Arquitetura de informação para sistemas distribuídos
Information architecture for Distributed Systems

Rodrigo Ferreira de Carvalho[1]
João Fernando Marar[2]
José Olympio Pinheiro[3]

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,* Sistemas Distribuídos, Arquitetura de Informação.

Abstract:

The Internet impact has reached a great number of users, consequently, it is generating a very big data information; however, this same information is not always easy to find. To search for information, web search engines classification algorithms need better adaptations in order to guarantee the right information, applications in artificial intelligence, etc. In this way, Web designers can act in a decisive way, providing a better answer in the search engines classification. So the objective of this paper is to describe procedures to improve the digital document classification, which are available to the person responsible for the creation of the site.

Keywords: Web design, Distributed Systems, information Architecture.

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[4],

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 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).

Desta forma, 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 variados tipos de suportes eletrônicos. Quando estes sistemas se combinam, a informação tem uma chance maior de se tornar conhecimento, muito mais rapidamente que qualquer uma das formas já citadas individualmente.

2- 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 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".

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.

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[5], responsável por permitir que o usuário encontre o que procura com o menor número de interações possíveis.

3- 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 da maneira que foi planejado. Ao contrário, desperdiça-se muito tempo na pesquisa e, muitas vezes, não se encontra nela aquilo que se deseja. Assim, 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. Os mecanismos de busca, que são responsáveis pelo rastreamento, cadastramento e indexação 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 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*.

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. O restante é atribuído, por exemplo, a linguagem XML, ou *Javascript* e também a 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 do código HTML 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).

4- Técnicas de Auxílio à Classificação de Documentos Digitais

Pesquisas desenvolvidas (Carvalho, 2003, p.114) comprovam que para que um documento digital possa ter relevância na classificação é necessário uma série de elementos combinados simultaneamente para torná-lo acessível. Tais técnicas abordaram:

- *Meta tag* de descrição**:** descrição do conteúdo do material disponibilizado no documento digital. <*META NAME="Description" CONTENT*="descrição_da_página_ou_site">

- *Meta tag keyword*: descrição das possíveis palavras-chaves que podem dar acesso ao conteúdo.
<*META NAME="Keywords" CONTENT*="palavras_separadas_por_vírgula">

- *Meta robot*: descrição para o programa do mecanismo de busca (*spider*) ser convidado a classificar a página e os *links* do documento digital.

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

A sintaxe do comando é discriminada a seguir: *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*;

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.

- *Meta tag* de identificação de idioma: para que o material possa ser classificado em *clusters* de idioma selecionado.
<*META HTTP-EQUIV="Content-Language" CONTENT="br"*>

Além de outras que podem ser utilizadas dependendo do objetivo.

- *Tag Title*: *Tag* de título, um importante parâmetro que identifica ou que pode identificar o assunto do documento digital. Esta *tag* é utilizada para identificar na barra de topo do navegador, o *site*, produto ou informação que trata o documento, é uma das primeiras *tags* que são lidas pelos *spiders* dos mecanismos de busca.

- *Tags Alt*: *Tag* de texto alternativo, esta *tag* quando bem utilizada pode, além de oferecer melhor navegação ao usuário, pois pode oferecer dicas do que está do outro lado do *link* sem que o usuário efetue o *link* apenas colocando o *mouse* por cima do botão e/ou imagem. Neste caso, mostrando uma caixa de texto com uma breve descrição do que poderá ser encontrado se o *link* for efetuado. Deve ser comentado que isso poderá acontecer se o responsável pelo desenvolvimento planejou o uso adequado da respectiva *tag*. Além disso, o conteúdo da *tag alt* pode ser visualizado quando por algum motivo o navegador não estiver ativado para mostrar as imagens do ambiente gráfico, possibilitando a navegação em modo texto (através das identificações da *tag alt*). E finalizando este item, o que torna a *tag alt* importante para o *site* e para os mecanismos de busca, é a aplicação da palavra-chave e / ou categoria chave em seu interior. desta forma, realizando positivamente a pontuação dentro da classificação das bases de informação.

- Nomenclatura de arquivos e pastas de forma orgânica:Todos os elementos relacionados ao mesmo documento, com por exemplo, pastas, subpastas e arquivos sejam de imagem, ou arquivos HTML, ASP, SWF, etc, devem possuir a aplicação de um nome referente a palavra-chave e / ou categoria-chave, para que também possam realizar a pontuação em relação a classificação nos mecanismos de busca.

- Textos visíveis na interface com o usuário: O texto que aparece no navegador também é classificado nas bases e se nesse texto a palavra-chave estiver contida, o mesmo proporcionará possibilidades de pontuação do material. Outro detalhe é que, quanto mais a palavra-chave estiver próximo do topo da página, mais relevância ela fornecerá para a pontuação no mecanismo de busca (este é um dos vários fatores relacionados ao *webwriting*).

- Analise dos *sites* concorrentes: A analise dos *sites* concorrentes deve ser realizada para verificar a quantidade de palavras-chaves que foram utilizadas para que esses mesmos documentos digitais pudessem ser classificados em posições relevantes. Neste caso, um detalhe fundamental a observar é se o *site* classificado tem ou não sua posição otimizada através de compra de posição. Essa analise é importante, pois com ela se pode chegar a um coeficiente referente à quantidade de palavras-chaves que devem ser utilizadas para que um novo *site* possa estar entre aqueles que se classificam em boas posições. Assim, da mesma forma que se pode fazer um documento digital ser classificado em posições mais otimizadas, os mesmos concorrentes podem adotar um processo continuo para que seus materiais estejam sempre atualizados em relação à informação e a classificação.

4.1- Estudo de viabilidade da Técnica

Em um período de dois meses (fevereiro a abril de 2004) 86 alunos do curso de Informática do Colégio Técnico Industrial da Unesp de Bauru, desenvolveram 86 *sites* institucionais, ao qual foram empregados as técnicas descritas do código HTML para a descrição das informações contidas no documento digital. O prazo para o envio do

documento digital foi estipulado para o final da 4ª semana, pois o tempo previsto para cadastramento e indexação de informações nos mecanismos de busca podem variar, e o tempo mínimo para cadastramento gratuito está entre 3 e 4 semanas. Desta forma, ao final do período de dois meses já seria possível colher resultados das classificações obtidas.

Entretanto, ao final da 4ª semana, apenas sete *sites* foram enviados dentro do período e os outros 86 foram enviados entre a 5ª e 6ª semanas, portanto fora do período mínimo para classificação. Estes 86 *sites* possuem boas chances de serem classificados pois, também utilizaram as técnicas descritas no item 4. Mas para efeito de nossa pesquisa, serão apenas relatados os dados obtidos dos documentos digitais que seguiram as recomendações iniciais. Desta forma, dos sete *sites* enviados dentro do período previsto, seis foram classificados em posições relevantes (classificados entre um dos 10 primeiros itens da página de resposta do mecanismo de busca) e, um outro *site* classificado em 12ª posição. Ou seja, dos *sites* que foram enviados dentro do prazo previsto, 85 por cento foram classificados em primeiras posições.

5- conclusão

5.1- Resultados obtidos

O que podemos observar, foi que um tempo mínimo de quatro semanas é necessário para conseguir uma classificação nos mecanismos de busca, se os procedimentos descritos no item 4 forem utilizados, pois levando em consideração que a proposta de classificação foi desenvolvida sem que se haja custos para a classificação dos documentos digitais. O que pode colaborar para que informações de âmbito não apenas comerciais possam estar bem classificadas, e assim, outros conteúdos possam ter a chance de serem encontrados de maneira a provocar um modo mais otimizado de procurar e encontrar, utilizando um tempo mínimo para a procura.

Outro detalhe que pode ser notado foi o número de classificação dos itens enviados no prazo estipulado, com exceção de um documento digital que ficou classificado em 12ª posição, todos os outros foram classificados entre os primeiros 10 itens listados, comprovando desta forma a eficiência de se usar simultaneamente vários recursos de identificação do documento digital.

É oportuno relatar que a obtenção de uma classificação relevante usando parâmetros do código HTML e uma arquitetura de informação orgânica, onde cada elemento individual contido no *site* possa colaborar para a classificação do documento digital, torna-se fundamental para que a informação seja encontrada de maneira a oferecer rapidez no processo de pesquisa e retorno de informações relevantes. Adicionalmente, se estes métodos forem utilizados os responsáveis pelo documento digital não precisam arcar com despesas adicionais para que seus conteúdos possam estar classificados em posições relevantes.

6- Bibliografia

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.

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

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

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

CARVALHO, Rodrigo Ferreira de. 2003. 194p. Dissertação (Mestrado – Desenho Industrial) – Faculdade de Arquitetura, Artes e Comunicação, Universidade Estadual Paulista.

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.

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

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.

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

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.

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.

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.

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

(1) Ms, CTI – Colégio Técnico Industrial, Unesp – Bauru, SP, e-mail: flash@feb.unesp.br.
(2) Dr, FC – Faculdade de Ciências, Dpto. Computação, Unesp – Bauru, SP, e-mail: fermarar@fc.unesp.br.
(3) Dr, FAAC – Faculdade de Arquitetura Artes e Comunicação, Dpto. Representação Gráfica, Unesp – Bauru, SP
e-mail: holihn@uol.com.br.

(4) 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. Disponível em: <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm> ou <http://www.unicamp.br/~hans/mh/memex.html>. Acesso em: 22 jan. 2005.
(5) Arquitetura de informação, a estrutura e organização lógica de funcionamento de um sistema computacional.