas em todas as páginas do *site*, veja tabela 7.

Meta tags	**Preenchimento**
Meta tag description	Laboratório SACI, sistemas adaptativos e computação inteligente, Unesp Bauru, São Paulo, Brasil, Prof Dr João Fernando Marar, Redes neurais artificiais, Processamento de Imagens, Inteligência Artificial, Reconhecimento de sinais, Sistemas distribuidos de informação, Artes e novas tecnologias, web design, arquitetura de informação, Interfaces, Usabilidade.
Meta tag keyword	Redes Neurais, redes neurais, neural network, Sistemas Adaptativos, sistemas adaptativos, computação, Inteligência Artificial, João Fernando Marar, Unesp, Unesp Bauru, unesp, FC, fc, faculdade de ciências, mestrado, doutorado, orientação, orientador, pesquisa, Reconhecimento de sinais, webdesign, Interfaces
Meta tag robots	*all*
Meta tag revisit after	15 *days*
Meta tag language	br
Meta tag type	*text*/html; *charset*=iso-8859-1

Tabela 7. *Meta-tags* utilizadas nas páginas do *site* SACI. Todas as descrições referentes às *meta-tags* foram feitas no capítulo 4 item 4.9.

5.3.4 Navegabilidade

5.3.4.1 *Links* Internos

Em relação à navegabilidade do *site* SACI, todos os seus itens internos estão interligados, proporcionando um controle maior do usuário, pois a estrutura montada, além de oferecer controle total para qualquer item que o usuário deseja acessar, proporciona, através da estrutura de hipertexto interna, uma ligação total de todas as páginas para todas as páginas, evitando levar o usuário para um item sem retorno.

5.3.4.2 *Links* externos

Em relação aos *links* externos, todas as páginas possuem três que são: Unesp, Faculdade de Ciências e Fapesp e mais uma página dedicada a encaminhamento para outras instituições que possuem o mesmo ponto em comum com o material do *site* SACI. Todos os *links* externos que possuam algum tipo de vínculo com a informação divulgada, podem oferecer para determinados mecanismos de busca uma probabilidade maior de classificação. Desta forma, essa particularidade também foi utilizada para possibilitar que o *site* SACI pudesse utilizar esse recurso para pontuação.

5.4 Períodos de análise

Com o documento digital finalizado utilizando todos os recursos anteriormente comentados, foi enviado em 1º de maio de 2003 para o provedor da Unesp de Bauru e para um provedor gratuito. Após o envio, foi realizado o cadastramento do documento digital nos mecanismos de busca *Google*, AltaVista, Radix, Acheiaqui, Achei e, também nos diretórios *Open Diretory*, *Yahoo* e Cadê.



Em 21 de maio de 2003, o *Yahoo* classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente"; 1º lugar com palavra chave "sistemas adaptativos" e 1º com palavra chave "computação inteligente".

Em 26 de maio de 2003, o Cadê classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente"; 1º lugar com palavra chave "sistemas adaptativos" e 1º lugar com palavra chave "computação inteligente".

Em 30 de maio de 2003, o Altavista classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente"; 1º lugar com palavra chave "sistemas adaptativos"; 1º lugar com palavra chave "computação inteligente"; 7º lugar com palavra chave "saci"; 1º lugar com palavra chave "adaptativos unesp" e 1º lugar com palavra chave "adaptativos".

Em 04 de junho de 2003, o Acheiaqui classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente"; 1º lugar com palavra chave "sistemas adaptativos"; 1º lugar com palavra chave "computação inteligente" e 4º lugar com palavra chave "saci".

Em 07 de junho de 2003, o MSN (www.msn.com), classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente" e em 14º lugar com palavra chave "computação inteligente".

Em 09 de junho de 2003, o *Ixquick* (www.*ixquick*.com) classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente".



Em 09 de junho de 2003, o *Looksmart* (www.*loocksmart*.com) classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente"; 2º lugar com palavra chave "sistemas adaptativos" e 3º lugar com palavra chave "computação inteligente".

Estas e outras informações de classificação podem ser vistas nas tabelas 8 e 9. Vale lembrar que os mecanismos que foram utilizados para cadastrar o *site* SACI foram o *Google,* o Altavista, o *Yahoo*, o Cadê, o Achei, o Radix, o *Open Diretory e* o Acheiaqui. A classificação obtida nos mecanismos de busca e diretórios citados permitiu ainda a classificação em outros mecanismos tais como: MSN, *Looksmart. Ixquick,* AOL, *Excite, Dogpile, Infospace, Webcrawler, Fast Search,* Tay, *Profusion, Metacrawler e Lycos,* foram classificadas através da utilização de acesso a bancos de dados externos, comprovando desta maneira o acesso que outros mecanismos fazem a partir de mecanismos que possuam uma grande base de dados.

Entretanto, mesmo seguindo essas recomendações, há no mínimo duas variáveis que estão fora de controle do responsável pelo documento digital e que são o algoritmo de classificação dos mecanismos de busca, que pode ser alterado sem prévio aviso e a inclusão de novos *sites* que possuam melhores referências de palavras-chave e códigos internos.

No entanto, é preciso ressaltar que, um documento digital que possa estar bem classificado em um momento, poderá não estar em outro. É por isso que todo documento digital precisa de manutenção periódica para evitar seu desaparecimento repentino e/ou sua substituição por outro. Da mesma forma que um *site* pode ser projetado para estar bem classificado, outros *sites* podem ser projetados para substituir esses *sites* já classificados.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Tabela de Classificação do *site* SACI nos mecanismos de busca

No ar	envio	YH	CD	AV	ACAQ	MSN	LS	IXQ	GG	OAL	EXT	DGP	ISP	WCR	FS	TAY	PFS	MCR	LYC	RDX
1/5																				
	13/5																			
		21/5																		
			26/5																	
				3/6																
					4/6															
						7/6														
							9/6	9/6												
									20/6	20/6										
											27/6									
												30/6								
													5/7							
														9/7						
															16/7					
																17/7				
																	18/7	18/7	18/7	
																				19/7

Tabela 8. Visualização das datas de indexação dos mecanismos de busca.

YH-*Yahoo*, CD-Cadê, AV-AltaVista, ACAQ-AcheiAqui, MSN-*Microsoft*, LS-*Looksmart*, LXQ-*Ixquick*, GG-*Google*, AOL-America *On Line*, EXT-*Excite*, DGP-*Dogpile*, ISP-*Infospace*, WCR-*Webcrawler*, FS-*Fast search (Alltheweb)*, TAY-TAY, PFS-*Profusion*, MCR-*Metacrawler*, LYC-*Lycos*, RDX-*Radix*.



Tabela com informações complementares sobre os mecanismos de busca

***Status* de envio**	Refe-rência	Nome do mecanismo	Endereço do mecanismo	Data de indexação	Tem-po / Dias
enviado	YH	***Yahoo***	**http://www.*yahoo*.com.br**	**21/05**	**08**
enviado	CD	**Cadê**	**http::/www.cadê.com.br**	**26/05**	**13**
enviado	AV	**AltaVista**	**http://www.altavista.com.br**	**03/06**	**21**
enviado	ACAQ	**AcheiAqui**	**http://www.acheiaqui.com.br/**	**04/06**	**22**
não enviado	MSN	*Microsoft*	http://www.msn.com.br	07/06	25
não enviado	LM	*Looksmart*	http://www.*looksmart*.com	09/06	27
não enviado	LXQ	*lxquick*	http://www.*lxquick*.com	09/06	27
enviado	GG	***Google***	**http://www.*google*.com**	**20/06**	**38**
não enviado	AOL	America *On Line*	http://www.oal.com.br	20/06	38
não enviado	EXT	*Excite*	http://www.*excite*.com/	27/06	45
não enviado	DGP	*Dogpile*	http://www.*dogpile*.com	30/06	48
não enviado	ISP	*Infospace*	http://www.*infospace*.com	05/07	53
não enviado	WCR	*Webcrawler*	http://www.*webcrawler*.com/	09/07	57
não enviado	FS	*Fast search (Alltheweb)*	http://www.*alltheweb*.com	16/07	64
não enviado	TAY	TAY	http://www.tay.com.br	17/07	65
não enviado	PFS	*Profusion*	http://www.*profusion*.com	18/07	66
não enviado	MCR	*Metacrawler*	http://www.*metacrawler*.com/	18/07	66
não enviado	LYC	*Lycos*	http://www.*lycos*.com	18/07	66
enviado	RDX	**Radix**	**http://www.radix.com.br**	**19/07**	**67**

Tabela 9. Informações complementares sobre os mecanismos.



Tabela de identificação das palavras chaves encontradas

mecanismo	Palavra-chave	Classificação	característica
Yahoo Consulta em 21/05/2003	sistemas adaptativos e computação inteligente	1º	1º entre 1
	sistemas adaptativos	1º	1º entre 2
	computação inteligente	1º	1º entre 9
Cadê Consulta em 26/05/2003	sistemas adaptativos e computação inteligente	1º	1º entre 1
	sistemas adaptativos	1º	1º entre 2
	computação inteligente	1º	1º entre 9
AltaVista Consulta em 03/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 149
	sistemas adaptativos	1º	1º entre 1283
	computação inteligente	1º	1º entre 3947
	Saci	7º	7º entre 6907
	Adaptativos	1º	1º entre 1937
	unesp adaptativos	1º	1º entre 94
AcheiAqui Consulta em 04/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 1
	sistemas adaptativos	1º	1º entre 1
	computação inteligente	1º	1º entre 1
	Saci	8º	8º entre 8
MSN Consulta em 07/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 159
	computação inteligente	14º	14º entre 6781
Looksmart Consulta em 09/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 789
	sistemas adaptativos	2º	2º entre 1740
	computação inteligente	3º	3º entre 4339
Ixquick Consulta em 09/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 157
Google Consulta em 20/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 256
	sistemas adaptativos	3º	3º entre 2260
	computação inteligente	5º	5º entre 8110
	computação adaptativos	2º	2º entre 801
	sistemas unesp	17º	17º entre 15900
	adaptativos unesp	1º	1º entre 148
	adaptativos	6º	6º entre 3420
	saci marar	2º	2º entre 5

América *On Line* Consulta em 20/06/2003	sistemas adaptativos e computação inteligente	1º	1º entre 250
	sistemas adaptativos	5º	5º entre 2180
	computação inteligente	16º	16º entre 7740
	adaptativos	8º	8º entre 3290
	adaptativos unesp	1º	1º entre 147
	computação adaptativos	2º	2º entre 792
	sistemas unesp	20º	20º entre 15600
Excite Consulta em 27/06/2003	sistemas adaptativos e computação inteligente	4º	4º não divulga
	sistemas adaptativos	8º	8º não divulga
	computação inteligente	8º	8º não divulga
	adaptativos unesp	3º	3º não divulga
	unesp adaptativos	1º	1º não divulga
	unesp inteligente	2º	2º não divulga
	unesp saci	3º	3º não divulga
	saci unesp	4º	4º não divulga
Dogpile Consulta em 30/06/2003	sistemas adaptativos e computação inteligente	1º	1º não divulga
	sistemas adaptativos	7º	7º não divulga
	computação inteligente	6º	6º não divulga
	unesp adaptativos	1º	1º não divulga
	adaptativos unesp	1º	1º não divulga
Infospace Consulta em 05/07/2003	sistemas adaptativos e computação inteligente	5º	5º não divulga
	sistemas adaptativos	9º	9º não divulga
	computação inteligente	10º	10º não divulga
	unesp adaptativos	1º	1º não divulga
	adaptativos unesp	1º	1º não divulga
Webcrawler Consulta em 09/07/2003	sistemas adaptativos e computação inteligente	5º	5º não divulga
	sistemas adaptativos	9º	9º não divulga
	computação inteligente	8º	8º não divulga
	unesp adaptativos	1º	1º não divulga
	adaptativos unesp	1º	1º não divulga
Fast search *(Alltheweb)* Consulta em 16/07/2003	sistemas adaptativos e computação inteligente	3º	3º entre 124
	unesp adaptativos	9º	9º entre 77



TAY Consulta em 17/07/2003	sistemas adaptativos e computação inteligente	1º	1º entre 10
Profusion Consulta em 18/07/2003	sistemas adaptativos e computação inteligente	1º	1º com 60%
	sistemas adaptativos	3º	3º com 76%
	unesp adaptativos	1º	1º com 69%
	adaptativos unesp	1º	1º com 60%
	adaptativos	22º	22º com 49%
	redes neurais sistemas adaptativos	7º	7º com 53%
	sistemas adaptativos redes neurais	7º	7º com 53%
Metacrawler Consulta em 18/07/2003	sistemas adaptativos e computação inteligente	5º	5º entre 19
	sistemas adaptativos	14º	14º entre 38
	computação inteligente	15º	15º entre 50
	unesp adaptativos	2º	2º entre 21
	adaptativos unesp	4º	4º entre 19
Lycos Consulta em 18/07/2003	sistemas adaptativos e computação inteligente	3º	3º entre 126
	adaptativos unesp	8º	8º entre 77
	unesp adaptativos	9º	9º entre 77
Radix Consulta em 18/07/2003	sistemas adaptativos e computação inteligente	1º 2º 3º	1º entre 54 2º entre 54 3º entre 54
	sistemas adaptativos	3º 4º 5º	3º entre 485 4º entre 485 5º entre 485
	computação inteligente	1º 2º 3º	1º entre 2452 2º entre 2452 3º entre 2452
	unesp adaptativos	1º 2º 3º	1º entre 33 2º entre 33 3º entre 33
	adaptativos	2º 3º 4º	2º entre 763 3º entre 763 4º entre 763
	redes neurais sistemas adaptativos	7º 8º 9º	7º entre 104 8º entre 104 9º entre 104
	unesp inteligente	1º 2º 3º	1º entre 582 2º entre 582 3º entre 582
	saci marar	1º 2º 3º	1º entre 4 2º entre 4 3º entre 4
	Marar	4º 5º	4º entre 67 5º entre 67

Tabela 10. Identificação dos mecanismos de busca, palavras-chave utilizada e sua posição de classificação.



A tabela 11 mostra o resultado obtido pelos mecanismos de busca em relação às palavra(s)-chave utilizadas nos testes. Deste modo, podemos verificar comparativamente quais são as palavra(s)-chave melhores classificadas e a resposta de cada mecanismo. Paralelamente, no gráfico 2 podemos verificar o desempenho dos mecanismos referentes às 16 palavras-chave utilizadas no teste.

Tabela de resultados obtidos através de palavras chaves

Mecanismos	**Palavra(s)-chave**																**Resultados de palavra(s)-chave encontrada(s)**
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	sistemas adaptativos e computação inteligente	sistemas adaptativos	computação inteligente	saci	adaptativos	unesp adaptativos	computação adaptativos	sistemas unesp	adaptativos unesp	saci marar	unesp intelligente	unesp saci	saci unesp	redes neurais sistemas adaptativos	sistemas adaptativos redes neurais	marar	
Yahoo	x	x	x														3
Cadê	x	x	x														3
AltaVista	x	x	x	x	x	x											6
AcheiAqui	x	x	x	x													4
MSN	x		x														2
Looksmart	x	x	x														3
Ixquick	x																1
Google	x	x	x		x		x	x	x	x							8
América *On Line*	x	x	x		x		x	x	x								7
Excite	x	x	x			x			x		x	x	x				8
Dogpile	x	x	x			x			x								5
Infospace	x	x	x			x			x								5
Webcrawler	x	x	x			x			x								5
Fast search	x					x											3
TAY	x																1
Profusion	x	x			x	x			x					x	x		7
Metacrawler	x	x	x			x			x								5
Lycos	x					x			x								3
Radix	x	x	x		x	x				x	x			x		x	9

Tabela 11. Visualização de palavra(s)-chave encontrada(s)

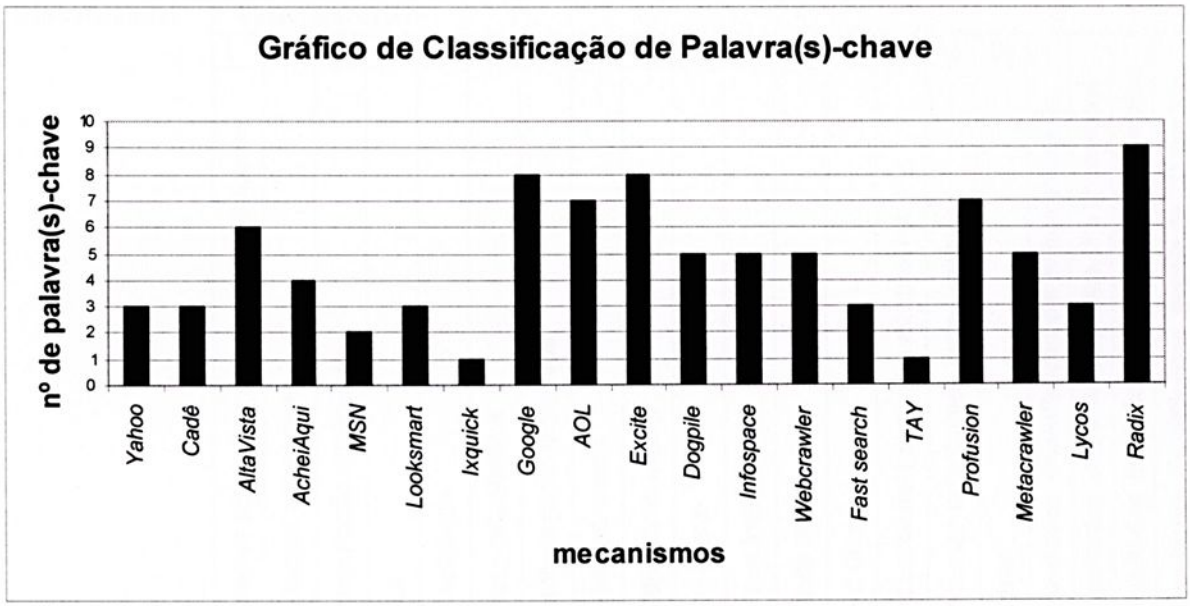


Gráfico 2. Comparativo da indexação do *site* SACI em alguns mecanismos de busca.

A tabela 12 mostra os resultados obtidos em quatro colunas independentes para melhor discriminar os mecanismos de busca e a classificação ocorrida em primeira posição, de segunda à quinta posição, de sexta à décima posição e por último os resultados da classificação ocorridos a partir da 11ª posição.

Tabela de classificação das palavras-chave

Mecanismos	Palavra(s)-chave																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16				
	sistemas adaptativos e computação inteligente	sistemas adaptativos	computação inteligente	saci	adaptativos	unesp adaptativos	computação adaptativos	sistemas unesp	adaptativos unesp	saci marar	unesp inteligente	unesp saci	saci unesp	redes neurais sistemas adaptativos	sistemas adaptativos redes neurais	marar	**Resultados classificados em primeira posição**	**Resultados classificados entre 2ª a 5ª posições**	**Resultados classificados entre 6ª a 10ª posições**	**Resultados classificados depois da 11ª posição**
Yahoo	1º	1º	1º														3	-	-	-
Cadê	1º	1º	1º														3	-	-	-
AltaVista	1º	1º	1º	7º	1º	1º											5	-	1	-
AcheiAqui	1º	1º	1º	8º													3	-	1	-
MSN	1º		14º														1	-	-	1
Looksmart	1º	2º	3º														1	2	-	-
Ixquick	1º																1	-	-	-
Google	1º	3º	5º		6º		2º	17º	1º	2º							2	5	1	1
América *On Line*	1º	5º	16º		8º		2º	20º	1º								2	2	1	1
Excite	4º	8º	8º			1º			3º		2º	3º	4º				1	5	2	-
Dogpile	1º	7º	6º			1º			1º								3	-	2	-
Infospace	5º	9º	10º			1º			1º								2	1	2	-
Webcrawler	5º	9º	8º			1º			1º								2	-	2	-
Fast search	3º					9º											-	1	1	-
TAY	1º																1	-	-	-
Profusion	1º	3º			22º	1º			1º					7º	7º		3	1	2	1
Metacrawler	5º	14º	15º			2º			4º								-	2	-	2
Lycos	3º					9º			8º								-	1	2	-
Radix	1º	3º	1º		2º	1º				1º	1º			7º		4º	5	3	1	-

Tabela 12. Visualização de palavra(s)-chave e sua posição de classificação.

Capítulo 6

6.1 Conclusões e trabalhos futuros..... 133

Capítulo 6

6.1 Conclusões e trabalhos Futuros

A classificação verificada do *site* hospedado no provedor de acesso na Unesp possui diferentes classificações dependendo da base de dados pesquisada e da palavra-chave utilizada.

O *site* "espelho", hospedado em um provedor de acesso gratuito no endereço http://www.sacisistemasadaptativos.hpg.com.br, até o momento (07/10/2003) permanece sem classificação em nenhum mecanismo de busca. Assim, mesmo que teoricamente vários agentes (*spiders*) estejam rastreando o conteúdo da *Web*, esse rastreamento não é suficientemente rápido para acessar e classificar um documento digital que esteja disponibilizado em uma base de acesso sem que se faça o envio do endereço do *site* para uma base de dados.

Assim, mesmo o *site* "espelho" possuindo todos os atributos de identificação como o uso de palavra(s)-chave, *meta-tags* e *tag alt* sendo utilizadas da mesma forma que o *site* hospedado no provedor da Unesp, o resultado de classificação depois de 144 dias é nulo. Contudo isso não quer dizer que o material hospedado no

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

provedor gratuito não poderá vir a ser classificado. Revela-nos, sim, a importância do processo de se fazer o cadastramento do documento digital nas bases de dados, para agilizar o processo de cadastramento e indexação do material desenvolvido.

A partir deste resultados motivadores, nosso trabalho futuro será a implementação de métodos de classificação utilizando redes neurais artificiais e agentes inteligentes para otimizar a busca de informação em sistemas distribuídos de informação.

Também é de meu interesse, desenvolver estudos direcionados para distribuição de informações em sistemas de multimeios e suas características próprias de *design,* usabilidade, arquitetura de informação e interfaces com o usuário.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Referências Bibliográficas

ALBERTINI, Alberto L. **Comércio Eletrônico: Modelos, Aspectos e Contribuições de sua Aplicação**. São Paulo, ed. Atlas, 2001. 280p.

ANDERSON, David P. Kubiatowics, J. **Um computador de abrangência mundial.** *Scientific American* Brasil nº 4. p.54-61, 2002.

BACHIOCHI, D. Berstene, M. Chouinard, E. Conlan, N. Danchak, M. Furey, T. Neligon, C. Way, D. ***Usability and designing navigational aids for the World Wide Web***. *Computer Networks and ISDN Systems*, vol 29, p.1489-1496, 1997.

BERGMAN, Michael K. ***The Deep Web: Surfacing Hidden Value***. *The Journal of Electronic Publishing. The University of Michigan Press*. Vol 7, Issue 1, 2001 Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em: 17 set. 2002.

BLACK, Roger. ***Websites* que funcionam**. São Paulo, ed. Quark, 1997. 237p.

BHARAT, Krishna. **SEARCHPAD: *Explicit capture of search context to support web search***. Computer Networks, vol 33, p.493-501, 2000.

BODNER, Richard C. CHIGNELL, Mark H. CHAROENKITKARN, Nipon. GOLOVCHINSKY, Gene. KOPAK, Richard W. ***The impact of text browsing on text retrieval performance***. *Information Processing & Management,* vol 37, p.507-520, 2001.

BOUERI Filho, J. J. - **Antropometria aplicada à arquitetura, urbanismo e desenho industrial**. São Paulo, FAUUSP,1991. Volume I.

BRICKIN, Meredith. ***Virtual Words: no interface design***. In: Benedikt, Michael: *Cyberspace first steps*. MIT *Press*: Boston 1992, p 363-383.

BRIAN, Sergey. PAGE, Lawrence. ***The anatomy of a large scale hypertextual web search engine***. *Computer Networks and* ISDN *Systems*, vol 30. p.107-117, 1998.

BONSIEPE, Gui. ***Design* do material ao digital**. Florianópolis, ed. Fiesc/Iel, 1997. 191p.

CASANOVA, Marco A, Moura, Arnaldo V. **Princípios de Sistemas de Gerência de Bancos de Dados Distribuídos**. Rio de Janeiro, ed. Campus, 1985. 355p.

CASTRO, Elizabeth. **HTML para a *World Wide Web***. São Paulo, ed. Makron *Books,* 2000. 534p.

CERI, Fraternali S, BONGIO, P. A. ***Web Modeling Language (WebML): A modeling language for designing web sites***. *Computer Networks and* ISDN *Systems*, vol 33, p.137-157, 2000.

CHANG, Yue S. YUAN Shyan M. LO, Winston. ***A new multi search engine for querying data through an internet search service on CORBA***. *Computer Networks*, vol 34, p.467-480, 2000.

CHIANG, Roger H. L. CHUA, Cecil E. H. STOREY, Veda C. ***A smart web query method for semantic retrieval of web data***. *Data & Knowledge Engineering*, vol 38, p. 63-84, 2001.

COUCEIRO, Luiz A. C. C, BARRENECHA, Hugo F. S. **Sistemas de Gerência de Banco de Dados Distribuídos**. Rio de Janeiro, Livros Técnicos e científicos Ed. S.A., 1984. p77.

CRONEY, J. - **Antropometría para diseñadores**. São Paulo, Gustavo Gilli, 1971.

DALAL, N.P. QUIBLE, Z. WYATT, K. ***Cognitive design of home pages: an experimental study of comprehension on the World Wide Web***. *Information Processing and Management*, vol 36, p.607-621, 2000.

DEMPSEY, Bert J. VREELAND, Robert C. SUMMER JR, ROBERT, G. YANG, Kiduk. ***Design and empirical evaluation of search software for legal professionals on the WWW***. *Information Processing & management*, vol 36, p. 253-273, 2000.

DONDIS, Donis A. **Sintaxe da Linguagem Visual**. São Paulo: ed. Martins Fontes, 2000. 234p.

DOTTA, Sílvia. **Construção de Sites**. São Paulo, ed. Global, 2000. 144p.

DUL, J. & WEERDMEESTER, B. - **Ergonomia prática.** São Paulo, Edgard Blucher, 1993. 148p.

EELKO, K.R.E.***The content and desing of web sites: an empirical study***. *Faculty of Economics, Information & Management*, vol 37, p.123-134, 2000.

ERSKINE, Lewis E., TOD, David R. N. CASTER, Burton, JOHN K. ***Dialogical techniques for the design of web sites***. *Int. J. human computer Studies*, vol 47, p.169-195, 1997.

FURGERI**,** Sérgio. **Ensino Didático da Linguagem XML**. São Paulo, ed. Érica, 2001. 278p.

GANASCIA, Jean Gabriel. *L´Intelligence Artificielle.* Paris, Flammarion, 1993.

GANDAL, Neil. ***The dynamics of competition in the internet search engine market***. *International Journal of Industrial Organization*, vol 19, p.1103-1117, 2001.

GARDNER, Howard. **Inteligência um conceito reformulado**. Rio de Janeiro, ed. Objetiva,1999. 347p.

GARRATT, Andrea. JACKSON, Mike. BURDEN, Peter. WALLIS, Jon. ***A survey of alternative designs for a search engine storage structure***. *Information and Software Tecnology*, vol 43, p.661-677, 2001.

GRANDJEAN, E. - **Manual de ergonomia - adaptando o trabalho ao homem**: Porto Alegre, ed. Bookman,1998. 338p.

HALL, E. - **A dimensão oculta**: Rio de Janeiro, ed. Francisco Alves. 1977.

HOLSCHER, Christoph. STRUBE, Gerhard. ***Web search behavior of internet experts and newbies***. *Computer Networks*, vol 33, p.337-346, 2000.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

HORN, Robert E. ***Visual language. Global communication for the 21th century***. *Bainbridge Island*, Washinton, ed. Macro VU press,1998. 270p.

JOHNSON, Steven. **Cultura da interface**. Rio de Janeiro, ed. Jorge Zahar, 2001. 189p.

KLEINROCK, Leonard. ***Creating a mathematical theory of computer networks***. *Institute for Operations Research and the Management Sciences*, vol. 50, No. 1, p.125-131, 2002.

KRUG, Steve. **Não me faça pensar. Uma abordagem do bom senso à navegabilidade da *Web***. São Paulo, ed. *Market Books*, 2001. 187p.

KWOK, Cody. ETZIONI, Oren. WELD, Daniel S. ***Scaling question answering to the web***. *Capes. The Gale Group*. ACM *Transactions on Information Systems*, vol 19, i3, p.242-260, 2001.

LAUFER, R., Scavetta, D. Textos, Hipertextos, Hipermédia. Paris, Presses Universitaries Françaises, 1992. 225p.

LEÃO, Lucia. O Labirinto da hipermídia. São Paulo, Iluminuras,1999. 158p.

LEE, Foina. BRESSAN, Stéphane. OOI, Beng, C. ***Hybrid Transformation for indexing and searching web documents in the cartographic paradigm***. *Information Systems*, vol 26, nº2, p.75-92, 2001.

LÉVY, Pierre. **Cibercultura**. São Paulo, ed. 34, 1999. 260p.

_____,_____. Tecnologias de Inteligência, São Paulo, ed. 34, 1997. 208p.

LUZ, Iraci B. P. **Acesso à informação: um assunto polêmico**. Bauru, 1997. 110p. Dissertação (Mestrado – Comunicação e Poéticas Visuais) – Faculdade de Arquitetura, Artes e Comunicação, Universidade Estadual Paulista.

MARCHIORI, Massimo. ***The quest for correct information on the hyper search engines***. *Computer Networks and* ISDN *Systems*, vol 29, p.1225-1235, 1997.

MCLUHAN, Robert. ***Search for a top ranking. (strategies for getting a page one placement from a search engine).*** *Capes. The Gale Group*, p.47-49, 2000.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

MCLUHAN, Marshall. **Os meios de comunicação como extensão do homem**. ed. Cultrix, São Paulo, 1996. 407p.

MEGHABGHAB, George. ***Interative radial basis functions neural networks as metamodels of stochastic simulations of the quality of search engines in the World Wide Web.*** *Information Processing & Management*, vol 37, p.571-591, 2001.

MINASI, Mark. **Segredos de projeto de interface gráfica com o usuário**. Rio de Janeiro, ed. InfoBook,1994. 223p.

NIELSEN, Jackob. **Projetando *websites. Designing web usability***. Rio de Janeiro, ed. Campus, 2000. 416p.

_____._____ TAHIR, Marie. ***Homepage*: Usabilidade. 50 *websites* desconstruídos**. Rio de Janeiro, ed. Campus, 2002. 315p.

NOTESS, R.G. ***Search Engine Statistics: Relative Size Showdown***. Disponível em <http://www.searchengineshowdown.com/stat/size.shtml>. Acesso em: 03 set. 2003.

POO, Danny C.C. TOH, Tech K. KHOO, Christopher, S,G. ***Enhancing online catalog searches with an electronic referencer***. *The Journal of Systems and Software*, vol 55, p.203-219, 2000.

RADFAHER, Luli. ***Design/web/design2***. São Paulo, ed. Market Press, 2001. 265p.

RISDEN, Kirsten. CZERWINSKI, Mary P. ***An initial examination of ease of use for 2D and 3D information visualizations of web content.*** Int. J. *Human Computer Studies*,) vol 53, p.695-714, 2000.

ROSENSTIEHL, P. Enciclopédia EINAUDI, vol. 13, 1988.

SAVOY, Jacques. PICARD, Justin. ***Retrieval effectiveness on the web***. *Information Processing and Management*, vol 37, p.543-569, 2001.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

SIEGAL, David. **Criando *Sites* Arrasadores na *Web* II**. São Paulo, ed. Quark,1998. 305p.

SILVA, J. **Uso de gabaritos Configuráveis para desenvolvimento de interfaces Virtuais**. Anais do 1º Workshop de Realidade Virtual, São Carlos, SP, 1997, p. 81-90.

SILVEIRA, Marcelo. ***Web Marketing*, Usando Ferramentas de Busca**. São Paulo, ed. Novatec, 2002. 159p.

SHNEIDERMAN, B. ***Desiging information abundant web sites: issues and recommendations***. *Human Computer interaction laboratory, Departament of computer Science & Intitute for Systems Research* Int J. *Human Computer Studies*, vol 47, p.5-29, 1997.

SULLIVAN, Danny. ***Web Pages Tweaking: Will the Best Rise to the Top?*** Danny Sullivan's *Search Engine Strategies* 2000 *conference*. *Capes. The Gale Group. The Information Advisor*, Vol 12, i5 p.4-10, 2000.

SUGIURA, Atsushi. Etzioni, Oren. ***Query routing for Web search engines: arquitecture and experiments***. *Computer Networks*, vol 33, p.417-429, 2000.

THOBIAS, Maria A. L. S. **A *internet* e o ensino de ciências**. Bauru, 2000. 126p. Dissertação (Mestrado - Ensino de Ciências) – Faculdade de Ciências, Universidade Estadual Paulista.

TU, Hsieh Chang. HSIANG, Jieh. ***An architecture and category knowledge for intelligent information retrieval agents***. *Decision Support Systems*, vol 28, p.255-268, 2000.

VAN DAM, A. ***Post WIMP User Interfaces.*** *Communications of* ACM, vol.40, N. 2, Feb. 1997, p. 63-67.

WALKER, Regianald L. ***Search engine case study: searching the web using genetic programming and MPI***. *Parallel Computing*, vol 27, p.71-89, 2001.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Glossário

C

Cache Tipo de memória de alta velocidade que um processador pode acessar mais rapidamente do que a memória principal. Os dados utilizados com freqüência são armazenados na memória *cache*, que se localiza próxima à UCP (Unidade Central de Processamento) e funciona em conjunto com a memória principal.

Cluster A menor unidade de leitura ou gravação em disco do sistema operacional. Seu tamanho é variável, em função do tamanho do disco. Em relação ao texto, uma área de alocação de armazenamento de informações.

D

Design Concepção de um projeto ou modelo; planejamento. O produto desse panejamento.

DOS (*Disk Operating System*) - O DOS - como os outros sistemas operacionais - supervisiona operações como o *Input/Output* de disco, o suporte ao vídeo o controle do teclado e diversas funções internas relacionadas à execução e à manutenção de arquivos. O DOS é um sistema operacional mono-usuário e monotarefa com uma interface de linha de comandos, lançado em 1981 para IBM PCs e equipamentos compatíveis.

E

E-commerce (comércio eletrônico) - Escolha, compra e pagamento de um produto, realizados por meio da *Web*.

F

Frame (quadro/moldura) - Quadro de uma página HTML. Cada *home page* pode ter vários *frames*, sendo cada um deles um documento distinto.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

G

GiF Sigla para *Graphics Interchange Format*. Formato de arquivos de imagens mais utilizado na *Web*. O formato GIF cria arquivos de imagens de tamanho relativamente pequeno em relação aos demais formatos. Graças à essa compactação, é um formato ideal para a utilização na rede. O tipo de compactação utilizada no formato GIF funciona melhor quando a imagem tem áreas contínuas da mesma cor e, principalmente, poucas cores. Por isso, o formato não é muito recomendado para fotos (nesse caso, o formato JPEG é bem mais eficiente).

H

Heurística Metodologia, ou algoritmo, para resolver problemas por métodos que, embora não rigorosos, geralmente refletem o conhecimento humano e permitem obter uma solução satisfatória.

Hipertexto Documento capaz de incluir em seu conteúdo ligações com outras partes do mesmo documento ou com documentos diferentes. As ligações normalmente são indicadas através de uma imagem ou texto em uma cor diferente ou sublinhado. Ao clicar na ligação, o usuário é levado até o texto ligado.

Hyperlink Nome que se dá às imagens ou palavras que dão acesso a outros conteúdos em um documento hipertexto. O *hyperlink* pode levar a outra parte do mesmo documento ou a outros documentos.

I

Imagemaps Imagem com vários *hyperlinks* que levam a destinos diferentes. Existem dois tipos de mapas clicáveis. No modelo mais antigo, o usuário não sabe qual arquivo será descarregado ao colocar o *mouse* sobre determinada região da imagem (a menos que esteja escrito na imagem). A informação que é passada para o usuário enquanto ele movimenta o cursor sobre a imagem é a posição do cursor, indicada em *pixels* na margem inferior do navegador. Quando a imagem é clicada, a informação sobre a localização do clique é enviada ao servidor, que consulta uma tabela para descobrir qual arquivo enviar de volta. No modelo mais recente de mapa clicável, introduzido pelo *Microsoft Internet Explorer* e seguido pelo *Netscape Navigator*, a informação sobre qual arquivo o servidor deve mandar está já na página HTML. Quando o usuário passa o mouse sobre a imagem, o nome do arquivo aparece na margem inferior (barra de *status*) do navegador. Assim é possível receber uma resposta mais rápida do servidor.



J

Javascript Linguagem "interpretada" de criação de *scripts* desenvolvida pela *Netscape* e pela *Sun Microsystems*. É uma linguagem de programação feita para complementar as capacidades do HTML. O código de *JavaScript* é enviado ao cliente como parte do código HTML de uma página, e pode ser utilizado para criar efeitos especiais, como botões animados, sons etc.

JPG Sigla para *Joint Photographic Experts Group*, o nome original do comitê que escreveu o padrão desse formato de compressão de imagens. JPEG foi criado para comprimir imagens tiradas do mundo real. Funciona bem com fotos e desenhos naturalísticos, mas não é tão eficiente com desenhos de letras, linhas e *cartoons*. O formato JPEG permite uma alta compressão das imagens devido ao seu processo de compressão com perdas. Isso significa que a imagem final pode ficar com qualidade pior do que a original.

L

Léxico O vocabulário de uma língua. Exame de cada parte de um todo, tendo em vista conhecer sua natureza, suas proporções, suas funções, suas relações, etc.

Linux Sistema operacional criado pelo finlandês Linus Torvalds. Vem sendo implementado por uma ativa comunidade de programadores em todo o mundo. Integra o *Free Software Movement* (Movimento pelo Software Livre). É grátis e seu código é aberto, ou seja, qualquer pessoa pode modificá-lo.

M

Multimídia O termo multimídia é utilizado para definir um documento de computador composto de elementos de várias mídias, como áudio, vídeo, ilustrações e texto. Também é importante que esses documentos sejam interativos, ou seja, que permitam a participação do usuário. Para ser mais preciso, utiliza-se também o termo multimídia interativa.

P

PDF (*Portable Document Format*) - Formato em que as especificações do arquivo (negrito, itálico, tipo e tamanho de letra) são armazenadas identicamente em qualquer plataforma, sistema operacional e aplicativo.



Pixel Nome dado para *picture element* (elemento de imagem). É a menor área retangular de uma imagem. Cada *pixel* é uma cor diferente. Com essa combinação de cores, é possível mostrar qualquer cor. No entanto, a capacidade de mostrar todas as combinações de cores possíveis vai depender da qualidade e da configuração do monitor do usuário.

Provedores Computador ligado permanentemente à rede, que, entre outras coisas, armazena arquivos e permite o acesso de usuários. Também chamado de nó.

S

Search engine (Mecanismo de busca) Programa que permite pesquisar na *Web* ou em um banco de dados por meio de palavras-chaves em arquivos, documentos ou páginas *Web*. Alguns serviços permitem a pesquisa em vários documentos simultaneamente. São os meta buscadores.

Semântico Estudo das mudanças ou translações sofridas, no tempo e no espaço, pela significação das palavras. O estudo da relação de significação nos signos e da representação do sentido dos enunciados.

Site Conjunto de documentos apresentados ou disponibilizados na *Web* por um indivíduo, instituição, empresa, etc, e que pode ser fisicamente acessado por um computador e em endereço específico da rede.

U

Url (*Uniform Resource Location*) - Padrão de endereçamento da *Web*. Permite que cada arquivo na *Internet* tenha um endereço próprio, que consiste de seu nome, diretório, máquina onde está armazenado e protocolo pelo qual deve ser transmitido. Por isso se diz que cada página da rede tem sua própria URL.

Unix Sistema operacional desenvolvido, em 1969, pela empresa americana AT&T (uma das mais poderosas do mundo na área de telecomunicações), capaz de executar ao mesmo tempo várias tarefas (multitarefa) solicitadas por diferentes usuários simultaneamente (multiusuário). Foi criado para o desenvolvimento das redes remotas, na medida em que a formação de uma "fila"de pedidos para atendimento um a um (como nas primeiras redes locais) tornaria o tráfego da rede extremamente lento.



V

Vetorial Conjunto de n quantidades que dependem de um sistema de coordenadas n-dimensionais e que se transforma segundo leis bem determinadas quando se muda o sistema. Em relação ao desenvolvimento do trabalho, arquivo vetorial é aquele que se adapta melhor as condições desfavoráveis aumentando ou diminuindo seu tamanho.

W

Web (*World Wide Web* ou WWW) - Área da *Internet* que contém documentos em formato de hipermídia, uma combinação de hipertexto com multimídia. Os documentos hipermídia da WWW são chamados de páginas de *Web* e podem conter texto, imagens e arquivos de áudio e vídeo, além de ligações com outros documentos na rede. A característica multimídia da *Web* tornou-a a porção mais importante da *Internet*.

Windows (janela) Sistema operacional introduzido pela *Microsoft* em 1983. Possui ambiente multitarefa e interface gráfica com o usuário.

Webdesign É o grande diferencial entre um *site* concorrido ou uma página esquecida em um servidor ou um provedor de acesso qualquer.Uma página com *webdesign* é feita buscando o melhor para seu usuário , desde velocidade até visual. O aprimoramento de um *site* com *webdesign* deve ser contínuo, com crescente evolução a cada atualização.

Webmaster Nos *sites* de grande porte, com movimentadas aplicações de *e-commerce* e portais de alto tráfego, o cargo acabou desaparecendo. Isso acontece porque grandes empresas possuem equipes técnicas altamente segmentadas, onde cada profissional é responsável por uma parte específica do trabalho. Um *Webmaster* deve possuir conhecimento bastante abrangente. HTML, DHTML e *JavaScript* São premissas básicas. Além disso, é importante ter conhecimentos em tecnologia de banco de dados e algumas linguagens de *script*, como *Perl*, PHP, ASP, *Cold Fusion* etc.

Webwriting Forma de transmitir um conteúdo para *Web* de forma diferente do texto de revista, livro ou jornal, onde se usa o conceito de pirâmide invertida de informação, ao qual, no início do texto já se coloca o motivo do acontecimento, entre outros inúmeros detalhes.



X

XML Acrônimo de *eXtensible Markup Language*. Versão compacta da SGML (*Standard Generalized Markup Language*). Permite que os programadores e projetistas da *Web* criem *tags* personalizadas que porporcionem maior flexibilidade na organização e apresentação das informações do que é possível obter com o antigo sistema de codificação de documentos HTML. No início de 1997, a especificação XML foi publicada em um formato preliminar por um grupo de trabalho do W3C (*World Wide Web Consortium*) e contou com o apoio de várias das principais empresas do setor de informática.

Anexo 1

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Anexo 1

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

Arquitetura de informação para documentos digitais para *Web*

Rodrigo F. CARVALHO
Mestrando Desenho Industrial FAAC Unesp
CTI/Unesp Av. Nações Unidas 58-50, Bauru-SP, Brasil, fone (14) 2302 0161
flash@feb.unesp.br

João F. MARAR
Doutor
Laboratório de Sistemas Adaptativos e Computação Inteligente Dpto de Computação FC – Unesp
Av Dr. Edmundo C. Coube s/n, Bauru-SP, Brasil
fermarar@fc.unesp.br

RESUMO

Este artigo tem com objetivo, revelar funções internas de um documento digital, que podem ser usadas para favorecer a melhor classificação e o acesso de informação mais relevante em sistemas distribuídos de informação. E através de resultados, comprovar a necessidade de construção de documentos digitais não apenas de forma mecânica, mas com toda uma concepção estrutural orgânica. Assim, todos os elementos que estiverem contidos na estrutura dos documentos, poderão oferecer uma possibilidade maior de classificação nos mecanismos de busca. E através dos procedimentos mostrados neste artigo, permitir que as chances de classificação dos *sites,* sejam eles pessoais, comerciais corporativos, educacionais, etc, possam ser rastreados internamente e indexados nos sistemas distribuídos de informação. Adicionalmente, a substituição de *sites* mecânicos por estruturas orgânicas com probabilidade de melhor classificação, torna a utilização dos mecanismos de busca por parte dos usuários uma experiência menos desagradável, em relação ao tempo necessário para a encontrar a resposta da pesquisa, visto que com o crescente número de informações disponibilizadas na *web*, as chances de encontrar uma informação segura são cada vez mais difíceis.

Palavras chave: *webdesign*, *spiders*, classificação, sistemas distribuídos, arquitetura de informação.

1. INTRODUÇÃO

Os diversos algoritmos (procedimentos lógicos) dos mecanismos de busca existentes na *Web* (*World Wild Web*), tais como *Google*, Altavista, *Yahoo,* etc, realizam uma função de catalogar e indexar informações obtidas dos documentos digitais que trafegam nos sistemas distribuídos (como a www), buscando separar em *clusters* (pastas), as informações que possam estar ligadas através de um determinado assunto. Alguns mecanismos realizam esse trabalho de forma mais eficiente que outros. Entretanto, existe muito ruído de informação, os quais influenciam na baixa qualidade das indexações realizadas.

Existe uma forte tendência no ambiente digital distribuído em utilizar recursos da linguagem HTML. Pesquisas [4][6][7] revelam que do total de informações existentes na *Web*, em média, 44% é referente a conteúdo com base em HTML, e o percentual restante é atribuído, a linguagem XML ou *Javascript* e também conteúdo multimídia (filmes, vídeo-clips, animações, música, ruídos, etc,) e hipermídia (os mesmos conteúdos da multimídia com recursos de interatividade). Além de outras formas de conteúdo como PDF(*portable document file*), dados dinâmicos, programas executáveis, planilhas de cálculos, arquivos de textos de diversos formatos, entre outros [1].

Desta forma, quando os atributos de identificação de elementos são utilizados incorretamente ou, não são aproveitados os melhores recursos de uma boa classificação nos mecanismos de busca, o documento digital fica escondido no provedor de acesso, sem servir ao seu propósito, que é o de ser encontrado para utilização e transferência de informação.

Adicionalmente, é importante deixar claro que seja qual for o mecanismo de busca utilizado, a classificação é realizada através da análise de texto[6]. Assim, qualquer conteúdo formal que não seja texto, oferece dificuldade para ser rastreado e classificado nas bases de dados dos mecanismos de busca. Por esse motivo, conteúdos em forma de imagens fixas, filmes, vídeo-clips, animações, sons, ruídos, programas executáveis, entre outros, acabam sendo prejudicados em relação ao seu formato, no que tange à identificação e classificação nos mecanismos de busca. Em sua essência, não podem ser classificados simplesmente pelo material oferecido, justamente porque os métodos de classificação utilizam padrões de análise léxica, semântica, e em alguns casos heurística (método que ao contrário do algoritmo, realiza buscas aleatórias e por tentativa e erro) e, que pela própria natureza dos outros arquivos que não possuam base textual, não podem ser analisados para classificação nas bases de dados[3]. As bases de dados são todas as informações que estão contidas dentro do índice dos mecanismos de busca, e que são atualizadas periodicamente.

2. DOCUMENTOS DIGITAIS MECÂNICOS

De maneira análoga, um *site* pode ser considerado mecânico quando não existe a preocupação inicial de se vincular uma ou mais palavras-chave, nos diversos itens que um *site* possa oferecer para pontuação nos mecanismos de busca e, através dessa

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

metodologia, vincular uma melhor qualidade em relação ao conteúdo do *site,* diminuindo a quantidade de ruídos nas listagens de respostas.

Deste modo, um *site* é mecânico quando não se utiliza uma palavra-chave no título da página (*tag title*), no endereço do *site* (URL), no texto visível (texto que aparece no navegador), nas caixas de textos auxiliares de botões e imagens (*tag Alt*), nos nomes dos diretórios do *site*, nos nomes dos arquivos de imagens, em nomes de arquivos de descarregamento (*downloads*) e todos os outros arquivos que possam ser usados. Ao contrário, deveriam possuir um nome vinculado com a palavra chave do *site* para que possam ser classificados com uma melhor pontuação. A pontuação nos mecanismos de busca é diferente para cada um deles, mas mesmo sendo diferente existem elementos em comum que podem ser usados para atribuir uma melhor pontuação para o *site*. Em contrapartida ao *site* mecânico devemos considerar o que poderíamos chamar de *sites* orgânicos.

Por *sites* orgânicos entendemos, em analogia com um organismo vivo e complexo, onde as suas partes ou órgãos funcionam de modo inter-dependente e sinérgico em relação à sua auto-realização, enquanto neste *site* (orgânico) é o sistema orgânico que leva à otimização nas listagens de classificação dos *search engines*.

3. DOCUMENTOS DIGITAIS ORGÂNICOS

Os *sites* orgânicos são aqueles que vinculam em todo seu conteúdo a palavra chave que poderá dar acesso ao documento digital através de um mecanismo de busca. Desta forma, a palavra chave deve, dentro do possível, ser utilizada de forma lógica dentro de todo código interno do *site,* sem, no entanto, ser repetida de forma aleatória, pois poderá ser detectada pelos algoritmos de inteligência artificial existentes nos mecanismos de busca [4] [1]. E caso o mecanismo detecte a existência de repetição aleatória simplesmente para ampliar a pontuação do *site,* o mecanismo de busca penalizará o documento digital com sua exclusão de seu banco de dados. A repetição aleatória da palavra-chave é considerada *"Spam"* e é rigorosamente combatida pelos mecanismos de busca, pois um dos objetivos destes mecanismos é oferecer um serviço de busca que possa ser o mais preciso possível.

Entretanto, não se deve esquecer que certos mecanismos de busca classificam documentos patrocinados e que podem estar bem classificados nas listagens de busca, independentemente do conteúdo apresentado. Por outro lado, certos mecanismos oferecem uma separação dos dados patrocinados e dos dados que são realmente classificados pelos agentes (rastreadores de informação na *web*).

4. CARACTERÍSTICAS DE BASES DE DADOS

As bases de dados podem oferecer três tipos de serviços ao usuário:

Catálogo: Este tipo de base de dados, funciona como uma lista de *sites* sobre vários assuntos. Um dos mais conhecidos, o *Yahoo* (www.*yahoo*.com.br). Funciona como



catálogo de outros *sites*, nos quais também é possível fazer uma busca por palavra-chave[6].

***Web-robots*:** Este tipo de *site* funciona com palavras-chave, mas usa *spiders* (programas rastreadores de informação) para alimentar seus índices[7].

***Metasearch*:** Trata-se de uma pesquisa dentro dos *sites* de mecanismos de busca. Ao invés de fazer uma consulta em cada um dos principais *sites*, o metabuscador, analisa informações de todos os *sites* de mecanismo de busca, ou em alguns deles, ao mesmo tempo[6].

5. METÁFORA DA CLASSIFICAÇÃO

Para entender melhor o processo dos mecanismos de busca, devemos imaginar muitas pessoas aguardando perguntas em uma sala (o banco de dados). Estas pessoas ficam esperando o mecanismo de busca fazer uma pergunta. Tão logo o mecanismo receba a pergunta, ele a transmite para as pessoas que estão na sala e neste momento, as pessoas da sala começam a conversar entre si, para saber quem é que possui alguma informação que possa satisfazer a pergunta. A pessoa que reconhecidamente responder com mais propriedade e conhecimento de causa, terá a resposta melhor posicionada em relação à pergunta. Se, entretanto não houver pessoas experientes na sala, será realizada uma votação entre algumas das pessoas que possuem algum conhecimento sobre o assunto. Neste caso, o resultado deste procedimento pode acabar desapontando a resposta à pergunta formulada. Está é a ação que provoca o protocolo de comunicação entre o mecanismo de busca e os agentes rastreadores de informação.

6. MECANISMOS DE BUSCA COM BANCO DE DADOS DE ÍNDICE

Os elementos classificados pelos mecanismos de busca são armazenados em um banco de dados de índice, com o objetivo de facilitar a consulta do usuário através de palavras-chave.
Este tipo de mecanismo de busca utiliza um *software* conhecido como *spider* (agentes rastreadores de informação), responsável pela visita e indexação das páginas da *Web*. O *spider* visita cada página e retira o conteúdo a ser indexado de cada documento, armazenando-o no banco de dados de índice, de forma adequada para sua eficiente recuperação.

7. MECANISMOS DE BUSCA SEM BANCO DE DADOS DE ÍNDICE

Uma outra alternativa de mecanismo de busca são os meta-buscadores ou *meta searches.* A diferença fundamental entre os mecanismos de busca com banco de dados de índice e o meta-buscador, é que o serviço oferecido por este não possui um banco de dados próprio. Assim, o meta buscador acessa o banco de dados de outros mecanismos de busca, reunindo informações de vários outros mecanismos e seleciona, dentro de critérios próprios, opções para formar sua listagem de elementos que



satisfaçam a requisição efetuada pelo usuário. Desta forma, ao invés de pesquisar em apenas um mecanismo de busca, ou individualmente fazer a mesma pesquisa usando vários mecanismos um de cada vez, o meta-buscador faz este trabalho, realizando a consulta do usuário em vários mecanismos e, retornando dados mais relevantes dentro do critério estipulado por ele (meta-buscador).

8. PROPRIEDADES DOS MECANISMOS DE BUSCA

As propriedades dos mecanismos de busca podem variar muito de um para outro. De forma geral, estão sempre classificando os documentos digitais através de algum elemento relacionado às meta *tags* (códigos específicos, passíveis de leitura pelos agentes), códigos do HTML, e texto visível. Cada um dos mecanismos apresenta suas características próprias com relação à forma de classificação de sua base de dados. Entretanto, todos utilizam elementos semelhantes entre eles para realizarem a sua classificação. Entre estes elementos comuns podem ser usados a *tag title* (título de identificação de página no navegador), a *meta tag description* (descrição concisa do conteúdo do *site*), *meta tag keyword* (palavras-chave que remetem ao conteúdo interno) o texto visível da página e a denominação da *tag alt* (atributo textual de identificação) para elementos não textuais. Assim, a diferença entre os mecanismos de busca estará na forma em que seu programa de identificação trabalha, qual é a quantidade de caracteres lidos em cada *tag,* quais são as *tags* de classificação do mecanismo e se o mecanismo é de rastreamento profundo ou raso. Apesar dos itens de classificação de cada mecanismo serem muito variados, a grande parte dos mecanismos de busca classifica, basicamente, com os mesmos atributos, diferenciando o número de elementos escolhidos, a quantidade de caracteres lidos e o algoritmo usado para classificação.

Desta forma, é possível conseguir uma boa classificação do documento digital em relação ao mecanismo de busca, pelo conhecimento da estrutura da base através de qual se pretende enviar o documento (*Google*, Altavista, *yahoo,* entre outros). Assim, se os atributos de classificação do documento digital forem usados de forma correta, poderá acarretar boa classificação em outros mecanismos, levando em consideração que os atributos de classificação que se pode inserir nos documentos digitais são sempre os mesmos.

9. OTIMIZANDO A CLASSIFICAÇÃO DOS DOCUMENTOS DIGITAIS

O diagrama da Figura 1, ilustra a possibilidade de um documento digital ser classificado nos bancos de dados, desde que, importantes recomendações para melhoria na qualidade da classificação de documentos digitais, descritas abaixo, sejam seguidas:

1. Utilizar a *tag alt* corretamente, para que se possa somar pontos na classificação do mecanismo de busca. Para se fazer uso correto da *tag alt*, é necessário vincular a palavra-chave do *site* com o nome do item ou uma pequena descrição do que será



encontrado na outra ponta do *link* (nós de vinculação interligados no hipertexto e hipermídia), isso se o elemento for um *link.*

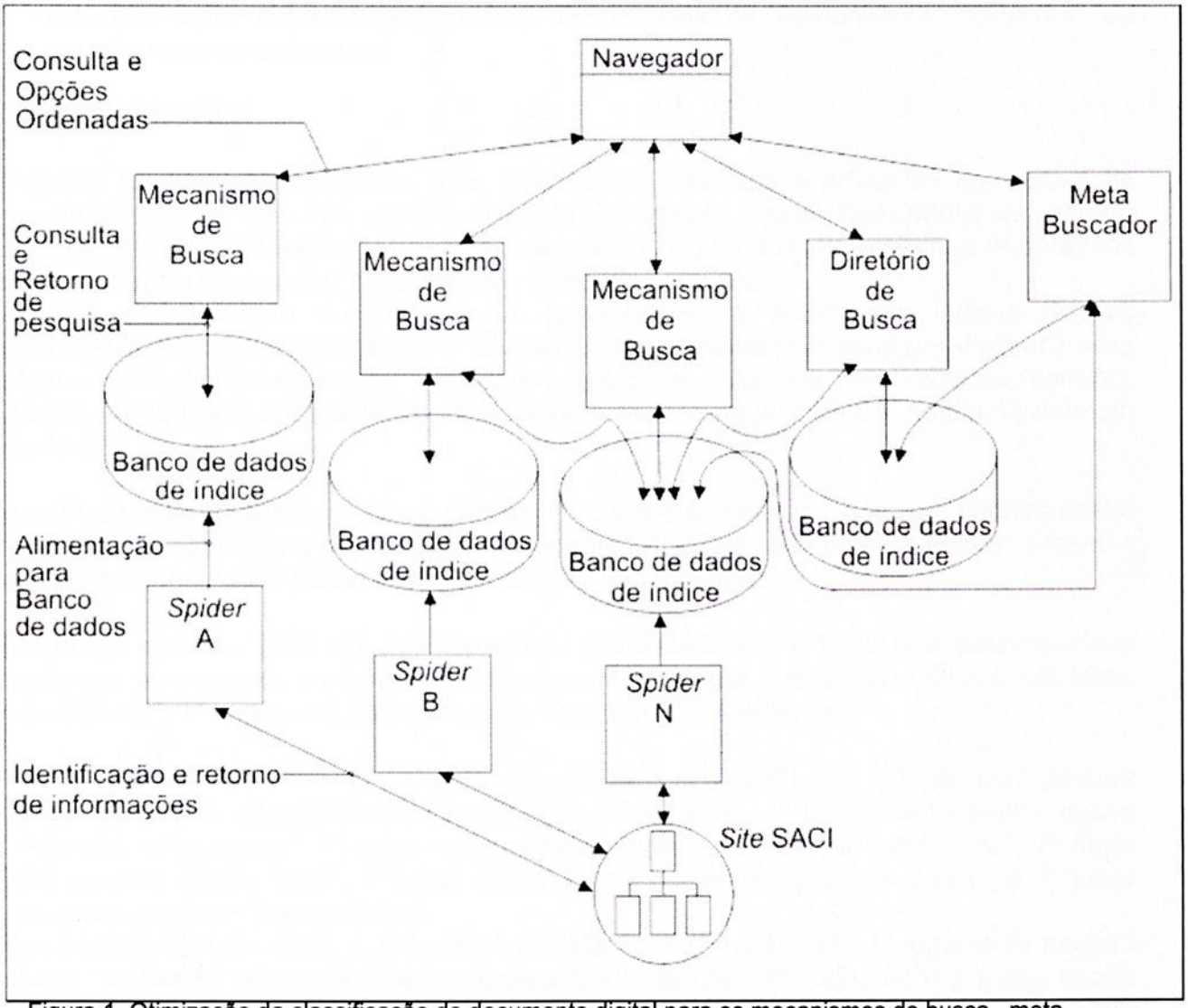


Figura 1. Otimização da classificação do documento digital para os mecanismos de busca , meta buscadores e diretórios.

2. Quanto mais próximo o *link* da informação desejada, mais pontos são oferecidos para a classificação do *site*. Assim, deve-se evitar posicionar a informação em camadas de *links* internos, fazendo o usuário clicar várias vezes. Como regra geral sobre *links,* deve-se estabelecer uma distância máxima de três cliques, quando possível. O objetivo é tornar a informação o mais próxima do usuário, pois que, um documento digital que oferece uma informação com um caminho muito longo não é bem pontuado.

3. Os *Links* que estão no documento digital que apontem para outros *sites* bem acessados e que possuam ligações com o material oferecido, pode aumentar a

probabilidade de classificação. Outros *sites* que possam apontar para o documento digital também oferecem pontuação na classificação, se o *site* que aponta for de uma grande empresa, ou de um *site* que tenha um grande índice de acessos, pode acarretar uma classificação melhor. Nem todos os mecanismos possuem, no entanto essa característica.

10. CONCLUSÃO

Estudos experimentais revelam que, as técnicas descritas e adotadas aumentam as possibilidades de uma boa classificação do documento digital. Realizamos um estudo através do *site* http://wwwp.fc.unesp.br/~fermarar/ , do laboratório Sistemas Adaptativos e Computação Inteligente, acessível em 1º de maio de 2003.
Através da utilização de padrões já comentados na elaboração interna de um documento digital, foi submetido o *site* SACI (http://wwwp.fc.unesp.br/~fermarar/) para alguns mecanismos de busca e diretórios tais como: Altavista (www.altavista.com.br), *Yahoo*, Cadê (www.cade.com.br), Acheiaqui (www.acheiaqui.com.br) no dia 13 de maio de 2003.

Em 21 de maio de 2003, o *Yahoo* classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente", 1º lugar com palavra chave "sistemas adaptativos" e 1º com palavra chave "computação inteligente".

Em 26 de maio de 2003, o Cadê classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente", 1º lugar com palavra chave "sistemas adaptativos", 1º lugar com palavra chave "computação inteligente".

Em 30 de maio de 2003, o Altavista classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente", 1º lugar com palavra chave "sistemas adaptativos", 1º lugar com palavra chave "computação inteligente", 7º lugar com palavra chave "saci", 1º lugar com palavra chave "adaptativos unesp" e 1º lugar com palavra chave "adaptativos".
Em 04 de junho de 2003, o Acheiaqui classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente", 1º lugar com palavra chave "sistemas adaptativos", 1º lugar com palavra chave "computação inteligente" e 4º lugar com palavra chave "saci".

Em 07 de junho de 2003, o MSN (www.msn.com), classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente" e em 14º lugar com palavra chave "computação inteligente".

Em 09 de junho de 2003, o *Ixquick* (www.*ixquick*.com) classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente".

Em 09 de junho de 2003, o *Looksmart* (www.*loocksmart*.com) classificou o *site* SACI em 1º lugar com palavra chave "sistemas adaptativos e computação inteligente", 2º lugar com palavra chave "sistemas adaptativos" e 3º lugar com palavra chave "computação inteligente".



Vale lembrar que os mecanismos que foram utilizados para cadastrar o *site* SACI foram apenas o Altavista, *Yahoo*, Cadê e Acheiaqui, e a classificação obtida nos mecanismos MSN, *Looksmart* e *Ixquick* foram classificadas através da utilização de acesso a bancos de dados externos, comprovando desta maneira o acesso que outros mecanismos fazem a partir de mecanismos que possuam uma grande base de dados.

Adicionalmente, pode-se notar que, a indexação do *site* nos mecanismos de busca foram excelentes, em se tratando do tempo de exposição do *site* e a classificação do mesmo.

Entretanto, mesmo seguindo essas recomendações, há no mínimo duas variáveis que estão fora de controle ao responsável pelo documento digital, que são: o algoritmo de classificação dos mecanismos de busca, que pode ser alterado sem prévio aviso e, a inclusão de novos *sites* que possuam melhores referências de palavras-chave e códigos internos.

Portanto, um documento digital que possa estar bem classificado em um momento, poderá não estar em outro. É por isso que, todo documento digital precisa de manutenção periódica, para evitar seu desaparecimento repentino e/ou sua substituição por outro. Da mesma forma, que um *site* pode ser projetado para estar bem classificado, outros *sites* criados posteriormente, seguirão os mesmos procedimentos para alcançarem uma boa colocação o que significa que quanto mais fatores de boa classificação forem utilizados, melhores benefícios alcançarão os usuários dos sistemas distribuídos.

11. REFERÊNCIAS

[1]BERGMAN, Michael K. (2001) *The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing. The University of Michigan Press*. Vol 7, Issue 1,Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em 17 de setembro de 2002.

[2]FURGERI, Sérgio. (2001) Ensino Didático da Linguagem XML. ed. Érica, São Paulo, 278p.

[3]KWOK, Cody. ETZIONI, Oren. WELD, Daniel S. (2001) *Scaling question answering to the web. Capes. The Gale Group. ACM Transactions on Information Systems*, vol 19, i3, pp.242-260.

[4]NIELSEN, Jackob. (2000) Projetando websites. *Designing web usability*. Ed. Campus, Rio de Janeiro, 416p.

[5]RADFAHER, Luli. (2001) *Design/web/design2*. Ed. Market Press, São Paulo, 265p.

[6]SILVEIRA, Marcelo. (2002) *Web Marketing*, Usando Ferramentas de Busca. ed. Novatec, São Paulo, 159p.

cm 1 2 3 4 5 6 7 unesp 11 12 13 14 15 16 17 18

[7]SULLIVAN, Danny. (2000) *Web Pages Tweaking: Will the Best Rise to the Top? Danny Sullivan's Search Engine Strategies 2000 conference. Capes. The Gale Group. The Information Advisor*, Vol 12, i5 pp.4-10.

Agradecimento

Este trabalho contou com o suporte financeiro da Fapesp processo 97/13309-5.

Anexo 2

Anexo 2

Interfaces inteligentes para mecanismos de busca: Design de informação para Web

Smart interfaces for searching engines: information design for the Web

Rodrigo Ferreira de Carvalho, João Fernando Marar

Palavras-chave: webdesign, spiders, classificação, sistemas distribuidos

Resumo

Os diversos algoritmos dos mecanismos de busca existentes na web realizam uma função de catalogar e indexar informações obtidas dos documentos digitais que trafegam nos sistemas distribuídos, buscando separar em clusters as informações que possam estar ligadas através de um determinado assunto. Alguns mecanismos realizam esse trabalho de forma mais eficiente que outros. Entretanto, existe muito ruído de informação. Este artigo tem com objetivo revelar funções internas de um documento digital que podem ser usadas para favorecer a melhor classificação e o acesso de informação mais relevante.

Key word: Webdesign, spiders, classification, retrieval systems

Abstract

The several algorithms of search engines existents in the web make a function to catalog and index informations captured of digital documents in which traffic in distributed system, separating in clusters the informations that can be connected through a determined subject. Disappear engines make this work in lives her efficient way than others. However, a lot noise of information exists. In this paper have like a main goal to reveal internal functions of the digital document that can to be used to favor the best classification and the access to the relevant information.



1 Introdução

Existe uma forte tendência no ambiente digital distribuído em utilizar recursos da linguagem HTML, em Bergman, (2001) pesquisas revelam que do total de informações existentes na *Web* em média 44% é referente a conteúdo com base em HTML, e o restante é atribuído, por exemplo, a linguagem XML ou *Javascript* e também conteúdo multimídia como filmes, animações, músicas, além de outras formas de conteúdo como PDF, dados dinâmicos, programas executáveis, planilhas de cálculos, arquivos textos de diversos formatos, etc.

Desta forma, quando os atributos de identificação de elementos são utilizados incorretamente ou não são utilizadas as chances de uma boa classificação nos mecanismos de busca são eliminadas. E o documento digital fica escondido no provedor de acesso, sem servir ao propósito de ser encontrado para utilização e transferência de informação.

Além disso, é importante deixar claro que seja qual for o mecanismo de busca utilizado, a classificação é realizada através da análise de texto (Silveira, 2002, p.30). Assim, qualquer elemento que não seja texto oferece dificuldade para ser rastreado e classificado nas bases de dados dos mecanismos de busca. Por esse motivo, elementos como por exemplo: imagens, filmes, animações, sons, programas executáveis, etc, acabam sendo prejudicados em relação ao seu formato para que possam ser identificados e classificados nos mecanismos de busca, pois em usa essência não podem ser classificados simplesmente pelo material oferecido, justamente porque os métodos de classificação utilizam padrões de analise semântica, léxica e em alguns casos heurística, e que pela própria natureza dos outros arquivos que não possuem base textual não podem ser analisados para classificação nas bases de dados. (Kwok et al., 2001).

Na verdade, se o elemento não textual deixar de possuir identificação apropriada,



não será motivo para que o documento digital deixe de funcionar. Entretanto, será um elemento a menos que não colaborará para melhor classificação do material como um todo. E em alguns casos, esse elemento não textual pode ser a chave de acesso ou a entrada para o documento digital para o usuário que possa estar procurando alguma informação, não pelo nome do *site*, mas pelo nome de um possível elemento interno. Desta forma, realizando a entrada no documento digital não pela página inicial, mas por um acesso paralelo, o que deve ser uma possibilidade para favorecer o usuário em relação à informação procurada (Sullivan, 2000).

Outros itens que não podem ser classificados são os dados oriundos de bases de dados dinâmicas, pois são dados criados através de informações personalizadas pelo usuário no instante de sua utilização e as páginas são criadas no momento da consulta, não existindo anteriormente para classificação prévia.

Com relação às bases de dados, elas podem ser gerais e normalmente grandes, como por exemplo o *Google* e o Alta Vista, ou específicas, direcionadas a determinados assuntos e em certos casos com acesso restrito, como por exemplo: *Web of Knowledge* (http://isiknowledge.com), anteriormente conhecido com *Web of Science* e o Probe (www.probe.br).

2 Características de Bases de Dados

As bases de dados podem oferecer três tipos de serviços ao usuário:

Catálogo: Este tipo de base de dados funciona como uma lista de *sites* sobre vários assuntos. Um dos mais conhecidos é o *Yahoo*. Funciona como catálogo de outros *sites* também é possível fazer uma busca por palavra-chave.

***Web-robots*:** Este tipo de *site* funciona com palavras-chave, mas usa *spiders* para alimentar seus índices.

cm 1 2 3 4 5 6 7 unesp 1 12 13 14 15 16 17 18

***Metasearch*:** é uma busca dentro dos *sites* de busca. Ao invés de realizar uma pesquisa em cada um dos principais *sites,* o *metasearch* busca em todos os *sites* (ou em alguns deles) ao mesmo tempo.

3 Metáfora da Classificação

Para entender melhor o processo do mecanismo de busca, devemos imaginar muitas pessoas em uma sala (o banco de dados). Estas pessoas ficam esperando o mecanismo de busca fazer uma pergunta. Tão logo o mecanismo receba a pergunta, ele a transmite para as pessoas que estão na sala, neste momento, as pessoas na sala começam a conversar entre si para saber quem possui alguma informação que possa satisfazer a pergunta. A pessoa que responder com mais propriedade e conhecimento de causa será a resposta melhor posicionada para responder a pergunta. Mas, se não houver pessoas experientes na sala, as pessoas farão uma votação entre algumas pessoas que possuem algum conhecimento sobre o assunto, entretanto, esse procedimento pode acabar desapontando a resposta à pergunta formulada.

4 Mecanismos de Busca com Banco de Dados de Índice

Os elementos classificados pelos mecanismos de busca são armazenados em um banco de dados de índice, com o objetivo de facilitar a consulta do usuário através de palavras-chaves.

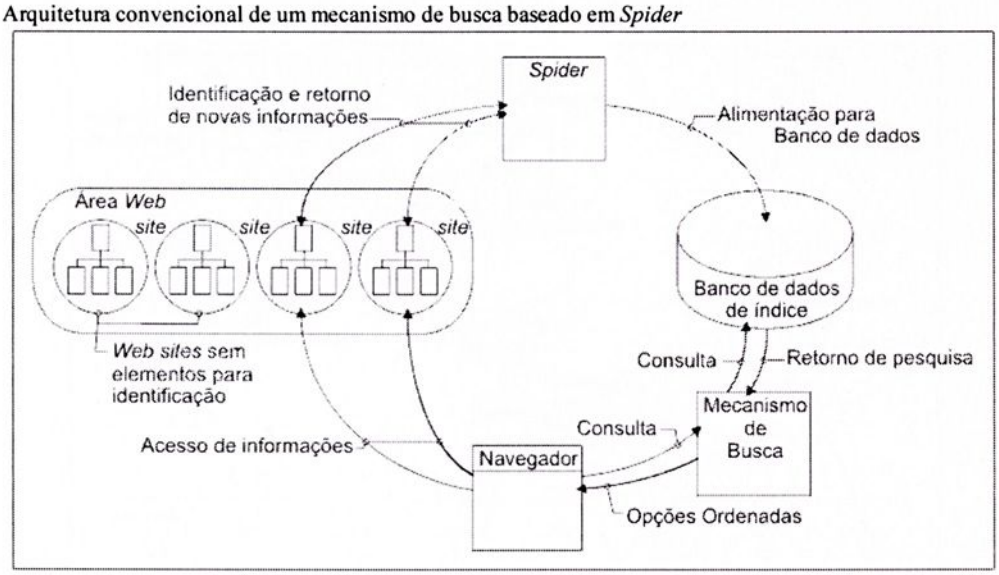
Este tipo de mecanismo de busca utiliza um *software* conhecido como *spider* responsável pela visita e indexação das páginas da *Web*. O *spider* visita cada página e retira o conteúdo a ser indexado de cada documento, armazenando-o no banco de dados de índice de forma adequada para sua eficiente recuperação. É importante salientar que no banco de dados de índice não estão armazenados os documentos na sua integra, (mesmo porque a capacidade de armazenamento teria que ser muito grande), mas sim, algumas palavras-chave e outros dados que sejam julgados necessários como a posição da palavra no documento, sua URL de origem, etc.



A consulta ao índice do mecanismo de busca é efetuada através do navegador do usuário, no *site* do mecanismo de busca. O mecanismo de busca analisa a consulta e retorna de forma ordenada, as referências aos documentos indexadas por seu spider que satisfaçam a consulta. Normalmente, além da referência, é apresentada uma pequena descrição do documento (retirada das primeiras linhas do mesmo), seu título e tamanho.

A arquitetura convencional deste tipo de mecanismo de busca está apresentada na Figura 1.

Figura 1. Arquitetura convencional de um mecanismo de busca baseado em *Spider*

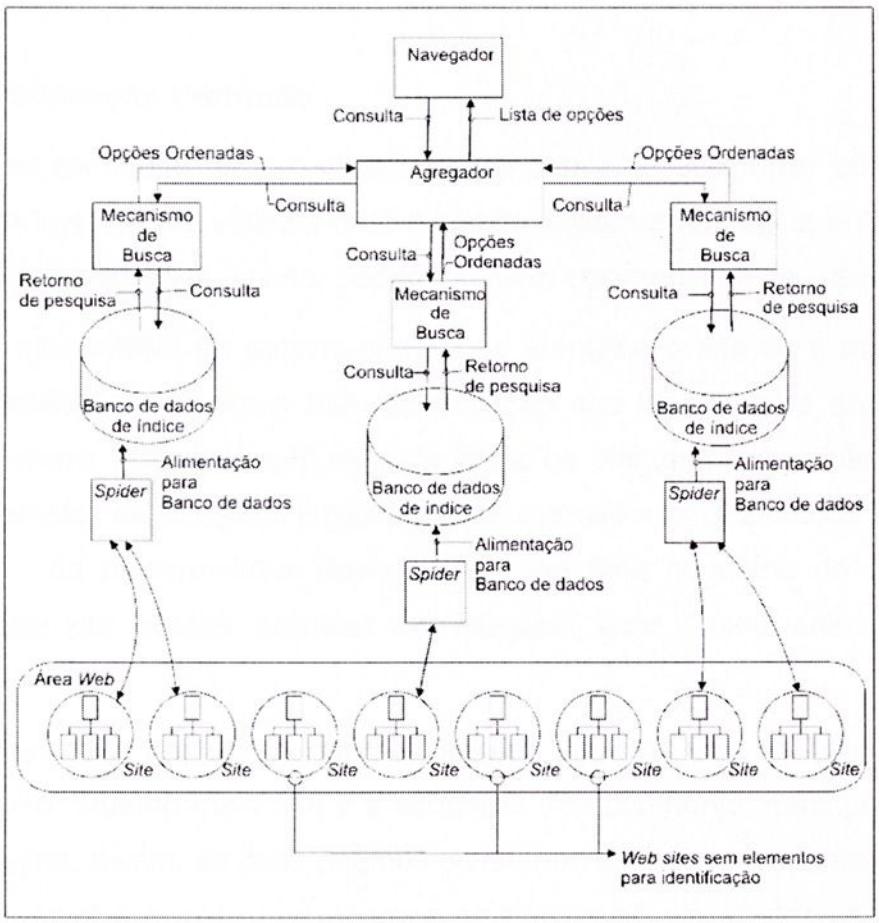


5 Mecanismos de Busca sem Banco de Dados de Índice

Uma outra alternativa de mecanismo de busca é o meta-buscador ou *meta-search.* A diferença fundamental entre os mecanismos de busca com banco de dados de índice e o meta-buscador, é que o serviço oferecido pelo meta-buscador não possui um banco de dados próprio. Assim, o meta-buscador acessa o banco de dados de outros mecanismos de busca reunindo informações de vários mecanismos e selecionando dentro de critérios próprios, opções para formar sua listagem de elementos que satisfaçam a requisição efetuada pelo usuário. Desta forma, ao invés de pesquisar em

apenas um mecanismo de busca, ou individualmente fazer a mesma pesquisa usando vários mecanismos um de cada vez, o meta-buscador faz esse trabalho, realizando a consulta do usuário em vários mecanismos e retornando dados mais relevantes dentro do critério estipulado pelo meta-buscador. A estrutura básica do meta-buscador pode ser vista na Figura 2.

Figura 2. Arquitetura convencional de um meta-buscador.



Entre outras situações, o meta-buscador pode ser usado para fazer uma pesquisa de forma mais rápida do que usando um ou vários mecanismos de busca, uma consulta de preços de livros, no qual o meta-buscador pode listar várias opções relacionadas aos *sites* que possam estar vendendo um determinado livro, assim, agilizando o processo de pesquisa.

Como resposta, o meta-buscador trás uma seqüência de opções em que se pode escolher o mesmo título pela empresa ou pelo valor, podendo ser comparado antes da compra, entre as opções oferecidas. Para fazer essa mesma operação utilizando um mecanismo de busca, seria necessário entrar em vários *sites* e realizar a pesquisa individualmente em cada livraria virtual existente. Adicionalmente, também seria necessário previamente conhecer o endereço de cada livraria virtual.

6 Método de Classificação Profundo

Quando a base de dados for classificada como profunda, quer dizer que o *spider* do mecanismo de busca pode verificar todo o código existente na página e no *site*. Assim, todos os elementos textuais ou não, podem oferecer oportunidade de classificação.

Entretanto, o uso correto da palavra-chave que identifica o *site* ou o material do *site* possui fator fundamental para a boa classificação nas listagens de busca. Assim, é essa palavra-chave que deve ser repetida entre os elementos que não são textuais como por exemplo, as imagens ilustrativas, os marcadores gráficos,os botões e etc. Essa repetição da palavra-chave pode e deve ser feita no nome do elemento não textual (por exemplo: botões, arquivos de imagens, sons, filmes, animações etc), e também na *tag Alt.*

Outra característica muito importante é planejar em qual lugar da página a palavra-chave aparecerá. Quanto mais alta e a esquerda do documento, mais pontuação será atribuída à página. Assim, se duas páginas possuírem as mesmas informações no texto visível (texto visível é o texto que aparece na página no navegador), mas a primeira concentra a palavra-chave o mais alto e a esquerda, e a segunda coloca a palavra-chave em qualquer outro lugar diferente das proximidades do lado superior esquerdo, a primeira página será classificada na frente da segunda. Esse procedimento classifica a palavra-chave mais alta e a esquerda com mais importância do que texto da segunda opção.

unesp

7 Método de Classificação Rasa

Quando a base de dados for classificada como rasa, significa que o mecanismo de busca rastreia apenas algumas partes do código da página ou do *site*. Entre outros detalhes, se uma página utiliza muitos códigos adicionais como, por exemplo, *Javascript*, e se for uma quantidade muito grande, essas linhas de código adicionais podem acabar atrapalhando o rastreamento de classificação, pois os mecanismos rasos irão ler apenas uma quantidade pequena de linhas para fazer sua classificação.

Outras considerações ao método de classificação rasa podem ser feitas em relação ao uso da estrutura de *Frames.* Quando se utiliza *Frames* é possível visualizar várias páginas ao mesmo tempo, utilizando para isso, o mesmo espaço de tela oferecida pelo navegador. Esse procedimento pode ser útil em alguns projetos, quando, por exemplo, certos *frames* podem ficar inalterados para a exibição de *links* de navegação, e uma outra área maior pode alterar seu conteúdo em relação ao item selecionado.

Entretanto, de forma geral, para a classificação nas bases de dados rasas e profundas o uso de *frames* não é recomendado, pois quando se utiliza *frames*, os mesmos dividem a área de navegação para chamar o conteúdo correspondente de cada parte dividida, e os códigos de divisão dos *frames* apenas possuem como códigos ou atributos de identificação o nome do arquivo que será chamado para ser mostrado na parte destinada ao *frame* em questão. Essa característica de montagem prejudica a leitura dos mecanismos rasos, pois os mesmos não continuam seguindo a estrutura de ligação que os *frames* indicam, e algumas características de mecanismos de busca rasos ou profundos não aceitam a classificação de páginas que utilizam *frames*. (Sullivan, 2000; Nielsen, 2000; Radfaher, 2001).

8 Propriedades dos mecanismos de busca

As propriedades dos mecanismos de busca podem variar muito entre um e outro, de forma geral estão sempre classificando os documentos digitais através de algum elemento relacionado às meta *tags*, códigos do HTML, e texto visível. Cada um dos mecanismos apresenta suas características próprias com relação à forma de classificação de sua base de dados, entretanto, todos utilizam elementos semelhantes



entre eles para realizarem a sua classificação, entre os elementos comuns podem ser usados a *tag title*, as *meta tags description* e *keyword* o texto visível e a denominação da *tag alt* para elementos não textuais. Assim, a diferença entre os mecanismos de busca estará na forma em que seu programa de identificação trabalha, qual é a quantidade de caracteres lidos em cada *tag,* quais são as *tags* de classificação do mecanismo e se o mecanismo é de rastreamento profundo ou raso. Apesar dos itens de classificação de cada mecanismo serem muito variados, a grande parte dos mecanismos de busca classifica basicamente com os mesmos atributos, diferenciando o número de elementos escolhidos a quantidade de caracteres lidos e o algoritmo usado para classificação.

Desta forma, é possível conseguir uma boa classificação do documento digital em relação ao mecanismo de busca, através do cohecimento da estrutura da base que se pretende enviar o documento. Assim, se os atributos de classificação do documento digital forem usados de forma correta, poderá acarretar boa classificação em outros mecanismos, levando em consideração que os atributos de classificação que se pode inserir nos documentos digitais são sempre os mesmos.

9 *Meta tags*

Quando Tim Berners-Lee criou o HTML e o conceito de *World Wide Web* juntamente com Robert Cailliau no início da década de 1990 no CERN (Organização Européia para Pesquisas Nucleares, localizada entre a França e a Suíça), resultado dos esforços em criar uma linguagem para definição da estrutura de um documento que pudesse ser conectado a outros, utilizando-se dos conceitos de hipertexto e hipermídia, um dos recursos criados foi à possibilidade de catalogação das páginas que compunham um *site* para busca posterior. (Furgeri, 2001, p33)

Para isso, especificou uma série de *tags* chamadas de *meta tags*, pois são *tags* de informação sobre a própria página onde estão inseridas. Essas *meta tags* formam um dos tópicos talvez mais incompreendidos e interessantes dentre os recursos da linguagem HTML.

unesp

De acordo com a especificação oficial da HTML 4.01 (a última versão), que pode ser obtida no *site* do W3 *Consortium* (http://www.w3c.org), as *meta tags* devem ser inseridas na seção iniciada pela *tag* <*head*>. Não faz diferença se serão escritas antes ou depois do título da página.

10 Importância das *Meta tags*

Existem vários recursos disponíveis com o uso das *meta tags*, mas os principais relacionam-se com os *sites* de busca baseados em *search engines* como o AltaVista, o RadarUOL, o *Google* o *Lycos* e vários outros. Basicamente, o que Berners Lee pretendeu foi definir uma forma de criar um banco de dados com endereços (URLs) de páginas que pudessem ser acessados através de alguma espécie de software de busca.

Algumas outras *meta tags* influenciam a forma como a página será lida, se será atualizada e com que periodicidade, também influindo na forma como a página é armazenada no *cache* do navegador, ou se não será armazenada.

Cada mecanismo de busca possui métodos próprios para cadastrar as informações das páginas do *site*, mas a maioria deles costuma pelo menos catalogar o conteúdo da *tag* <*title*> e as *meta tags Description* e *Keywords*. Alguns mecanismos de busca podem também catalogar os textos alternativos de imagens (atributo *alt* da *tag* <*img*>).

11 *Tag Title*

A *tag title* deve ser utilizada para identificar sem sombra de dúvida o *site* ou o material que será encontrado no documento digital. No código HTML a *tag title* é uma das primeiras *tags* a serem lidas pelos mecanismos de busca. Para o título da página ou *site* pode ser usado até 85 caracteres. Para os mecanismos de busca, a *tag title* tem grande importância, pois será analisada e comparada com a palavra procurada em uma pesquisa. Alguns títulos usam caracteres especiais com seqüências de sinais de maior ou outros símbolos antes da palavra ou frase de título, essa técnica de decoração



parece inofensiva, mas pode comprometer a classificação do documento digital na pontuação da listagem de resposta do mecanismo de busca.

12 *Tag Alt*

A *tag alt* é outro recurso que possui alto poder de aumentar a classificação de um documento se for utilizado de forma adequada. E a forma adequada de usar a *tag alt* é primeiro identificar qual é a palavra-chave do *site*. Segundo, identificar o que faz ou o que é o item, e reunir os dois itens no texto da *tag alt*. É possível também identificar apenas o que é ou o que faz o elemento, mas desta forma a classificação não terá um vinculo com a palavra-chave selecionada.

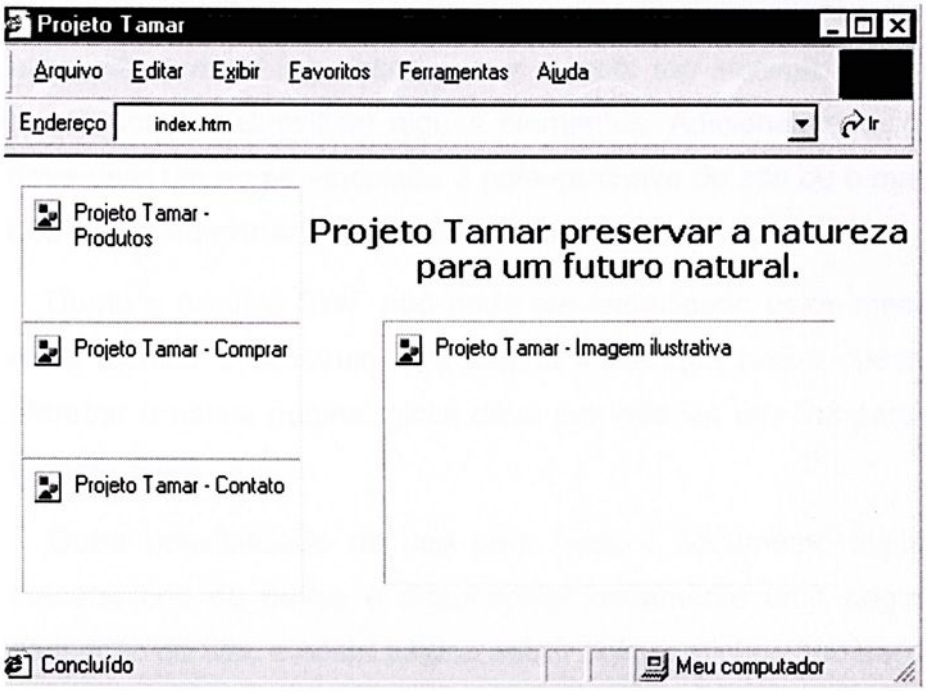
Entretanto, a *tag alt* não se destina a ser utilizada apenas para tornar um documento mais bem classificado. Assim, é possível operacionalizar com a *tag alt* o conceito de usabilidade, no qual é possível transmitir para o usuário o que será encontrado do outro lado do *link*, sem que o mesmo tenha que efetuá-lo. Para que esse recurso funcione adequadamente é necessário descrever com poucas e adequadas palavras o que se poderá encontrar do outro lado do *link* utilizando a *tag alt*.

Outro fator de usabilidade é a possibilidade do conteúdo ser processado pelo navegador, transformando o conteúdo da *tag alt* em som e assim, permitir o uso de usuários com deficiência visual.

E por último, em alguns casos, determinados usuários desconsideram a navegação com imagens, para ganhar tempo no descarregamento dos elementos do documento digital, desta forma com o uso da *tag alt*, os lugares que usam imagens apresentam o texto explicativo para que mesmo sem a imagem, o usuário consiga navegar no *site*. Adicionalmente, alguns usuários por preferência, utilizam navegadores totalmente textuais e que não apresentam nenhum tipo de elemento gráfico como por exemplo, o navegador *Lynxs*. Desta forma, para que se possa utilizar este tipo de navegador, com um documento construído usando elementos gráficos com itens navegacionais, é de extrema importância o uso da *tag alt* como identificador. Figura 3.



Figura 3. Visão de como a *tag alt* pode facilitar a navegação para usuários que configuram o navegador para não apresentar imagens.



13 Identificação de elementos não textuais e mecanismos de busca

Os mecanismos de busca classificam o conteúdo dos documentos digitais exclusivamente através de texto. Assim, todos os elementos não textuais deveriam possuir um nome vinculado a palavra-chave do *site* e/ou um nome significativo que identificasse esse elemento da forma mais clara possível para o usuário. Desta forma, contribuindo para a melhor classificação nas listagens de resposta dos mecanismos de busca.

Mais outros elementos não textuais que pode prejudicar a classificação são os documentos digitais construídos utilizando o formato SWF (*Shock Wave Flash*) conhecidos também como *sites* em *flash*. Com o formato SWF é possível criar documentos digitais inteiros, parciais ou mesclados com o HTML, dinâmicos ou estáticos, com ou sem acesso a banco de dados, mas que podem comprometer a classificação do documento, pois sua arquitetura interna vetorial impede qualquer utilização pelos mecanismos de busca que utilizam apenas elementos textuais.

Todos os *sites* que utilizam o formato SWF precisam de uma base HTML que chama o arquivo SWF. Essa base HTML precisa oferecer uma descrição bem elaborada utilizando a *meta tag description* e a *meta tag keyword* para que os mecanismos de busca possam classificar alguns elementos. Adicionalmente o nome do arquivo SWF deve usar um nome vinculado à palavra-chave do *site* ou o material desenvolvido, para que possa adicionar valor para a classificação dos *spiders*.

Como o arquivo SWF não pode ser identificado pelos mecanismos de busca, uma outra técnica é construir uma página inicial que possa descrever o que o *site* pode oferecer e nessa página inicial deve ser inserida um *link* para o *site* desenvolvido em formato SWF.

Outra possibilidade de uso para fazer o documento digital ser encontrado pelos mecanismos de busca é desenvolver novamente uma página inicial com uma boa descrição do *site,* e nesta página existir dois caminhos, um para a versão do documento digital com base HTML e outra com base SWF. Como o mecanismo de busca classifica o material feito em HTML, também classificará a versão SWF.

Outros elementos que podem ser inseridos na *Web* como músicas, imagens, animações, filmes, programas aplicativos, arquivos PDFs, planilhas, arquivos de texto, etc, precisam ser identificados para que possam ser encontrados em uma situação de pesquisa. E para que os mecanismos possam encontrá-los é necessário uma descrição textual adequada. Essa descrição pode ser feita no nome do arquivo, na *tag alt*, ou vinculando um texto visível que seja descritivo e que esteja ao redor do elemento disponibilizado(Kwok et al., 2001).

14 Características de uso das *meta tags* e *tags alt*

Entre as diversas *meta tags* existentes, as *meta tags description* e *keyword* são essenciais. A *meta tag description* deverá ser usada para descrever em até no máximo 255 caracteres quais são as características do material divulgado do documento digital. Alguns mecanismos classificam menos que 255 caracteres, mas isso não impede o funcionamento do *site,* apenas limitará a quantidade de caracteres lida pelo mecanismo de busca.

unesp

Ao contrário da *meta tag description* que é visualizada na listagem de resposta do mecanismo de busca, a *meta tag keyword* não aparece em nenhum lugar, mas é utilizada para classificação na base do mecanismo de busca. É recomendado não ultrapassar os 255 caracteres, pois o que passar desse limite pode ser ignorado por alguns mecanismos. Se for de interesse do responsável pela identificação do documento digital inserir mais *keywords*, é possível repetir a *tag* quantas vezes for necessário. Mas dependerá da política de classificação do mecanismo de busca classificar ou não mais de uma.

Entretanto, os mecanismos de busca podem detectar a repetição seqüencial de palavras destinadas a melhorar a classificação do documento digital de forma artificial. Isso é classificado como *Spam*, e é radicalmente combatido. O propósito dos mecanismos de busca é tentar trazer a informação procurada o mais próximo possível do interesse do usuário, e não classificar um documento digital porque seu responsável aumentou artificialmente a classificação do *site*.

Quando os mecanismos de busca detectam a utilização de *Spam* os *sites* que os utilizam são excluídos das bases de dados.

Uma técnica que pode ser utilizada na *meta tag keyword* é prever a digitação da palavra-chave pelo usuário. Assim, devem ser colocadas versões diferentes da palavra-chave, e também opções simulando digitação equivocada, sem, no entanto, ser considerado um *Spam.* Como por exemplo: sendo a palavra-chave "Design", poderíamos prever: design, DESIGN, dESIGN, Dsign e desin, entre outras.

Além disso, as palavras inseridas na *meta tag keyword* devem ser obrigatoriamente separadas por virgula para que sejam interpretadas como várias palavras, pois se não forem colocadas, o mecanismo interpretará como sendo um conjunto de uma só palavra.

15 *Spam*

Alguns responsáveis em divulgar o documento digital acreditam que podem, através de técnicas artificiais, aumentar a classificação nas listagens de busca. Entretanto, se o



mecanismo de busca detectar este procedimento, o *site* é excluído da base. Abaixo seguem alguns métodos de *Spam;*

- Seqüências idênticas de palavras ou frases;
- Preenchimento do texto visível ou não visível com a palavra-chave de forma aleatória no corpo do documento HTML, geralmente sem nenhum sentido textual;
- Preenchimento de palavras-chave com a mesma cor de fundo, tornando-as invisíveis para a visualização do usuário, mas visível para o mecanismo de busca;
- Outros *sites* criados exclusivamente com vários *links* idênticos para o *site* principal.

Alguns mecanismos de busca não revelam o que consideram como técnica de *Spam,* outros revelam parcialmente. Desta forma, não se deve julgar que porque um mecanismo de busca não forneça dados referentes a técnicas de inibição de *Spam* que ele não adote nenhuma.

16 Base de dados Patrocinados

Muitas bases de dados existentes na *Web* oferecem dados patrocinados, ou seja, em certas pesquisas os primeiros itens podem ter comprado sua posição na listagem de resposta. Essas bases que oferecem a possibilidade de comprar a posição de classificação estão descaracterizando o objetivo inicial da *internet,* que era um meio rápido de se encontrar informação.

Entretanto, por serem patrocinados, podem oferecer um serviço de acesso mais rápido e mais extenso em comparação a bases que não são patrocinadas, mas isso não significa que o serviço oferecido pelos não patrocinados possa ser de menor qualidade. Muitas vezes, por ser uma base de menor amplitude e de assunto restrito, as informações contidas nessas bases podem ter maior relevância que em bases convencionais.



Outra característica de documentos digitais patrocinados é a inclusão no topo da listagem de busca, de algum elemento gráfico que divulga algum material que em alguns casos, pode ter ligação direta com a pesquisa realizada.

Adicionalmente, essa atitude de vender a posição nas listagens de pesquisa não é interessante para o usuário, pois o mesmo pode levar em consideração uma informação patrocinada obtida nas listagens de busca como verdade e transformar essa informação em verdade absoluta, prejudicando seu julgamento em relação à informação encontrada.

Outros mecanismos de busca oferecem dificuldade em oferecer acesso a documentos digitais que não tenham em sua extensão de URL a complementação ".com", o que significa que o conteúdo tem alguma relação com comércio e assim, deixando de classificar outros documentos digitais, que podem trazer em sua extensão de URL o complemento .org, .gov, entre outros.

17 Proposta de metodologia para classificação otimizada de documentos digitais nos mecanismos de busca

Este artigo tem como objetivo apresentar uma metodologia para tornar um documento digital melhor classificado, utilizando como recursos seu código, adaptando, adequadamente, itens internos que podem contribuir para a melhor classificação nos mecanismos de busca.

Desta forma, alguns itens abaixo podem ser seguidos, para se obter uma classificação otimizada:

17.1 Escolher qual ou quais serão os mecanismos de busca que serão usados para indexação do conteúdo digital;

17.2 Extrair dos mecanismos de busca escolhidos, quais são os itens de maior importância para que os mesmos, realizem o processo de classificação de seus bancos de dados.

17.3 Com o material de desenvolvimento do *site* em mãos, identificar qual é a palavra-chave de maior importância (ou palavras-chave). De preferência, que o *site* ainda esteja



em processo de planejamento, pois caso contrário será necessário alterar muitos itens, como por exemplo, todos os elementos não textuais existentes no documento digital.

17.4 Usando os mecanismos de busca que indexarão o documento digital ainda a ser criado, é necessário realizar uma pesquisa com a palavra-chave identificada no item anterior, e verificar, nos primeiros *sites* encontrados na listagem de busca qual é a número de vezes que a palavra-chave é referenciada, para poder classificar os *sites* encontrados. Assim, se o objetivo for ser o primeiro na listagem de busca, o número de repetições da palavra-chave deve ser maior que a usada no *site* que foi classificado como primeiro, sem fazer utilização de técnicas de *spam.*

17.5 Os nomes internos dos possíveis diretórios para armazenar determinados conteúdos referentes ao documento digital, devem também ser nomeados, utilizando a palavra-chave do *site* e mais um complemento que possa caracterizar os elementos que serão guardados neste diretório. Por exemplo: *design*_imagens. Esses pequenos detalhes podem diferenciar um documento digital do outro na classificação.

17.6 O URL (*Universal Resource Locator*) ou simplesmente o endereço do *site* deve, dentro do possível, possuir uma referência a palavra-chave, sempre com o objetivo de pontuação. Exemplo de URL com a palavra-chave inserida: http://www.*design*.org.br.

17.7 Utilizar a *tag title* referenciando o nome do *site,* o assunto que trata o *site*, ou produto que se encontra no *site.* Não se deve usar artigo antes do substantivo que identifica a *tag title.* Não se deve colocar nenhum outro caractere antes ou depois do conteúdo da *tag title,* pois a ocorrência mais idêntica à palavra pesquisada pelo usuário será classificada em uma posição mais elevada que uma ocorrência similar. De preferência, o conteúdo da *tag title* deve ser o mais próximo possível da palavra-chave escolhida. Não se deve colocar mais de 85 caracteres na *tag title.*

17.8 A identificação do documento digital é essencial para o rastreamento dos mecanismos de busca, assim, as *meta tags* não podem ser esquecidas. Dependendo do conteúdo do *site*, não existe a necessidade da utilização de todas elas. Mas todo *site* deve usar pelo menos as seguites:

- ***<META NAME="Keywords" CONTENT="palavras_separadas_por_vírgula">***

- ***<META NAME="Description" CONTENT="descrição_da_página_ou_site">***
- ***<META NAME="Robots" CONTENT="all | index | noindex | follow">***
- ***<META HTTP-EQUIV="Content-Language" CONTENT="br">***

17.9 Utilizar a *tag alt* corretamente, para que se possa somar pontos na classificação do mecanismo de busca. Para se fazer uso correto da *tag alt*, é necessário vincular a palavra-chave do *site* com o nome do item ou uma pequena descrição do que será encontrado na outra ponta do *link*, isso se o elemento for um *link.*

17.10 Todos os elementos não textuais como botões, marcadores, arquivos de imagens fotográficas, arquivos de imagens de desenhos ou logomarcas, arquivos de música ou efeitos sonoros, animações, apresentações, arquivos PDFs, arquivos executáveis, planilhas, arquivos de textos, etc, devem estar vinculados a uma *tag alt* ou envolvidos por texto descritivo, para que possam ser detectados e classificados.

17.11 A posição do texto visível vinculado com a posição da palavra-chave é crucial para a boa pontuação do documento digital. Assim, respeitando o *layout* da página, quanto mais próximo do início da página e a esquerda, mais forte é a pontuação. Essa característica não é a única forma de pontuação, assim é possível compor um *layout* fora desses padrões estabelecidos, pontuando menos em relação ao texto, e pontuando mais, com outros elementos.

17.12 Quanto mais próximo o *link* da informação desejada, mais pontos são oferecidos para a classificação do *site*. Assim, deve-se evitar posicionar a informação em camadas de *links,* internos, fazendo o usuário clicar várias vezes. Como regra geral sobre *links,* estabelecer uma distância máxima de três cliques quando possível. O objetivo é tornar a informação o mais próxima do usuário, e um documento digital que oferece uma informação com um caminho muito longo não é bem pontuado.

17.13 Os *Links* que estão no documento digital e que apontem para outros *sites* bem acessados e que possuam ligações com o material oferecido pode aumentar a classificação. Outros *sites* que possam apontar para o documento digital também oferecem pontuação na classificação, se o *site* que aponta for de uma grande empresa, ou de um *site* que tenha um grande índice de acessos, isto pode acarretar uma



classificação melhor. No entanto, nem todos os mecanismos possuem essa característica.

18 Conclusão

Se o responsável ou a equipe responsável pelo desenvolvimento do *site* conseguir reunir partes desses elementos ou todos eles, as possibilidades de boa classificação do documento digital serão ampliadas nos mecanismos de busca, ampliando a alcance da informação ao usuário.

Com o objetivo de alcançar o usuário final ou o público-alvo e facilitar sua utilização de sistemas distribuídos de informação, o responsável pelo documento digital, possui ferramentas que podem e devem ser usadas para auxiliar a boa classificação das informações disponibilizadas, atendendo as necessidades de acesso rápido de informações. Não apenas sendo responsável pela elaboração do *design* e/ou da programação de conteúdos estáticos ou dinâmicos, mas sendo responsável por criar uma estrutura orgânica de ligação de elementos que possam agir não de forma individual, mas em conjunto, para um objetivo maior que deve ser alcançar o usuário com boa classificação e informação relevante. Entretanto, mesmo seguindo essas recomendações, há no mínimo duas variáveis que estão fora de controle ao responsável pelo documento digital que são o algoritmo de classificação dos mecanismos de busca, que podem ser alterados sem prévio aviso e, a inclusão de novos *sites* que possuem melhores referências de palavras-chave e códigos internos. Portanto, um documento digital que possa estar bem classificado em um momento, poderá não estar em outro. Contudo, todo documento digital precisa de manutenção periódica para evitar seu desaparecimento repentino e/ou substituição por outro. Da mesma forma que um *site* pode ser projetado para estar bem classificado, outros *sites* que possam ser criados seguirão os mesmos procedimentos, para alcançarem uma boa colocação, e quanto mais os fatores de boa classificação forem utilizados, melhores benefícios alcançarão os usuários dos sistemas distribuídos.

Como resultados práticos foram desenvolvidos três *sites* que podem comprovar a boa classificação em mecanismos de busca. O primeiro é o *site* do zoológico da cidade de

cm 1 2 3 4 5 6 7 unesp 1 12 13 14 15 16 17 18

Bauru SP, que pode ser acessado através do endereço: http://www.zoobauru.kit.net e que pode ser encontrado em 2º lugar no *Google*, 1º lugar no AltaVista, 1º lugar no *Yahoo* 1º lugar no Cadê - 2º lugar no AOL.com e 38º lugar Radar uol, com palavra-chave: zoobauru posição comprovada em 08/12/2002. O segundo *site* é o SIHOP, Sistema de Horário de Professores, que pode ser acessado no endereço http://www.sihop.kit.net e que pode ser encontrado em 1º lugar no AltaVista, 1º lugar no *Yahoo*, 1º lugar no Cadê, com palavra-chave: sihop, e 1º lugar no *Yahoo*, 1º lugar no Cadê, 1º lugar no AltaVista com palavra-chave: Sistema de Horário de Professores consulta comprovada em 08/12/2002. O terceiro *site* é o Crami, que pode ser encontrado no endereço: http://www.crami.cjb.net e que pode ser encontrado em: 1º lugar no Miner, *Yahoo* e no Cadê, no Alta Vista em 2º, em 3º no *Google* e no Radar Uol, e em 6º no *WebCrawler,* posição comprovada em 15/02/2003, com palavra-chave: crami bauru.

Agradecimento

Este trabalho contou com o suporte financeiro da Fapesp processo 97/13309-5

Referências

BERGMAN, Michael K. (2001) *The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing. The University of Michigan Press.* Vol 7, Issue 1,Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em 17 de setembro de 2002.

FURGERI, Sérgio. (2001) Ensino Didático da Linguagem XML. ed. Érica, São Paulo, 278p.

KWOK, Cody. ETZIONI, Oren. WELD, Daniel S. (2001) *Scaling question answering to the web. Capes. The Gale Group*. ACM *Transactions on Information Systems*, vol 19, i3, pp.242-260.

NIELSEN, Jackob. (2000) Projetando *websites*. *Designing web usability*. Ed. Campus, Rio de Janeiro, 416p.

RADFAHER, Luli. (2001) *Design/web/design*2. Ed. *Market Press*, São Paulo, 265p.

SILVEIRA, Marcelo. (2002) *Web Marketing*, Usando Ferramentas de Busca. ed. Novatec, São Paulo, 159p.

SULLIVAN, Danny. (2000) *Web Pages Tweaking: Will the Best Rise to the Top? Danny Sullivan's Search Engine Strategies* 2000 *conference*. *Capes. The Gale Group. The Information Advisor*, Vol 12, i5 pp.4-10.

Anexo 3

WebDesign Goal: Recuperação de Informações em Sistemas Distribuidos
WebDesign Goal: Information Retrieval Systems

Rodrigo Ferreira de Carvalho
Unesp, Colégio Técnico Industrial "Prof. Isaac Portal Roldan"
Dr. João Fernando Marar
Laboratório de Sistemas Adaptativos e Computação Inteligente
Depto de Computação –FC - Unesp

Palavras chave:

Webdesign, Mecanismos de busca, Sistemas Distribuídos.

Resumo

O impacto da Internet está atingindo uma grande quantidade de usuários, e seu crescimento gera uma quantidade de informação muito grande, o que não significa que poderá ser encontrada com facilidade. Atualmente para se encontrar a informação desejada na *Web* não constitui uma tarefa muito fácil; tornando o ato da pesquisa uma tarefa árdua. Afim de minimizar as dificuldades em encontrar informações, algoritmos de classificação para os mecanismos de busca da *Web* precisam de melhores adaptações, no que tange a garantia de procura de informação correta, aplicações em Inteligência Artificial, etc. Neste sentido, o *Webdesign* pode atuar de forma decisiva, proporcionando uma melhor resposta na classificação dos mecanismos de busca. Esse artigo tem por objetivo descrever procedimentos que promovam a melhoraria da classificação do documento digital, e que estão ao alcance do responsável pela elaboração do *site*.

Key words:
Webdesign, Search Engines, Distributed Systems.

Abstract

The Internet impact is reaching a great number of users, consequently, it is generating a very big data information, however, with relation about this subject, there are much informations with no relevance and that these same informations, sometimes, are not so easy to find. To find an information web search engines classification algorithms need better adaptations with relation to the guarantee to find the right information, applications in artificial inteligence, etc. In this way Webdesigners can act in a decisive way, providing a better answer in the search engines classification. So this article has for objective to describe procedures that promote the form to make better the digital documents classification in which is the hands of the responsible person to create the site.

* Este Artigo contou com o suporte financeiro da Fapesp, processo NRO 97/13.309-5.

cm 1 2 3 4 5 6 7 unesp 1 12 13 14 15 16 17 18

1- Introdução

Em 1969 foi realizado a primeira conexão entre computadores nos campi de quatro universidades situados em posições geográficas distantes. A experiência aconteceu em Stanford Researsh Institute (SRI), Universidade da Califórnia (UCLA), Universidade de Santa Barbara e a Universidade de Utah em Salt Lake City, formando quatro pontos em uma rede de computadores interligados. A experiência realizada, não tinha precedentes, o Professor Dr. Leonard Keinrock pioneiro em Ciência da Computação da UCLA, e seu pequeno grupo de estudantes esperavam se conectar com o computador de Stanford e tentar enviar alguns dados. O grupo começou a digitar algumas teclas e esperavam que essas letras aparecessem em um outro monitor de vídeo em outra universidade. Assim, para verificar se o que era digitado em um computador, estava sendo transmitido para um outro computador em outra universidade, foi utilizado o telefone, para confirmar verbalmente a chegada dos caracteres. Neste teste foram digitados sucessivamente os caracteres "L", "O", "G", em seguida o sistema interrompeu a conexão, e a partir deste momento originou-se a revolução da comunicação digital. Desta forma, acelerando o processo de "aldeia global", termo criado por Mcluhan, (1996), em 1964 para explicar a simultaneidade em que o tempo e o espaço desaparecem, quando os meios eletrônicos de comunicação começam a envolver o ser humano.

Os quatro pontos formados por aquelas universidades deram início ao que conhecemos hoje por *Internet*, não com a mesma aparência atual, mas funcionando estruturalmente da mesma forma que no passado. E por conseqüência desta experiência, a informação tem rompido várias barreiras, como tempo, idade, distâncias e custos assim, criando a sociedade da informação, ou a sociedade da velocidade da luz.

2- Sistemas Distribuídos como Suporte a Segurança de Informação

Para o funcionamento da *Internet* foi desenvolvida uma arquitetura para que nenhuma das bases possuísse a totalidade das informações, simplesmente para assegurar que os computadores conectados não parassem de funcionar se um deles, por algum motivo, sofresse algum dano, ou que o computador que armazenasse todos os dados pudesse ser atingido, e conseqüentemente, parar toda a comunicação realizada através da rede formada pelos computadores. A esta característica de comunicação e controle, denominamos Sistema Distribuído.

Adicionalmente, os documentos digitais que trafegam nessas rotas nos sistemas distribuídos não funcionam apenas com a elaboração do *design*, do conteúdo e da programação, há também a arquitetura de informação[1], responsável por permitir que o usuário encontre o que procura com o menor número de interações possíveis. Para conseguir isto, é necessário organizar o conteúdo de maneira clara e específica. A arquitetura de informação permite que os visitantes saibam onde estão e para onde podem ir, é a definição clara do caminho lógico para se encontrar a informação.

cm 1 2 3 4 5 6 7 unesp 1 12 13 14 15 16 17 18

A Arquitetura da Informação visa à organização de grandes massas de dados, preparando rotas de acesso a eles. Uma arquitetura eficiente torna a informação acessível e compreensível aos usuários, qualquer que seja seu nível de conhecimento.

3- Otimizar as Possibilidades de Classificação de Documentos Digitais e Encontrar Informação Segura

O propósito da *Internet* sempre foi o armazenamento e o acesso rápido sem, no entanto, armazenar a informação em apenas um único lugar, mas com o passar do tempo, podemos notar que seu funcionamento não se encontra hoje como foi planejado. Ou seja, encontrar o que se deseja de maneira a não desperdiçar muito tempo na pesquisa.

Apesar da *Internet* ter mais de 30 anos, sua abertura ao acesso pública chaga há quase 10 anos, e a quantidade de informação torna-se um grande problema (Bharat, 2000; Chang at. al., 2000; Gandal, 2001). Como encontrar a informação necessária em uma simples pesquisa que pode nos trazer muito mais de um milhão de alternativas? Segundo (Kwok at al., 2001) a crescente base de dados que amplia e dificulta o rastreamento de informações, tornando uma pesquisa simples na *Web*, uma tarefa às vezes problemática, ou pela falta ou por encontrar uma enorme quantidade de informações, que podem não estar bem classificadas em uma listagem de respostas, fazendo com que o usuário tenha que percorrer vários itens de uma classificação equivoca, antes de encontrar a informação que o satisfaça.

Alguns estudiosos[2] afirmam que, apenas 20 por cento de todo material depositado na *Internet* tem chance de ser acessado, pois certos métodos de cadastramento do documento digital ou são desprezados ou são desconhecidos por quem disponibiliza a informação figura 1. Assim, o material publicado na *Internet* permanece oculto, sem acesso, pelo fato de que procedimentos de identificação foram ignorados. Desta forma, mais um instrumento foi projetado para a *Internet*, o mecanismo de busca. Nos últimos anos a *Web* cresceu tanto que é impossível existir um único lugar que inclua todos os *sites*.

4 – Características dos mecanismos de busca

Os mecanismos de busca podem ser classificados em duas categorias, os que usam mecanismos chamados *spiders*, e os diretórios de busca. Os *spiders* são programas

---

(1) Arquitetura de informação, a estrutura e organização lógica de funcionamento de um sistema computacional.

(2) Brin, (1998), Castro, (2000), Sullivan, consultor de *Internet* e estudioso do comportamento dos mecanismos de busca e como fazem a indexação dos *sites* desde 1995, também foi o fundador do *site Search Engine Watch,* http://www.searchenginewatch.com. Leavitt, co-fundador da empresa *Lycos* e CEO (*Chief Executive Officer*) em tecnologia *Web*, http://www.steampunk.com/jrrl/vita.html

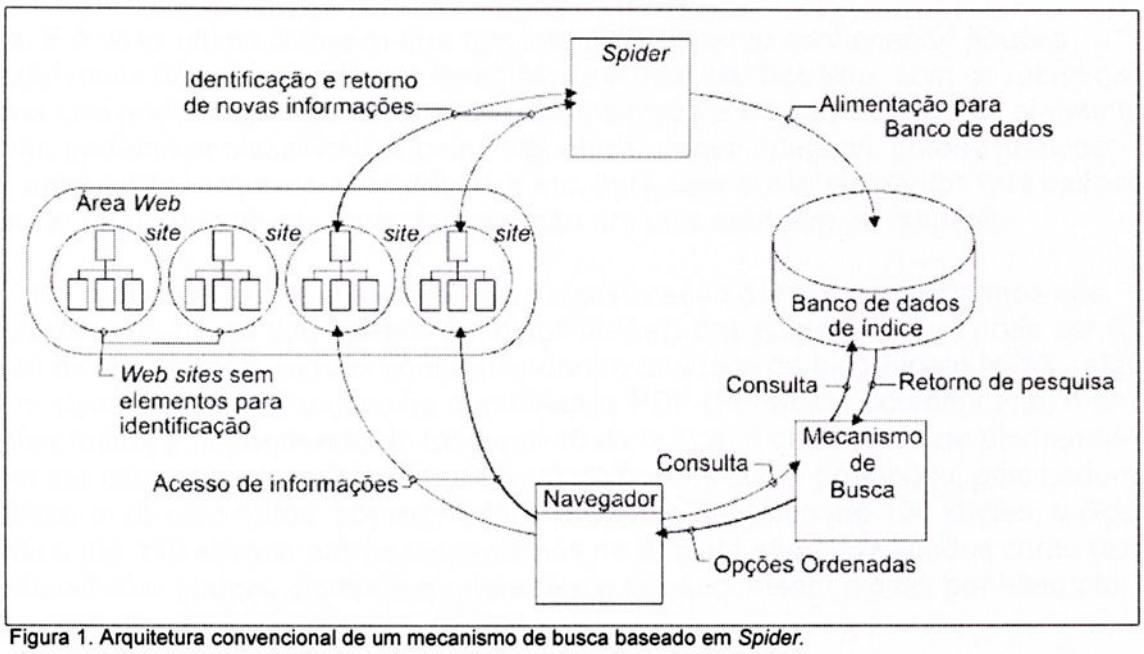


Figura 1. Arquitetura convencional de um mecanismo de busca baseado em *Spider*.

que buscam informações nos provedores de *Internet* e realizam copias do material existente, apenas textual e os classificam em seus bancos de dados. O *spider* foi uma das tentativas de organizar e facilitar o conteúdo da *Internet*, e foi criado em 1994 por John Leavitt co-fundador da empresa *Lycos* e CEO (*Chief Executive Officer*) em tecnologia *Web*. A diferença entre os *spiders* e os diretórios de busca, é que os mecanismos de busca que usam *spiders* farão a classificação do material através do conteúdo apresentado no *site*, enquanto que os mecanismos classificados como diretórios, classificam os *sites*, na grande maioria através de uma analise prévia de um editor humano, o que pode causar uma espera maior do que nos mecanismos que usam *spiders* na indexação dos banco de dados, e até mesmo correr o risco de não cadastramento do *site*, se o material contido no documento digital não estiver dentro do interesse do diretório. Outra característica importante, é que para um *site* aparecer nos diretórios de busca, o responsável necessita fazer o cadastramento obrigatório, enquanto que nos mecanismos que usam *spiders* esse procedimento não tem obrigatoriedade, mas também se torna adequado realizar a apresentação, pois desta forma a indexação do material poderá se realizar em um tempo menor.

Para que se possa utilizar adequadamente os recursos dos mecanismos de busca, seria necessário conhecer totalmente as características do algoritmo de classificação de cada um dos mecanismos. No entanto, apenas algumas características são reveladas, pois o interesse dos mecanismos é realizar dentro do possível a melhor classificação, sem que o responsável pelo documento digital possa o bem classificar apenas porque saiba com funciona o algoritmo de classificação. Na verdade, é possível comprar uma boa posição nas listagens de resposta em alguns mecanismos, mas o que se pode notar é que o grau de confiabilidade do mecanismo de busca acaba sendo prejudicado com esse tipo de ação (SILVEIRA, 2002). Assim sendo, as listagens começam a mostrar duas áreas de informação, uma com o título patrocinado e uma outra com dados realmente classificada através das informações adquirias pelos mecanismos de

busca. E é essa última listagem que nos interessa, mesmo conhecendo poucas características dos algoritmos de classificação é possível trabalhar com os recursos básicos que podem ser utilizados pelos mecanismos de busca, e utilizar os elementos que não podem ser classificados pelos algoritmos como: imagens, botões gráficos, sons, animações, arquivos executáveis e etc, para usar como elementos que possam colaborar na pontuação de uma classificação em uma listagem de resposta.

Na realidade, o que é utilizado para classificação dentro dos algoritmos dos mecanismos de busca são apenas os textos visíveis das páginas (o que pode ser lido através do navegador), alguns conteúdos dentro das *tags* da Linguagem HTML, alguns *spiders* classificam informações de documentos PDF (Portable Document File) e de arquivos textos e dependendo do mecanismo de busca, a quantidade de elementos que podem ser lidos, alguns mecanismos são classificados como profundos, pois podem classificar mais elementos, por exemplo: o AltaVista classifica até 100 kbytes, o *Google* classifica até 110 kbytes, outros mecanismos no entanto são classificados como rasos, pois classificam poucos elementos referentes a um documento digital, por exemplo: apenas o texto visível da primeira página.

5 – WebDesign Goal: Uma Metodologia para a garantia de Recuperação de Informação em Sistemas Distributivos

Algumas empresas que se dizem especializadas em classificação de documentos digitais afirmam que podem classificar um *site* em até 500 mecanismos diferentes, desta forma, tornando o material mais popular.
Esse serviço pode acabar prejudicando o documento digital, pois na verdade esse tipo de empresa pode colocar elementos para classificação que são alheios aos interesses iniciais do *site*, e na verdade uma visualização para 500 ou mais mecanismos de busca pode acabar não sendo muito útil, pois certos mecanismos são destinados a conteúdos específicos e que não trarão nenhuma vantagem em aparecer classificados em seus índices, outros por sua vez, podem ser mecanismos de regiões ao qual o conteúdo não oferecerá o menor retorno para nenhum dos lados. Assim, o mais viável seria um projeto desenvolvido voltado para necessidades do documento digital dentro dos mecanismos que poderão alcançar o maior número de interessados no produto ou na informação oferecida pelo *site*.

Para que se possa atingir o público alvo, um estudo detalhado sobre o material do *site* é necessário, para que se possa definir a palavra ou palavras chave do *site*, e através da mesma, vincular todos os nomes de elementos não textuais com a palavra chave selecionada.

É importante lembrar que o uso adequado da palavra chave na estrutura do *site*, não é a único meio para classificar o documento digital, na verdade, apenas a palavra chave sendo utilizada não garante boa classificação. Desta forma, um conjunto de outros fatores reunidos, acabam se tornando fundamental para que se possa ter uma boa classificação, tais como:

unesp

5.1 Escolher qual ou quais serão os mecanismos de busca que serão usados para indexação do conteúdo digital. Tabela 1.

MECANISMOS DE BUSCA	NÚMERO DE PÁGINAS
Google	1.500.000.000
Fast Search All the Web	625.000.000
AltaVista	550.000.000
Webtop.com	500.000.000
Inktomi	500.000.000
Northern Light	390.000.000
Excite	250.000.000

Tabela 1. Tamanhos dos mecanismos de busca. Dados de dezembro de 2001.
Fonte: http://searchenginewatch.internet.com/reports/sizes.html

Atualmente o mecanismo de busca *Google,* é o maior em relação ao tamanho de seu banco de dados e que indexa seu conteúdo independente de dados patrocinados. O *Google* fornece dependendo do assunto, alguns itens pagos que aparecem nas listagens de busca, mas estes dados estão identificados separadamente como informação patrocinada, sendo o acesso de dados patrocinados uma opção aos dados classificados pelos *spiders.* O aparecimento do *site* na listagem de busca pode demorar de uma a três semanas, quando os dados internos para indexação são utilizados. O *Google* também analisa a importância dos *links* internos e externos, como sendo uma característica das mais importantes para classificação. Não é necessário o cadastramento de todas as páginas do *site,* pois como o *Google* é um mecanismo profundo, o mesmo classifica todas as páginas a partir do cadastramento da *home page.* O *Google* classifica não apenas *sites* com domínio comercial ".com", mas qualquer outro domínio, o que abre uma maior quantidade de opções para os usuários.

O mecanismo de busca Alta Vista, também mudou sua estratégia de listagem de resposta, para uma apresentação semelhante ao do *Google.* É interessante lembrar que no ano 2000, O *Google* quase não era conhecido e não oferecia concorrência ao Alta Vista ou a qualquer outro mecanismo.

Desta forma, podemos avaliar que a separação das informações patrocinadas e das realmente classificadas oferece um grau maior de confiabilidade das informações ao mecanismo de busca, o que ainda não é suficiente para qualificar a informação.

5.2 Extrair dos mecanismos de busca escolhidos, quais são os itens de maior importância para que os mesmos realizem o processo de classificação em seus bancos de dados.



5.3 Com o material para desenvolvimento do *site* em mãos, identificar qual é a palavra chave de maior importância (ou palavras chave). De preferência, que o *site* ainda esteja em processo de planejamento, pois caso contrário será necessário alterar muitos itens, como por exemplo, todos os nomes dos elementos não textuais existentes no documento digital.

5.4 Usando os mecanismos de busca que indexarão o documento digital ainda a ser criado, é necessário realizar uma pesquisa com a palavra chave identificada no item anterior, e verificar nos primeiros *sites* encontrados na listagem de busca qual é a número de vezes que a palavra chave é referenciada, para poder classificar os *sites* encontrados, e assim, se o objetivo for ser o primeiro na listagem de busca, o número de repetições da palavra chave deve ser maior que a usada no *site* que foi classificado como primeiro, sem a utilização de técnicas de *spam. Spam*, é a técnica que utiliza repetição de elementos textuais para poder aumentar a pontuação na classificação do *spider* de forma artificial e sem sentido, mas se detectado pelo mecanismo de busca, pode causar a eliminação do conteúdo digital do banco de dados. É importante deixar claro que, apenas o uso da quantidade maior de palavra chave do documento digital não garante a melhor classificação, isso dependerá do conjunto de fatores reunidos em favor da melhor pontuação do *site*.

Adicionalmente, alguns *sites* podem ser classificados em privilegiadas posições sem terem utilizado recursos de palavras chaves. Quando for encontrado esse tipo de situação, significa que o documento digital pode ter comprado sua posição no mecanismo de busca.

5.5 Os nomes internos dos possíveis diretórios para armazenar determinados conteúdos referentes ao documento digital, devem também ser nomeados utilizando a palavra chave do *site* e mais um complemento que possa caracterizar os elementos que serão guardados neste diretório. Por exemplo: projeto_tamar_imagens, "projeto_tamar" seria a palavra chave e "imagens" seria o nome do local (pasta) em que seria armazenado as imagens utilizadas no *site*. Esses pequenos detalhes podem diferenciar um documento digital de um outro na classificação.

5.6 O URL (*Universal Resource Locator*) ou simplesmente o endereço do *site* deve dentro do possível possuir uma referência a palavra chave, sempre com o objetivo de pontuação. Exemplo de URL com a palavra chave inserida: http://www.projeto_tamar.org.br ou http://www.projeto_tamar.hpg.com.br.

5.7 Utilizar a *tag title* referenciando o nome do *site,* o assunto que trata o *site*, ou o produto que se encontra no *site.* Não se deve usar artigo antes do substantivo que identifica a *tag title.* Não se deve colocar nenhum outro caractere antes ou depois do conteúdo da *tag title,* pois a ocorrência mais idêntica à palavra pesquisada pelo usuário será classificada em uma posição mais elevada que uma ocorrência similar. De preferência, o conteúdo da *tag title* deve ser o mais próximo possível da palavra chave escolhida. Não se deve colocar mais de 85 caracteres na *tag title.*



5.8 A identificação do documento digital é essencial para o rastreamento dos mecanismos de busca, assim, as *meta tags* não podem ser esquecidas. Dependendo do conteúdo do *site*, não existe a necessidade da utilização de todas elas. Mas todo *site* deve usar pelo menos as seguintes:

- ***<META NAME="Keywords" CONTENT="palavras_separadas_por_vírgula">***
- ***<META NAME="Description" CONTENT="descrição_da_página_ou_site">***
- ***<META NAME="Robots" CONTENT="all \| index \| noindex \| follow">***
- ***<META HTTP-EQUIV="Content-Language" CONTENT="br">***

5.9 Utilizar a *tag alt* corretamente, para que se possa somar pontos na classificação do mecanismo de busca. Para se fazer uso correto da *tag alt*, é necessário vincular a palavra chave do *site* com o nome do item ou uma pequena descrição do que será encontrado na outra ponta do *link*, isso se o elemento for um *link*.

5.10 Todos os elementos não textuais como botões, marcadores, arquivos de imagens fotográficas, arquivos de imagens de desenhos ou logomarcas, arquivos de música ou efeitos sonoros, animações, apresentações, arquivos PDFs, arquivos executáveis, planilhas, arquivos de textos, etc, devem estar vinculados a uma *tag alt* ou envolvidos por texto descritivo, para que possam ser detectados e classificados.

5.11 A posição do texto visível vinculado com a posição da palavra chave é crucial para a boa pontuação do documento digital. Assim, respeitando o *layout* da página, quanto mais próximo do início da página e a esquerda mais forte é a pontuação. Essa característica não é a única forma de pontuação, assim é possível compor um *layout* fora desses padrões estabelecidos e pontuando menos em relação ao texto, e pontuando mais, com outros elementos.

5.12 Quanto mais próximo o *link* da informação desejada, mais pontos é oferecido para a classificação do *site*. Assim, deve-se evitar posicionar a informação em camadas de *links* internos, fazendo o usuário clicar várias vezes. Como regra geral sobre *links* estabelecer uma distância máxima de três cliques, quando possível. O objetivo é tornar a informação o mais próxima do usuário, e um documento digital que oferece uma informação com um caminho muito longo não é bem pontuado. Esse item possui relação direta com o termo arquitetura de informação.

5.13 *Links* que estão no documento digital e que apontem para outros *sites* bem acessados e que possuam ligações com o material oferecido pode aumentar a classificação. Outros *sites* que possam apontar para o documento digital também oferecem pontuação na classificação, se o *site* que aponta for de uma grande empresa, ou de um *site* que tenha um grande índice de acessos, pode acarretar uma classificação melhor. Nem todos os mecanismos possuem essa característica.



6- Conclusão

Se o responsável ou a equipe responsável pelo desenvolvimento do *site* conseguir reunir partes dos elementos citados no item cinco ou todos eles, as possibilidades de boa classificação do documento digital serão ampliadas nos mecanismos de busca.

As 13 etapas citadas no item cinco seriam obrigações de um projeto *Web*, que muitas vezes acabam não sendo feitas, e o resultado da não aplicação destes itens é a má classificação do documento digital, dificultando o acesso de informação para o usuário, e eliminando a chance de visualização do conteúdo pretendido pelo responsável pelo *site*.

Assim, o *Webdesigner*, que tem a responsabilidade de conseguir deixar o material mais compreensivo, de fácil navegação e atrativo para o usuário deveria também possibilitar que o material possa ser encontrado pelos mecanismos de busca através de palavras chaves que estabelecem vínculos com as informações apresentadas no *site*, e/ou elementos não textuais que possam atrair o usuário para uma das partes internas do *site*, tornando desta forma, a informação mais próxima das necessidades de quem procura a informação. Desta forma, transformando o momento da pesquisa apenas em um breve instante para a obtenção do conteúdo procurado. Assim, se o processo de criação de um documento digital para a *Web* for elaborado organicamente ao qual, todos os elementos podem contribuir para um bom resultado nas classificações das listagens dos mecanismos de busca, a informação do documento digital estará mais próxima do usuário, facilitando sua utilização, não importando se o material desenvolvido for de caráter pessoal, educacional, corporativo, comercial ou etc.

Como resultados práticos foram desenvolvidos dois sites que podem comprovar a boa classificação em mecanismos de busca. O primeiro é o site do zoológico da cidade de Bauru SP, que pode ser acessado através do endereço: http://www.zoobauru.kit.net e que pode ser encontrado em 2º lugar no *Google*, 1º lugar no AltaVista, 1º lugar no *Yahoo* 1º lugar no Cadê - 2º lugar no AOL.com e 38º lugar Radar uol, com palavra chave zoobauru consulta comprovada em 08/12/2002. O segundo site é o SIHOP, Sistema de Horário de Professores, que pode ser acessado no endereço http://www.sihop.kit.net e que pode ser encontrado em 1º lugar no AltaVista, 1º lugar no *Yahoo*, 1º lugar no Cadê, com palavra chave: sihop, e 1º lugar no *Yahoo*, 1º lugar no Cadê, 1º lugar no AltaVista com palavra chave: Sistema de Horário de Professores consulta comprovada em 08/12/2002. Adicionalmente, a posição de um documento digital pode ser alterada por outros sites que possam trazer em seu conteúdo interno uma pontuação maior, e outra condição é a própria modificação dos algoritmos de classificação dos *spiders* o que torna o *Webdesign* uma função cíclica na qual o gerenciamento é um item fundamental.



Bibliográfica

BERGMAN, Michael K. ***The Deep Web: Surfacing Hidden Value***. *The Journal of Electronic Publishing. The University of Michigan Press*. Vol 7, Issue 1, 2001 Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em 17 de setembro de 2002.

BHARAT, Krishna. ***SEARCHPAD: Explicit capture of search context to support web search***. *Computer Networks*, vol 33, p.493-501, 2000.

BODNER, Richard C. CHIGNELL, Mark H. CHAROENKITKARN, Nipon. GOLOVCHINSKY, Gene. KOPAK, Richard W. ***The impact of text browsing on text retrieval performance***. *Information Processing & Management*, vol 37, p.507-520, 2001.

BRICKIN, Meredith. ***Virtual Words: no interface design***. *In: Benedikt, Michael: Cyberspace first steps.* MIT *Press*: Boston 1992, p 363-383.

BRIN, Sergey. PAGE, Lawrence. ***The anatomy of a large scale hypertextual web search engine***. *Computer Networks and* ISDN *Systems*, vol 30. p.107-117, 1998.

CASTRO, Elizabeth. **HTML para a *World Wide Web***. editora Makron *Books,* São Paulo, 2000. 534p.

CHIANG, Roger H. L. CHUA, Cecil E. H. STOREY, Veda C. ***A smart web query method for semantic retrieval of web data***. Data & Knowledge Engineering, vol 38, p. 63-84, 2001.

DALAL, N.P. QUIBLE, Z. WYATT, K. ***Cognitive design of home pages: an experimental study of comprehension on the World Wide Web***. *Information Processing and Management*, vol 36, p.607-621, 2000.

DOTTA, Sílvia. **Construção de Sites**. Editora Global, São Paulo 2000. 144p.

GANDAL, Neil. ***The dynamics of competition in the internet search engine market***. *International Journal of Industrial Organization*, vol 19, p.1103-1117, 2001.

KWOK, Cody. ETZIONI, Oren. WELD, Daniel S. ***Scaling question answering to the web***. *Capes. The Gale Group*. ACM *Transactions on Information Systems*, vol 19, i3, p.242-260, 2001.

KRUG, Steve. **Não me faça pensar. Uma abordagem do bom senso à navegabilidade da Web**. editora *Market Books*, São Paulo, 2001. 187p.

HOLSCHER, Christoph. STRUBE, Gerhard. ***Web search behavior of internet experts and newbies***. *Computer Networks*, vol 33, p.337-346, 2000.

cm 1 2 3 4 5 6 7 unesp 1 12 13 14 15 16 17 18

JOHNSON, Steven. **Cultura da interface**. Editora Jorge Zahar, Rio de Janeiro, 2001. 189p.

LUZ, Iraci B. P. **Acesso à informação: um assunto polêmico**. Bauru, 1997. 110p. Dissertação (Mestrado – Comunicação e Poéticas Visuais) – Faculdade de Arquitetura, Artes e Comunicação, Universidade Estadual Paulista.

MCLUHAN, Marshall. **Os meios de comunicação como extensão do homem**. editora Cultrix, São Paulo, 1996. 407p.

SILVEIRA, Marcelo. ***Web Marketing*, Usando Ferramentas de Busca**. editora Novatec, São Paulo, 2002. 159p.

SULLIVAN, Danny. ***Web Pages Tweaking: Will the Best Rise to the Top?*** Danny Sullivan's *Search Engine Strategies* 2000 *conference*. *Capes. The Gale Group. The Information Advisor*, Vol 12, i5 p.4-10, 2000.

TU, Hsieh Chang. HSIANG, Jieh. ***An architecture and category knowledge for intelligent information retrieval agents***. *Decision Support Systems*, vol 28, p.255-268, 2000.

Autores:
Mestrando Rodrigo Ferreira de Carvalho, e-mail: flash@feb.unesp.br
Professor Dr. João Fernando Marar, e-mail: fermarar@fc.unesp.br

Índice de Figuras

Figura 1. interface do navegador *MOSAIC*..... 47
Figura 2. Jornada 568, *Personal Organizer*..... 50
Figura 3. Arquitetura Linear..... 52
Figura 4. Arquitetura Não Linear..... 53
Figura 5. Arquitetura Hipertexto..... 53
Figura 6. Política de Privacidade..... 55
Figura 7. Formas de Pagamento..... 55
Figura 8. Vendas por telefone..... 56
Figura 9. Lojas Americanas..... 56
Figura 10. Melhor navegação..... 59
Figura 11. Correio..... 60
Figura 12. Nokia..... 69
Figura 13. Livraria Cultura..... 70
Figura 14. Arquitetura de Mecanismos de Busca..... 73
Figura 15. Estrutura do Meta Buscador..... 74
Figura 16. Exemplo de uso de palavra-chave..... 76
Figura 17. Estrutura em *Frames*..... 77
Figura 18. Código Interno..... 80
Figura 19. Exemplo *Tag Alt*..... 94
Figura 20. Visão da *tag Alt*..... 95
Figura 21. Elementos não textuais..... 96
Figura 22. Zôo Bauru..... 106
Figura 23. COT..... 106
Figura 24. SHIOP..... 107
Figura 25. Diagrama..... 108
Figura 26. Opção de interface 1..... 111
Figura 27. Opção de interface 2..... 111
Figura 28. Opção de interface 3..... 111
Figura 29. Opção de interface 4..... 111
Figura 30. Opção de interface 5..... 111

Figura 31. Opção de interface 6..... 111
Figura 32. Opção de interface 7..... 112
Figura 33. Opção de interface 8..... 112
Figura 34. Opção 5 ampliada..... 112
Figura 35. Interface atual do *site* SACI..... 113
Figura 36. Logo marca aplicada nas páginas do *site* SACI..... 117
Figura 37. Palavras-chave na estrutura interna do *site* SACI..... 118

Índice de Gráficos

Gráfico 1. Comparativo de Crescimento de Bases de Dados..... 23

Gráfico 2. Comparativo da indexação do *site* SACI em mec. de busca..... 130

Índice de Tabelas

Tabela 1. Dimensão dos Mecanismos de Busca..... 39
Tabela 2. Dimensão dos Diretórios..... 39
Tabela 3. Propriedades de Classificação do AltaVista..... 78
Tabela 4. Propriedades de Classificação do *WebCrawler*..... 81
Tabela 5. Propriedades de Classificação do *Excite*..... 82
Tabela 6. *Tag Title*..... 120
Tabela 7. *Meta-Tags*..... 120
Tabela 8. Datas de Indexação..... 124
Tabela 9. Informações complementares sobre os mecanismos..... 125
Tabela 10. Classificação dos mecanismos de busca e palavras-chave..... 128
Tabela 11. Visualização das Palavras-chave encontradas..... 129
Tabela 12. Visualização da Palavra-chave e posição de classificação..... 131

0810001181