Arquitetura de informação:
sistemas distribuídos

Rodrigo Ferreira de Carvalho
João Fernando Marar

MENEZES, MS., and PASCHOARELLI, LC., orgs. *Design e planejamento*: aspectos tecnológicos [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 277 p. ISBN 978-85-7983-042-6. Available from SciELO Books <http://books.scielo.org>.

8
Arquitetura de informação: sistemas distribuídos

Rodrigo Ferreira de Carvalho[1]
João Fernando Marar[2]

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 se torne 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 deve 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 nesse campo do conhecimento. Nessa época, Vannevar Bush 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 dessas informações de maneira rápida e eficiente (Johnson, 2001).

---

1 Mestre em *design*, Universidade Estadual Paulista.
2 Livre-docente, Universidade Estadual Paulista.

O volume de publicações, contudo, cresceu tanto que tomar conhecimento das novas técnicas e 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. Isso gerou a necessidade de uma instituição mais dinâmica, que se antecipasse às demandas dos usuários e que, além de selecionar, processar e armazenar o acervo, intermediasse também o fluxo da informação (Luz, 1997).

Assim, 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 e livros poderiam estar bem ordenadas ou indexadas em estantes, mas a criação constante de novas informações exigia cada vez mais espaço. Para eliminar esse 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 de forma alfabética ou numérica ineficiente. O pensamento é mantido em uma teia de conhecimento no cérebro. Assim, seria ideal encontrar uma forma de fazer algo análogo de forma automatizada (Gardner, 1999).

A informação pode implicar 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 esses sistemas se combinam, a informação tem uma chance maior de tornar-se conhecimento muito mais rapidamente que qualquer uma das formas já citadas individualmente.

Sistemas distribuídos como suporte à segurança de informação

A arquitetura desenvolvida para o funcionamento da transmissão de conteúdo por meio da *internet* foi elaborada 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, consequentemente, parasse toda a comunicação realizada por meio da rede formada pelos computadores. É o que se chama de sistema distribuído em rede ou hipermídia "distribuída".

Dessa forma, era possível um computador acessar informações contidas em 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. Ampliava-se, assim, o alcance do ser humano e começava-se a deixar virtualmente a distância da informação a um clique do usuário.

Por meio 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 se 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, responsável por permitir que o usuário encontre o que procura com o menor número de interações possíveis.

O problema: otimizar as possibilidades de classificação de documentos digitais e encontrar informação segura

O propósito da *internet* sempre foi o armazenamento de informação por meio de um acesso rápido. Mas com o passar do tempo, podemos notar que seu funcionamento não atingiu plenamente esse requisito da maneira como 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 et 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 et al. (2001, p.242), a crescente base de dados amplia e dificulta o rastreamento de informações, tornando uma pesquisa simples na *web* uma tarefa às vezes problemática, ou pela falta ou porque se encontra 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% de todo o 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.

Dessa forma, quando os atributos de identificação do código HTML são utilizados incorretamente, ou não são utilizados, as chances de uma boa classificação 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. 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 por meio 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 análise 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).

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 necessária 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-chave 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 *link*s do docu-

mento 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 *spider*;

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">

Há 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. Essa *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, essa *tag*, quando bem utilizada, pode, além de oferecer melhor navegação ao usuário, 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. Nesse caso, mostra 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 (por meio 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, 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, como, por exemplo, pastas, subpastas e arquivos, sejam de imagem ou arquivos HTML, ASP, SWF etc, devem possuir a aplicação de um nome referente à palavra-chave e/ou categoria chave para que também possam realizar a pontuação em relação à 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, ele proporcionará possibilidades de pontuação do material. Outro detalhe é que quanto mais a palavra-chave estiver próxima do topo da página, mais relevância ela fornecerá para a pontuação no mecanismo de busca (esse é um dos vários fatores relacionados ao *webwriting*).
- Análise dos *sites* concorrentes: a análise 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. Nesse caso, um detalhe fundamental a observar é se o *site* classificado tem ou não sua posição otimizada por meio de compra de posição. Essa análise é importante, pois com ela se pode chegar a um coeficiente referente à quantidade de palavras-cha-

ves 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 contínuo para que seus materiais estejam sempre atualizados em relação à informação e a classificação.

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, nos quais 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 quarta semana, pois o tempo previsto para cadastramento e indexação de informações nos mecanismos de busca pode variar, e o tempo mínimo para cadastramento gratuito está entre três e quatro semanas. Dessa forma, ao final do período de dois meses já seria possível colher resultados das classificações obtidas.

Entretanto, ao final da quarta semana, apenas sete *sites* foram enviados dentro do período, e os outros 79 foram enviados entre a quinta e a sexta semanas, portanto fora do período mínimo para classificação. Esses 79 *sites* possuem boas chances de serem classificados, pois também utilizaram as técnicas descritas anteriormente. Mas para efeito de nossa pesquisa, serão apenas relatados os dados obtidos dos documentos digitais que seguiram as recomendações iniciais. Dessa forma, dos sete *sites* enviados dentro do período previsto, seis foram classificados em posições relevantes (classificados entre um dos dez primeiros itens da página de resposta do mecanismo de busca), e o outro *site* foi classificado em 12ª posição. Ou seja, dos *sites* que foram enviados dentro do prazo previsto, 85% foram classificados em primeiras posições.

Conclusão

O que pudemos observar foi que um tempo mínimo de quatro semanas é necessário para conseguir uma classificação nos mecanismos de busca, se os procedimentos descritos anteriormente forem utilizados, pois levando em consideração que a proposta de classificação foi desenvolvida sem 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 pôde 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 dez itens listados, comprovando dessa forma a eficiência de 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, em que 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 esses métodos forem utilizados, os responsáveis pelo documento digital não precisam arcar com despesas adicionais para que seus conteúdos possam estar classificados em posições relevantes.

Referências bibliográficas

BERGMAN, M.K. The deep *web*: surfacing hidden value. *The Journal of Electronic Publishing*. The University of Michigan Press. v.7, Issue 1, 2001. Disponível em: <http://www.press.umich.edu/jep/07-01/bergman.html>. Acesso em: 17 set. 2002.

BHARAT, K. Searchpad: explicit capture of search context to support web search. *Computer networks*, v.33, p.493-501, 2000.

BLACK, R. *Websites que funcionam*. São Paulo: Quark, 1997.

BONSIEPE, G. *Design do material ao digital*. Florianópolis: Fiesc/Iel, 1997.

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

CHANG, Y. S.; YUAN, S. M.; LO, W. A new multi search engine for querying data through an internet search service on CORBA. *Computer networks*, v.34, p.467-80, 2000.

DONDIS, D. A. Sintaxe da linguagem visual. São Paulo: Martins Fontes, 2000.

GANDAL, N. The dynamics of competition in the internet search engine market. *International Journal of Industrial Organization*, v.19, p.1103-17, 2001.

GARDNER, H. *Inteligência, um conceito reformulado*. Rio de Janeiro: Objetiva,1999.

JOHNSON, S. *Cultura da interface*. Rio de Janeiro: Zahar, 2001.

KRUG, S. *Não me faça pensar*. Uma abordagem do bom senso à navegabilidade da *web*. São Paulo: Market Books, 2001.

KWOK, C.; ETZIONI, O.; WELD, D. S. *Scaling question answering to the web*. Capes. The Gale Group. ACM Transactions on Information Systems, v.19, i3, p.242-60, 2001.

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

NIELSEN, J. *Projetando websites*. Designing web usability. Rio de Janeiro: Campus, 2000.

NIELSEN, J.; TAHIR, M. *Homepage*: Usabilidade. 50 *websites* desconstruídos. Rio de Janeiro: Campus, 2002.

SILVEIRA, M. *Web marketing*: usando ferramentas de busca. São Paulo: Novatec, 2002.