Quem procura, acha?
o impacto dos buscadores sobre o modelo distributivo da World Wide Web

Suely Fragoso[*]

Resumo: *O artigo discute as implicações do impacto dos buscadores sobre o modelo muitos-muitos de distribuição de informação na World Wide Web. Para isso, revisa brevemente e de forma crítico-descritiv a história dos sistemas de busca, desde os primeiros buscadores até as práticas colaborativas. Em paralelo a esta revisão discute a vinculação entre os buscadores e o mercado publicitário. A seguir, apresenta a questão da concentração do tráfego na web em torno de um pequeno número de sistemas de busca, os quais pertencem a um número igualmente reduzido de grupos empreendedores. Finalmente, aborda a confiança que os usuários depositam nos buscadores para concluir que os sistemas de busca representam uma importante pressão verticalizadora, capaz de aproximar do modelo massivo a experiência da maioria dos usuários da WWW.*

Palavras-chave: *internet, web, buscadores, search engines, google, yahoo*

O século XX foi o século da comunicação de massa, durante o qual a imprensa, o cinema, o rádio e a televisão floresceram conforme o modelo irradiativo (um-muitos) de distribuição. Tecnologias de comunicação originalmente vocacionadas para funcionamento epidêmico (muitos-muitos) chegaram a ser reencaminhadas para o modo irradiativo. Ao final dos anos 1990, entretanto, uma nova prática emergiria das instalações militares e dos *campi* universitários: a comunicação mediada por computador (CMC). À primeira vista, parecia não se tratar de muito mais que a transposição para um novo ambiente tecnológico de alguns modos pré-existentes de comunicação interpessoal (um-um), como o correio ou o telefone. Entretanto, a configuração tecnológica (em rede) e o ambiente cultural (tanto o espaço universitário quanto a proximidade entre a comunidade hacker e os movimentos da contracultura) eram altamente propícios à comunicação epidêmica (muitos-muitos), que de fato viria a florescer. Com a popularização da internet, e em especial através da *World Wide Web*, as possibilidades de comunicação muitos-muitos estenderam-se a um número sem precedentes de pessoas. Em um contexto até então marcado pela hegemonia aparentemente intransponível do modelo massivo de comunicação, à época era praticamente impossível não saudar o potencial 'subversivo' da CMC.

[*] Unisinos, <suely@unisinos.br>

Embora os números absolutos obscureçam o fato de que apenas uma reduzidíssima parcela da população mundial tem pleno acesso às redes digitais de comunicação, é inegável que a CMC elevou exponencialmente o número de indivíduos capazes de desempenhar o papel de emissor em processos comunicacionais de grande escala, provocando um rearranjo no cenário midiático. Sem deixar de louvar os méritos dessa nova modalidade de comunicação tecnológica, é importante atentar também para os desdobramentos negativos do modelo muitos-muitos.

Antes de mais nada, um grande número de emissores implica um elevado número de mensagens. Em um texto que já se tornou um clássico do tema, Lawrence e Giles estimaram em 800 milhões o número de páginas indexáveis[1] disponíveis na web em 1999 (Lawrence e Giles, 1999, p.2). Um ano mais tarde, Murray calculava que o número de páginas indexáveis já teria ultrapassado os dois bilhões (Murray, 2000, p.3). Em janeiro de 2005, Gulli e Signorini calcularam a existência de pelo menos 11,5 bilhões de páginas (Gulli e Signorini, 2005, p.1). Não bastasse a grandeza desses números, é preciso lembrar que a web é essencialmente dinâmica e auto-organizada. No mesmo ano de 2000 em que o incremento diário no número total de páginas era estimado em cerca de 7,3 milhões (Murray, 2000, p. 3), Arasu *et al* constataram que a meia-vida das páginas com domínio '.com' não ultrapassava dez dias (Arasu *et al*., 2001, p.3). Além disso, é preciso considerar a imensa variedade de linguagens empregadas nas páginas (textos, sons, imagens estáticas e dinâmicas) e o dinamismo de seu conteúdo.

O cenário assim constituído é de uma tal exuberância que traz para o primeiro plano a diferença crucial entre a multiplicação das pessoas capazes de 'publicar' na *World Wide Web* e a visibilidade de cada uma delas. A questão não se resume à qualidade ou pertinência do material disponibilizado, mesmo porque é fundamental respeitar as diferentes concepções de pertinência. Na hipótese – altamente fantasiosa – de que todos os milhões de *terabytes* da web interessassem a todos e a cada um, o problema do excesso não se resolveria, pelo contrário. Na ausência de um controle por *gatekeeping*, 'na entrada', como é de praxe nos meios de comunicação analógicos, o ambiente muitos-muitos da web

[1] A expressão 'páginas indexáveis' designa o conteúdo da web normalmente acessível às ferramentas de busca. As páginas não-indexáveis compõem a web profunda (*deep web*), que agrega as páginas que não enviam (ou recebem) links; o conteúdo dinâmico, gerado em resposta a consultas a bancos de dados e o material de acesso restrito.

favorece a emergência de mecanismos de filtragem e de seleção 'na saída'. Nesse cenário, os sistemas de busca configuram uma solução óbvia e aparentemente inócua. Entretanto, não é exagero dizer que seus desdobramentos, sobretudo quando se leva em conta a configuração que assumiram nos últimos anos, põem em risco o próprio formato epidêmico da WWW. Para esclarecer devidamente esta última colocação, que corresponde à proposição fundamental deste texto, vale a pena revisitar algumas passagens da história dos sistemas de busca na internet.

Uma breve (e incompleta) história (comentada) dos buscadores

A necessidade de orientação em meio à profusão de material disponibilizado na internet é anterior à *World Wide Web*: o primeiro indexador, denominado ***Archie*** (1990), surgiu em 1990. Reunia informações de arquivos disponíveis em servidores ftp anônimos e mantinha-os atualizados checando os dados em intervalos de até 30 dias. Os usuários do *Archie* procuravam por sequências de caracteres nos nomes dos arquivos ou pastas disponíveis no índice. Inicialmente destinado a uso departamental, *Archie* foi anunciado publicamente quando abrangia pouco mais de 200 servidores (Deustch, 1990).

A facilitação da localização dos arquivos disponíveis para ftp pelo *Archie* inspirou a criação de um indexador semelhante para Gopher, que foi chamado ***Veronica*** (1992). *Veronica* era um banco de dados que reunia os menus dos servidores Gopher, permitindo a realização de buscas por tópico (com palavras-chave) ao invés de por servidor (como era inerente ao sistema). Pouco depois apareceu ***Jughead*** (1993), que teve o mérito de introduzir a possibilidade de realizar buscas booleanas (Salient Marketing, s.d.)

Um outro sistema, em vários aspectos mais avançado e reunindo características do próprio Gopher e dos buscadores que nele operavam, já estava em operação desde o ano anterior. Era o WAIS (*Wide Area Information Server*, 1992), desenvolvido por iniciativa conjunta de 4 empresas. Com o WAIS, era possível realizar buscas em bases de dados remotas, cujos resultados eram organizados em ordem decrescente de frequência das palavras-chave. Clientes WAIS foram criados para vários sistemas operacionais, incluindo Windows, Macintosh e Unix, mas a propriedade privada 'segurou' a popularização do WAIS. De fato, podia ser arriscado, à época, contradizer o caráter público da internet. Diversas boas idéias e implementações competentes sucumbiram devido à insistência em comercializá-las. Mesmo assim, é de se duvidar que o CERN (http://www.cern.ch) tivesse idéia da escala que

assumiriam as consequências de sua decisão de abrir mão, em 1993, do direito de propriedade dos códigos básicos do projeto de um sistema global de hipertexto que havia sido iniciado por Tim Berners-Lee em 1989 e que viria a tornar-se a *World Wide Web* como a conhecemos hoje. Combinado com a decisão de tornar a WWW um sistema de domínio público, o lançamento do primeiro browser para Windows, o *X Windows Mosaic* (1993) e sua posterior adaptação para plataformas Macintosh, ajudou a popularizar a web numa escala sem precedentes para todos os demais sistemas de informação.

Poucos meses após o lançamento do *Mosaic*, a primeira aranha começou a rastrear a web. Era o ***World Wide Web Wanderer*** (1993), o primeiro webrobot[2]. O *Wanderer* percorria a web mapeando cada página de um site e prosseguindo para uma das páginas conectadas a ela, para então mapeá-la e prosseguir para a próxima e assim sucessivamente[3] e armazenava os endereços que encontrava num banco de dados. A idéia inicial era mapear toda a web (Gray, 1995) e partia da premissa de que todas as páginas estariam conectadas a pelo menos uma outra, de modo que seria uma questão de tempo até que o *Wanderer* percorresse a web inteira[4].

Apesar da controvérsia causada pelo impacto da operação do *WWW-Wanderer* sobre os servidores da rede, antes do final de 1993 pelo menos mais três outros bots rastejavam pela web: *JumpStation*, *WWW-Worm* e *RBSE*. O ***Worm*** indexava os títulos e endereços das páginas, enquanto o ***JumpStation*** inovou ao arquivar também os cabeçalhos. Ambos apresentavam os resultados na ordem em que os encontravam. O ***RBSE*** foi o primeiro a implementar um sistema de ranqueamento baseado na relevância relativa à expressão utilizada para a busca (Mauldin, 1997; Wall, 2006).

Ainda em 1993 surgiu o primeiro indexador projetado especificamente para a web, o ***Aliweb.*** Fortemente inspirado pelo Archie, não possuía um rastreador, mas compunha seu banco de dados a partir das informações fornecidas diretamente pelos *webmasters*. Isso permitia que o sistema arquivasse descrições das páginas, que eram alimentadas pelos

[2] Webrobots, também chamados *crawlers*, *spiders* e, daqui para a frente referidos como rastreadores ou bots, são programas que percorrem a web passando de um documento para outro através dos hiperlinks.

[3] Esse tipo de rastreamento é conhecido como '*depth-first*' (em profundidade) e implica que o rastreador retorna à página inicial diversas vezes, o que coloca grande demanda sobre os servidores, comprometendo seu desempenho. Uma outra abordagem possível é a '*breadth-first*' (em abrangência), em que o rastreador segue todos os links de uma página e só depois prossegue para os links das páginas seguintes.

[4] A crença de que todos os endereços estão ao alcance de quem – ou o que – percorresse os links perdurou até recentemente, quando foi matematicamente demonstrado que a natureza direcional das hiperconexões da web implica necessariamente em sua fragmentação (Barabási, 2002, p. 167).

próprios criadores, mas por outro lado tornava a qualidade e atualidade do banco de dados dependentes da boa vontade de terceiros.

Também contando com um banco de dados construído sem o apoio de rastreadores, surgiu no ano seguinte o primeiro diretório web pesquisável, o ***Galaxy***. Como listava apenas URLs que tinham sido fornecidas diretamente, o Galaxy pôde organizar os endereços em categorias e sub-categorias, permitindo que os usuários restringissem a busca a sub-áreas de sua base de dados, o que acelerava e tornava mais preciso o processo.

Não demorou a surgiu um bot capaz de associar o registro do conteúdo completo das páginas à funcionalidade do rastreamento automático. Para fazê-lo, o ***WebCrawler*** (1994) adotou a indexação vetorial[5]. A estratégia foi um grande sucesso: após seis meses de uso, o *WebCrawler* já havia indexado milhares de documentos e efetuado quase um quarto de milhão de buscas, atribuídas a mais de 23 mil usuários diferentes (Pinkerton, 1994). Em novembro do mesmo ano, o número de buscas realizadas chegou à marca de um milhão (Pinkerton, s.d.). Logo o sistema da universidade de Washington deixou de ser capaz de dar suporte ao buscador, um problema que só seria resolvido com a venda do *WebCrawler*.

Outros sistemas de busca aperfeiçoaram ainda mais a combinação de funcionalidade e abrangência inaugurada pelo *WebCrawler*. Um dos mais significativos foi o ***Lycos*** (1994), que além de organizar os resultados das buscas conforme sua relevância, permitia consultas por prefixo e dava bônus por proximidade entre palavras (Mauldin, 1997). Um dos atrativos iniciais do *Lycos* foi o tamanho de seu banco de dados, cujo peso era aliviado pela estratégia de não arquivar o conteúdo completo das páginas mas apenas um resumo, que era construído automaticamente considerando as 100 palavras-chave mais freqüentes em cada página, combinadas com as palavras do título, do cabeçalho e as 20 primeiras linhas ou os primeiros 10% do documento. Os resumos podiam ser vistos junto com a lista dos resultados e ajudavam o usuário a decidir qual das páginas encontradas visitar primeiro.

Outro diferencial importante do *Lycos* foi o funcionamento de seu rastreador, que não operava *depth-first* nem *breadth-first*, mas conforme uma estratégia que Mauldin denominou '*best-first*'. Para definir qual era a 'melhor' página, e portanto a próxima a ser

[5] No modelo vetorial de indexação, documentos em linguagem natural são representados através de vetores (no caso, palavras-chave que funcionam como termos de indexação aos quais são atribuídas características vetoriais). O sistema avalia a relevância dos documentos conforme sua relação espacial com as palavras-chave utilizadas para a busca.

rastreada, a aranha do *Lycos* levava em conta o número de links que cada página recebia de outros servidores (*inlinks*).

Em meados dos anos 1990, a capacidade da web para atrair volumes significativos de tráfego começava a chamar a atenção de novos investidores. Os buscadores foram considerados particularmente interessantes pelo capital publicitário, inicialmente interessado em incluir *banners* e pequenos anúncios nas páginas de início. Logo os sistemas de busca descobriram que a intensificação do fluxo de público era o caminho para atrair mais anunciantes. Com vistas a gerar seu próprio tráfego e incrementar o tempo de permanência dos usuários em seu domínio, muitos assumiram o formato de portal, passando a oferecer uma variedade de serviços. Um dos primeiros e mais bem sucedidos portais da web foi, sem dúvida, o *Yahoo!*

O ***Yahoo***! começou muito modestamente, como uma lista de sites favoritos de dois primeiranistas de doutorado da University of Stanford em 1994. A prática de publicar listas de favoritos na web era bastante comum na época, e o grande diferencial do indíce de Yang e Filo era a disponibilização de breves descrições das páginas listadas. Com o aumento do número de indicações, a lista tornou-se pouco amigável e os autores criaram uma estrutura de árvore (categorias e sub-categorias), conferindo ao *Yahoo*! o perfil de um diretório. Para responder ao crescimento da popularidade da lista, adicionaram também uma ferramenta de busca e passaram a aceitar inscrições de websites que desejassem figurar em seu banco de dados. Com menos de um ano de funcionamento, a página do *Yahoo*! celebrou seu milionésimo acesso, com visitantes vindos de quase 100 mil endereços distintos. (Yahoo! Media Relations, 2005).

Tendo estreado tarde, o ***AltaVista*** (1995) enfrentou uma competição feroz. Era, no entanto, extremamente mais rápido que as outras ferramentas disponíveis à época e prometia aos webmasters atualizar as informações recebidas em no máximo 24 horas. Foi também a primeira ferramenta que permitiu buscas a partir de perguntas formuladas em linguagem natural, buscas em newsgroups e buscas específicas por palavras associadas a imagens, títulos e outros campos do código html. Foi também a primeira ferramenta a disponibilizar buscas por *inlinks* (Sonnenreich, 1998), uma possibilidade que tendia a passar desapercebida dos usuários comuns mas com importantes implicações para o marketing.

Além disso, o *AltaVista* acrescentou um campo de 'dicas' embaixo da área de busca, o que ajudou a aumentar a fidelidade à ferramenta.

A essa altura, novas formas de integrar o conteúdo publicitário aos resultados das buscas, adaptando-se ao caráter *push* da web começavam a se popularizar. A 'inclusão paga' (*paid inclusion*), em que o webmaster paga a ferramenta de busca ou diretório para garantir que seu site seja incluído no banco de dados, já era comum quando surgiu uma versão mais elaborada, a 'classificação paga' (*paid placement*), que consiste em pagar o buscador para garantir que o site figure entre os melhor classificados em buscas por uma determinada palavra (ou várias). Em 1997, o ***GoTo*** (1997) inaugurou um novo modelo de vendas, introduzindo o modelo de 'seleção paga' (*pay-per-click*), em que os anunciantes só pagam ao buscador quando o link para o seu site (do anunciante) é selecionado. Rapidamente, os sistemas de busca se tornaram os principais veículos para a publicidade online (FutureNow, Inc, 2003, p. 15).

O próprio sucesso do negócio de buscas fomentou a concorrência, e logo havia dezenas de buscadores diferentes na rede. Cada um deles operava com interface e algoritmos próprios e seus bancos de dados cobriam diferentes porções da Web. Por conseguinte, consultas a sistemas diferentes produziam resultados diferentes, e os usuários passaram a repetir as mesmas consultas em várias ferramentas, buscando maior amplitude de resposta. Para atender a essa nova demanda surgiram as ferramentas de meta-busca, que permitem buscar em vários sistemas de busca ao mesmo tempo. Os dois primeiros sistemas de meta-busca apareceram quase simultaneamente, em 1995. ***Savvy Search*** realizava buscas em até 20 outros buscadores por vez e inclusive permitia acesso a alguns diretórios temáticos. No entanto, simplesmente ignorava as opções avançadas dos vários sistemas de busca. Já o ***MetaCrawler***, que se tornaria mais popular, enfrentava as diferenças de sintaxe entre as opções avançadas dos sistemas de busca criando sua própria sintaxe e convertendo o *input* do usuário no comando correspondente em cada sistema de busca acessado. No sentido inverso, os resultados encontrados eram convertidos para um formato único na página de resposta (Selberg e Etzioni, 1995).

Do ponto de vista dos sistemas de busca originais os meta-buscadores eram uma péssima idéia, pois desviavam o público de suas páginas e por conseguinte afastavam os anunciantes. Junto aos usuários, entretanto, fizeram grande sucesso – em especial o

MetaCrawler, que logo ultrapassou a capacidade dos servidores do campus da University of Washington, tendo sido então licenciado para a *go2net*, que mais tarde se tornaria *InfoSpace*. Sob a gestão da *InfoSpace*, o *MetaCrawler* encontrou um modelo compatível com a meta-busca, passando a disponibilizar os resultados das várias ferramentas acompanhados dos anúncios originais de cada site. O grande impulso comercial para os meta-buscadores adveio, entretanto, da publicidade *pay per click*, que permitia diferenciar entre o tráfego originado pela ferramenta original e o oriundo do meta-buscador.

Em paralelo à manipulação dos resultados das buscas pela inserção de resultados pagos, surgiu também o *search spam*[6]. Do ponto de vista dos buscadores, era fundamental evitar o *spam*, pois a ocorrência de resultados improcedentes ou mal classificados afastava o público e, com ele, os anunciantes. Para isso, os sistemas de busca desenvolviam estratégias de indexação e classificação cada vez mais sofisticadas. Por outro lado, o número de inclusões pagas nas listas de resultados era cada vez maior. Logo a disseminação dessas práticas começaria a comprometer a confiança dos usuários nos sistemas de busca de um modo geral.

Àquela altura, a disputa pelo mercado parecia girar em torno do tamanho dos bancos de dados dos diferentes sistemas de busca. Números portentosos eram exibidos como argumento para a existência de grandes quantidades de usuários. Devido aos altos custos envolvidos na compilação de bancos de dados com tamanho competitivo, a sobrevivência das pequenas ferramentas tornou-se praticamente impossível. Muitas foram compradas pelos buscadores maiores, interessados tanto em aumentar ainda mais seus bancos de dados quanto, muitas vezes, em particularidades dos rastreadores e sistemas de classificação que, como de praxe na indústria da busca, as pequenas empresas mantidas em sigilo. A competição por maiores fatias do mercado publicitário era pesadíssima, mas as possibilidades de lucro também o eram. Os usuários, entretanto, haviam ficado em segundo plano, reduzidos, sob a forma de fluxo de público, a matéria-prima para negociação com os anunciantes.

No mundo acadêmico, estava em gestão um sistema de classificação que recolocava no centro da cena uma das características mais interessantes do *Lycos*: a "heurística de

[6] *Search spam* consiste em configurar o site de modo a 'enganar' os sistemas de busca para obter melhor classificação.

popularidade" (Mauldin, 1997). A estratégia foi aperfeiçoada no *BackRub*, que classificava os resultados de acordo com o número de '*back links'* que cada site recebia. O projeto cresceu rapidamente e foi renomeado ***Google*** (1998). A princípio, Page e Brin não pareciam estar interessados em criar uma empresa em torno de seu novo buscador, tanto é que tentaram vendê-lo ainda em 1998, sem sucesso. Um ano mais tarde, o *Google* continuava em versão beta, mas a reputação de ser um novo sistema de busca que fornecia resultados bastante mais confiáveis que as outras ferramentas e que não apenas não incluía resultados pagos entre os resultados orgânicos mas também utilizava um algoritmo de classificação inovador e cuja forma de atuação era de conhecimento público já começava a torná-lo um sucesso. Outros pontos fortes do *Google* eram a velocidade das buscas e a simplicidade da interface (começando pela ausência de banners e outro material publicitário, o que levava a página inicial a carregar muito mais rápido que a dos outros sites de busca). Logo o *Google* pôde enfrentar a concorrência também na batalha pelo maior banco de dados e passou a anunciar a quantidade de páginas indexadas imediatamente embaixo do campo de buscas.

Ao final de 2000, o *Google* começou a exibir alguns resultados pagos, mas, ao contrário da maiora das outras ferramentas, não os mesclou com os resultados orgânicos. Àquela altura o *Google* já havia se estabelecido como o melhor sistema de buscas na mente do público, que aceitou bem a diferenciação de gráfica entre os resultados orgânicos e os pagos. Os demais buscadores foram obrigados a encarar a superioridade da relevância dos resultados fornecidos pelo *Google* e a lealdade que aquela qualidade gerara entre os usuários: muitos outros sistemas de busca, inclusive alguns grandes como o *Yahoo*!, fariam acordos para incluir resultados vindos do *Google* em suas próprias páginas. Ao final de 2003, chegou-se a estimar que dois terços de todas as buscas realizadas na web retornavam resultados oriundos do *Google* (Thies, 2005).

Em setembro de 1999, o *Microsoft MSN Search* começara a aplicar seu próprio método de classificação aos dados obtidos junto a diferentes bancos de dados (Sullivan, 1999), dando início ao processo de desvinculação dos terceiros que até então impulsionavam suas buscas. Em 2003, a *Microsoft* anunciou a intenção de construir seu próprio rastreador (Sullivan, 2003) que só seria oficialmente anunciado dois anos mais tarde (Sullivan, 2005). Pouco mais de um ano depois, em outubro de 2006, a *Microsoft* lançou o *Windows Live Search*,

uma nova plataforma de busca com interface mais customizável e que permite a classificação dos resultados por mais recente, mais popular e mais exato) (Murray, 2006)

Em paralelo à vinda da análise de hiperlinks para o centro do palco e à entrada da *Microsoft* no negócio de buscas, os primeiros anos da década de 2000 vêm sendo marcados também pela redescoberta do potencial da criação colaborativa de listas de favoritos. A prática, que está na origem de buscadores importantes como o *Yahoo!*, ressurgiu aperfeiçoada pela marcação colaborativa, que consiste na associação de palavras-chave ao site apontado como favorito. Ferramentas baseadas em marcação social procedem buscas em bancos de dados alimentados pelos próprios usuários, tomando como base as marcações que os membros da comunidade escolheram associar aos elementos indexados. Um dos sites de social tagging mais populares é o *Del.icio.us* (http://del.icio.us), mas existem inúmeros outros.

Os sistemas colaborativos são típicos da chamada Web 2.0 e apostam no poder subversivo da 'cauda longa', uma característica há muito conhecida dos estatísticos e recentemente popularizada. A idéia da cauda longa se aplica perfeitamente à web, cuja estrutura de linkagem obedece a um padrão em que poucos sites são muito conectados enquanto a maioria dos sites recebe poucos links. Na contramão dos algoritmos que apostam na maior popularidade dos sites que concentram maior número de *inlinks*, a hipótese da cauda longa põe em foco justamente o enorme poder dos pequenos sites, cuja audiência pode, cumulativamente, superar a de um grande portal.

A força da grana

Infelizmente, no extremo oposto da cauda longa, bocas vorazes avançam sobre as esperanças de pluralização do poder na indústria das buscas. É por esta razão que, ao abordar a internet pelo ponto de vista da economia política, van Couvering enxerga na rede a mesma estrutura que caracteriza o modelo irradiativo dos meios de comunicação de massa:

> Pode-se argumentar que a internet não é um meio de massa no sentido clássico, que os milhares ou mesmo milhões de sites visíveis na web não são resultado de um processo industrial de produção e nem representam um substrato comum da vida cotidiana. (...)
> Eu sugiro que ao aceitar o argumento de que algum conteúdo é produzido em pequena escala [e escolher concentrar sua atenção nesse conteúdo] os acadêmicos estão negligenciando o estudo de um importante novo meio de comunicação de massa. (van Couvering, 2004)

De fato, o alcance global das ferramentas de busca e sua concentração nas mãos de um reduzidíssimo número de empreendedores, majoritariamente estadunidenses, ajudam a configurar um cenário extremamente semelhante ao dos grandes impérios midiáticos tradicionais. O movimento de concentração das ferramentas de busca nas mãos de alguns poucos grupos acelerou após o estouro da bolha da internet em 2000 e pode ser observado nas representações gráficas disponibilizadas por *Bruce Clay, Inc*. (Figura 1).

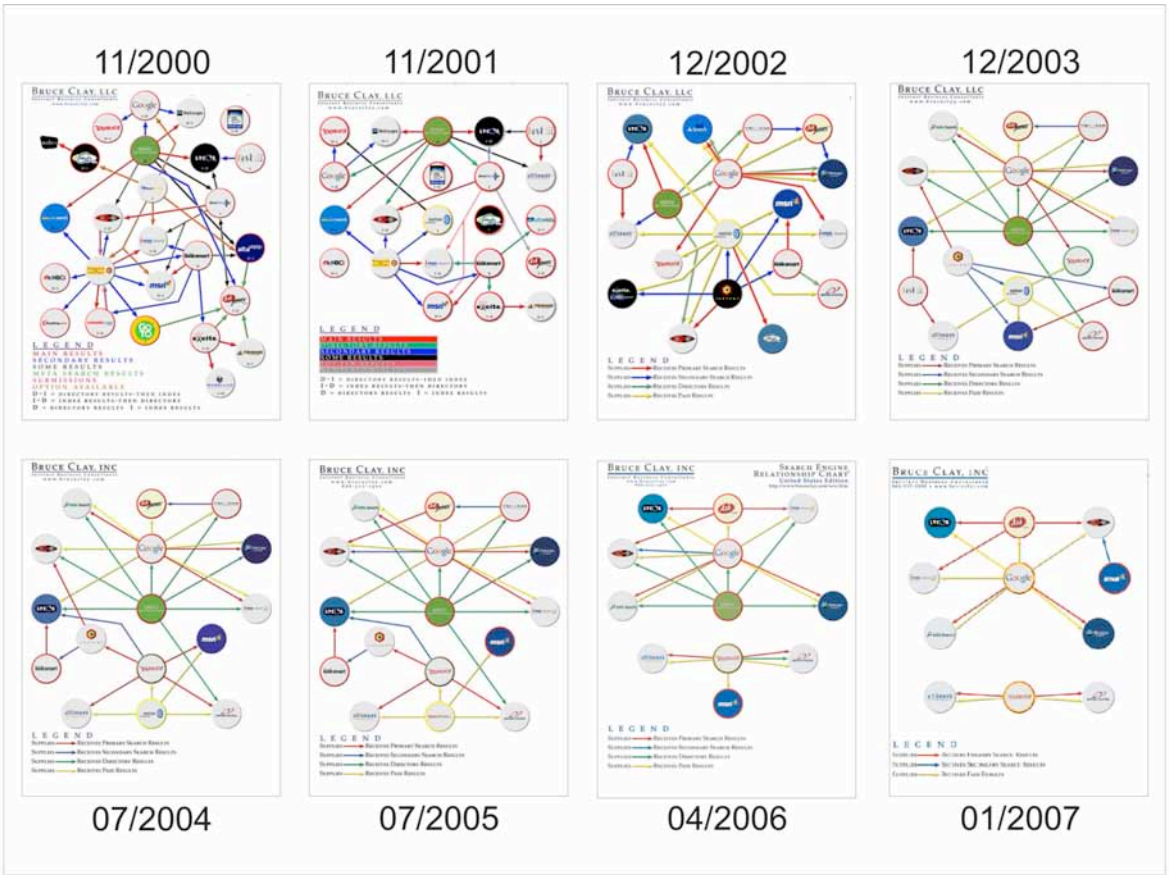

Figura 1 - Nas representações gráficas disponibilizadas por Bruce Clay, Inc. é possível visualizar a redução do número de grupos empreendedores envolvidos com o negócio das buscas na web entre os anos 2000 e 2006. ADAPTADO DE BRUCE CLAY, INC., 2006.

A concentração aparece de forma ainda mais intensa quando se passa do número geral de *players* para as relações existentes entre os onze principais buscadores identificados em janeiro de 2007: os resultados de todos provêm de apenas quatro fontes: *Google, Ask.com, MSN* e *Yahoo*! (Figura 2)

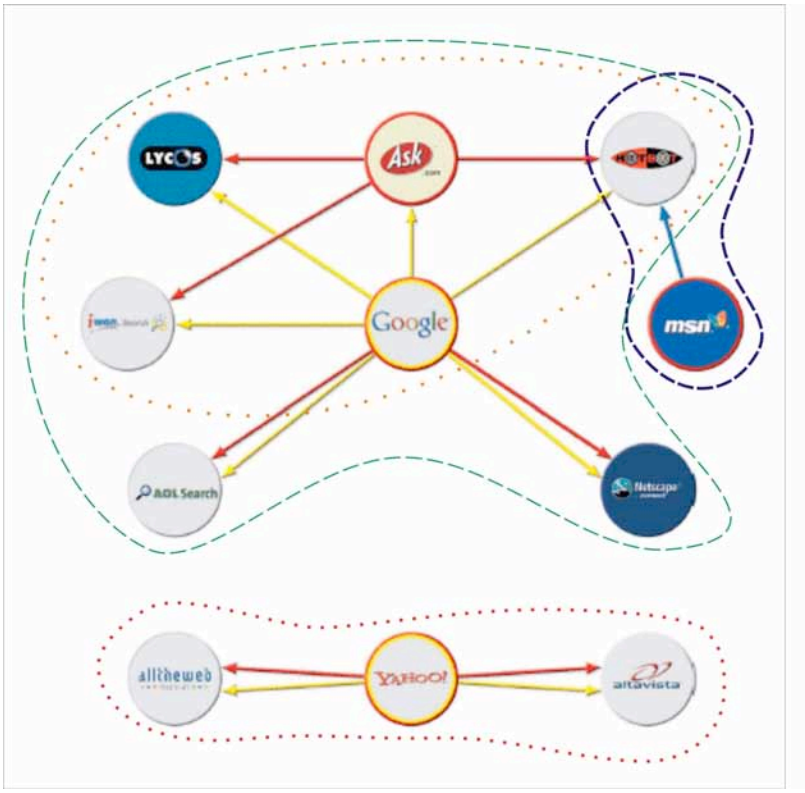

Figura 2 – Relações entre os buscadores ADAPTADO DE BRUCE CLAY, INC., 2007.

Evidentemente há uma variedade de pequenos empreendimentos de busca que não estão representados nos gráficos acima e não são levados em conta nas análises mercadológicas de van Couvering. São ferramentas experimentais ou temáticas, em sua maioria operando com bancos de dados pequenos e muitas vezes incubadas em universidades. Não seria inédito se algum deles viesse a tomar a frente da indústria das buscas no futuro – isso já aconteceu em ocasiões anteriores, por exemplo com o *AltaVista* e com o *Google*. No entanto, a crescente consolidação do negócio das buscas torna esse tipo de ocorrência cada vez mais difícil de acontecer. Como o capital da indústria das buscas provém majoritariamente da publicidade, a sobrevivência no mercado atual depende da capacidade de conquistar grandes afluxos de público. Os usuários, por sua vez, tendem a se concentrar nas ferramentas mais conhecidas.

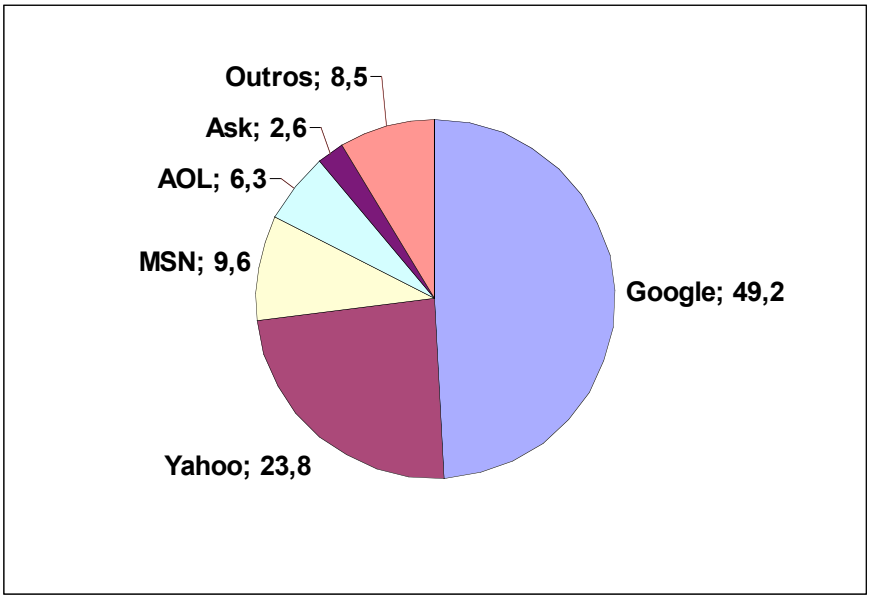

Figura 3: Porcentagens do total de buscas realizadas por usuários estadunidenses em diferentes buscadores em novembro de 2006. REPRODUZIDO DE SULLIVAN, 2006.

Incapazes de competir com as grandes no que diz respeito ao tamanho de seus bancos de dados, as ferramentas pequenas tendem a se especializar, concentrando-se em temas específicos ou na web dinâmica. Conforme uma dessas pequenas ferramentas se destaca, atrai a atenção das maiores, tornando-se uma aquisição em potencial Avanços nesse sentido já estão bastante consolidados nas ferramentas locais. Sites colaborativos também já começaram a ser adquiridos pelas grandes empresas de busca.

Os tentáculos dos maiores *players* não se restringem às outras ferramentas de busca. Inclusive o *Google*, originalmente uma alternativa ao modelo de portal, avança na direção da diversificação de atividades. A pletora hoje oferecida pelo *Google* é tão variada que sua grandeza chega a passar desapercebida pela maioria dos usuários. Para além das buscas especializadas (*GoogleFinance, Froogle*) inclui serviços como *GoogleCheckout*, *GoogleCalendar, GoogleTalk*, *Gmail* e aplicativos como *GoogleWebAccelerator*, *GoogleEarth*, *Picasa* and *GoogleDesktop.* A esta altura, o acúmulo das buscas em torno do *Google*, seus parceiros e subsidiários aponta para um perfil monopolista que tem conferido à empresa a reputação de "*Microsoft* da internet" (Mohney, 2003; Maney, 2005).

Num cenário altamente desregulamentado, o *Google* e seus concorrentes mais poderosos começam inclusive a ensaiar movimentos de convergência. Ao final de 2006, *Google,*

Yahoo! e *Microsoft* anunciaram uma primeira ação conjunta, com a adoção do *Google SiteMaps Protocol* como padrão comum às três empresas. Com essa unificação, os webmasters deixam de ter que informar separadamente os bancos de dados do *Google, Yahoo!* e *MSN* sobre suas páginas, passando a fazê-lo de forma unificada (Mills, 2006). Na prática, isso integra uma parcela dos bancos de dados das três empresas.

À mercê dos buscadores

Ano após ano, *Google, Yahoo*! e *MSN* figuram entre os dez sites mais visitados em todas as nações pesquisadas pela *Nielsen/Netratings* (htpp:/www.nielsen-netratings.com). Mais de 80% das buscas se concentram sobre essas mesmas empresas. Os usuários, por sua vez, utilizam essas ferramentas inclusive para navegar até os sites mais conhecidos:

> Existem dois tipos de usuários que digitam a URL no sistema de busca ao invés de no campo de endereços do browser: aqueles suficientemente inexperientes para não compreender a diferença entre os dois e aqueles que são tão experientes que estão habituados a usar os buscadores como um portal para a internet. (...) Não importa se este comportamento é motivado por ignorância ou destreza, o resultado final é o mesmo: o buscador é o ponto focal da experiência online para todos os tipos de usuários da internet. (Ken Cassar in Nielsen/Netratings, 2006)

Outros dois modos de encontrar os sites, digitando a URL diretamente na barra de endereços e atravessando os links de um site para outro, são praticados em escala bem mais modesta. Para a maioria dos usuários, tudo se passa como se a web se restringisse ao conteúdo dos bancos de dados dos grandes buscadores. Embora estes tenham dimensões expressivas, cobrem apenas uma parcela da WWW. Mesmo desconsiderando o conteúdo privado, estimado entre quinhentas (Cohen, 2006) a duas mil (Bergman, 2001) vezes maior que a Web indexável, Gulli e Signorini calcularam que em 2005 os bancos de dados dos principais buscadores não cobriam mais que 76,2% da web (*Google.* O alcance do *Yahoo*! seria 69,3%, do *MSN* 61,9% e do *Ask* 57,6%) (Gulli e Signorini, 2005, p. 2). As taxas de sobreposição entre os bancos de dados dos quatro sistemas mais populares é também significativa (Figura 4):

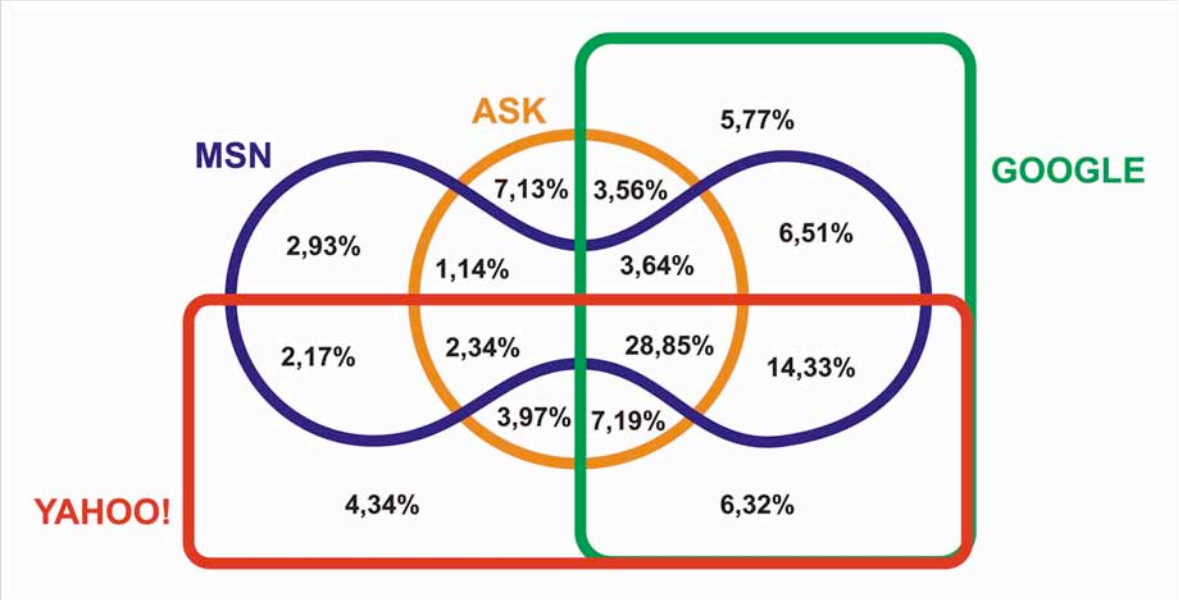

Figura 4: Representação gráfica das porcentagens da web indexável nos bancos de dados dos maiores buscadores. REPRODUZIDO DE GULLI E SIGNORINI, 2005, p. 2.

Mesmo indexados, muitos sites não chegam jamais a constar entre os resultados das buscas. Uma das razões para isso é a restrição do intervalo que as ferramentas efetivamente dedicam às consultas: para evitar que o usuário desista da busca e vá realizá-la em outro sistema, após um certo tempo de acesso a busca é interrompida, independente da cobertura da consulta (o Google inclusive indica tempo dedicado à pesquisa junto ao número de resultados encontrados) Essa restrição de tempo perde importância quando se verifica que, apesar de anunciar quantidades enormes de resultados para os usuários, os grandes buscadores de fato não disponibilizam mais que – no máximo – os mil primeiros. Além disso, apesar dos algoritmos de des-clusterização, mais de uma página de um mesmo site por vezes figura entre os resultados apresentados (Fragoso, 2006).

A maioria dos usuários não chega a perceber o limite de páginas efetivamente exibidas pelos buscadores, pois concentra sua atenção nos primeiros classificados. Verificações empíricas indicam que não mais de 10% dos usuários prosseguem para além da 3ª página de resultados, sendo que 62% tendem a selecionar um resultado que figura na primeira página (iProspect, 2006). O resultado é uma acentuadíssima canalização de tráfego em alguns poucos endereços, convergindo para os que se classificam melhor junto às principais ferramentas de busca.

Finalmente, é preciso dizer que os resultados das buscas podem ser bastante inconsistentes: buscas com os mesmos parâmetros realizadas em ocasiões diferentes, muitas vezes apresentam resultados diferentes, sobretudo no *Google* (Fragoso, 2006). Os usuários, no entanto,

> [s]entem-se no controle das buscas; quase todos expressam confiança em suas habilidades para utilizar os buscadores. Estão felizes com os resultados que encontram; mais uma vez, quase todos dizem ser bem sucedidos e encontrar o que estavam procurando. Além disso, os usuários confiam muito nos sistemas de busca: a grande maioria declarou que os buscadores são fontes de informação justas e neutras (Fallows, 2006, p. 2)

Evidentemente os sistemas de busca não podem deixar de proceder seleções e estabelecer hierarquias; afinal, esta é sua primeira finalidade. É verdade que sua operação não representa um re-aprisionamento do pólo da emissão e portanto não compromete a liberdade de expressão na WWW. É preciso estar alerta, entretanto, para o fato de que os buscadores funcionam como verdadeiros *gatekeepers* digitais - com o agravante de que operam conforme critérios cuidadosamente mantidos em sigilo e com objetivos estritamente comerciais. É amplamente sabido que as ferramentas de busca tendem a indexar mais sites dos EUA que dos demais países (Thellwall e Vaughan, 2004), misturam resultados pagos e orgânicos, seus algoritmos podem ser manipulados interna ou externamente, etc. Apesar disso, os usuários confiam candidamente nos buscadores, garantindo a condição final para que a Web reverta para um modelo de distribuição verticalizado, cujo funcionamento tende a ser ainda mais centralizado e tendencioso que o dos meios massivos de comunicação.

Referências

ANDERSON, C. The Long Tail, **Wired Magazine**, Issue 12.10, Outubro de 2004. Disponível online em http://www.wired.com/wired/archive/12.10/tail.html [14/01/2007]

ARASU, A. et al, Searching the Web. **ACM Transactions on Internet Technology**, Vol. 1, No. 1, Agosto de 2001, p. 2-43. Disponível a partir de http://.portal.acm.org [acesso restrito] [25/12/2006]

BERGMAN, M. K. The Deep Web: Surfacing Hidden Value. **The Journal of Electronic Publishing** Volume 7, Issue 1, Agosto de 2001. Disponível online em http://www.press.umich.edu/jep/07-01/bergman.html [25/12/2006]

BRIN, S. e L. PAGE, The Anatomy of a Large-Scale Hypertextual Web Search Engine, **Seventh International Conference on World Wide Web**, Brisbane, Australia, 1998. Disponível online em

BRUCE CLAY, INC, **Internet Business Consultants**, 2007**.** Disponível online em http://www.bruceclay.com [10/01/2007]

COHEN, L., **Internet Tutorials.** *University at Albany, SUNY.* Disponível online em http://www.internettutorials.net/ [25/12/2006]

DEUTSCH, P., A. EMTAGE e B. HEELAN, **archie - An Electronic Directory Service for the Internet**, 1990. Disponível online em http://tecfa.unige.ch/pub/documentation/Internet-Resources/short-guides/whatis.archie [04/01/2006]

FALLOWS, D. Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. **Pew Internet & American Life Project**, 2005 Disponível online em http://www.pewinternet.org/ [10/01/2007]

FRAGOSO, S., Sampling the Web: discussing strategies for the selection of Brazilian websites for quanti-qualitative analysis. In: M. Consalvo e C. Haythornthwaite (orgs.). **AoIR Internet Research Annual**. New York: Peter Lang, 2006, v. 4, p. 195-208

FUTURE NOW INC., **What converts search engine traffic: understanding audience, vehicle, message and perspective to optimize your ROI**. 2003. Disponível em http://jobfunctions.bnet.com/whitepaper.aspx?&tags=E-business%2FE-commerce&docid=161804 [acesso restrito]

GRAY, M., **Measuring the Growth of the Web : June 1993 to June 1995**. MIT Report , 1995. Disponível online em http://www.mit.edu/people/mkgray/growth/ [14/01/2007]

GULLI, A e A. SIGNORINI, The Indexable Web is more than 11.5 billion pages. **International Conference on the WWW 2005**, 10 a 14 de Maio, 2005, Chiba, Japão. Disponível online a partir de http://www.cs.uiowa.edu/~asignori/web-size/size-indexable-web.pdf [25/12/2006]

IPROSPECT, Inc., **Search Engine User Behavior Study**, Abril de 2006. Disponível online em http://www.iprospect.com [26/12/2006]

KAHLE, B. An Information System for Corporate Users: Wide Area Information Servers, **WAIS Corporate Paper version 3**. 8 de Abril de 1991.Versão para MS-Word disponível em ftp://think.com in /pub/wais/wais-overview-docs.sit.hqx. [04/01/2007]

KOSTER, M., ALIWEB - Archie-Like Indexing in the Web**, First International Conference on the World-Wide Web**, Genebra, Suíça, 1994. Disponível online em http://www.informatik.uni-stuttgart.de/menschen/sommersn_public/aliweb-paper.html [04/01/2007]

LAWRENCE, S. e L. GILES, Accessibility and Distribution of Information on the Web, **Nature**, Vol. 400, pp. 107-109, 1999. Versão reduzida disponível online em 2003 em http://wwwmetrics.com [02/01/2007]

MANEY, K. Google: The next Microsoft? Noooo! Cyberspeak, **USA Today**, 31 de Agosto de 2005. Disponível online em http://www.usatoday.com/tech/columnist/kevinmaney/2005-08-30-google-microsoft_x.htm [14/01/2007]

MAULDIN, M.L., Lycos: Design choices in an Internet search service. **IEEE Expert**, Jan-Fe*v* 1997, p. 8-11. *Disponível em* IEEE Expert Online, http://www.fuzine.com/lti/pub/ieee97.html [10/01/2007]

MILLS, E. Google, Yahoo, Microsoft adopt same Web index tool. **CNET News.com**, 15 de Novembro de 2006. Disponível online em http://www.cnet.com/?tag=hdrgif [02/01/2007]

MOHNEY, D: Is Google the next Microsoft? **The Inquirer,** 1 de Setembro de 2003. Disponível online em http://www.theinquirer.net/default.aspx?article=11305 [14/01/2007]

MURRAY, B., **Sizing the Internet: a Cyveillance White Paper**, 2000. Disponível online em http://www.cyveillance.com [02/01/2007]

MURRAY, R. Search Wars Salvo: Microsoft Launches Live Search, **Search Insider Media Post**, 6 de Outubro de 2006. Disponível online em http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticleHomePage&art_aid=49199 [14/01/2007]

NIELSEN/NETRATINGS, 2006. Top Search Terms Reveal Web Users Rely On Search Engines To Navigate Their Way To Common Web Sites. **Nielsen/Netratings Press Release, 18 de Janeiro de 2006**. *New York,*. Disponível online em http://www.nielsen-netratings.com [14/01/2007]

PINKERTON, B, .**WebCrawler Timeline**. Disponível online em http://thinkpink.com/bp/WebCrawler/History.html [14/01/2007]

PINKERTON, B. Finding What People Want: Experiences with the WebCrawler, **Second International WWW Conference**, Chicago, USA, 1994. Disponível online em http://thinkpink.com/bp/WebCrawler/WWW94.html [04/01/2007]

SALIENT MARKETING, ***History of Search Engines***. S.d. Disponível online a partir de http://www.salientmarketing.com/seo-resources/search-engine-history.html [03/01/2006]

SELBERG, E. e O. Etzioni, *Multi-Service Search and Comparison Using the MetaCrawler* - **Fourth International World Wide Web Conference**, Boston, USA, 1995. Disponível online em http://www.w3.org/Conferences/WWW4/Papers/169/ [05/12/2006]

SONNENREICH, W.e T. MACINTA, **A History of Search Engines,** Wiley Computer Publishing, 1998. Disponível online em http://www.wiley.com/legacy/compbooks/sonnenreich/history.html [26/12/2006]

SULLIVAN, D. Microsoft's MSN Search To Build Crawler-Based Search Engine, **SearchEngineWatch**, 1 de Julho de 2003. Disponível online em http://searchenginewatch.com/showPage.html?page=2230291 [14/01/2007]

SULLIVAN, D. MSN Search Officially Switches To Its Own Technology **SearchEngineWatch SearchDay**, 1 de Fevereiro de 2005. Disponível online em http://searchenginewatch.com/searchday/article.php/3466721 [14/01/2007]

SULLIVAN, D. Nielsen NetRatings Search Engine Ratings. **SearchEngineWatch Report**, 22 de Agosto de 2006. Disponível online em http://searchenginewatch.com/reports/article.php/2156451 [03/01/2007

THIES, D. The Search Engine Marketing Kit, 2005. Disponível online em http://www.sitepoint.com/books/sem1/ [04/01/2007]

YAHOO! MEDIA RELATIONS, 2005, **The History of Yahoo! - How It All Started.** Disponível online em http://docs.yahoo.com/info/pr/index.html [4 jan 2007]