Estudos da Língua(gem)
Número Especial

O Controle da palavra do outro nos dispositivos de busca

(The control of the other's speech on search engines)

Marco Antonio Gutierrez*
Universidade UnIgranRIO

RESUMO
Se "todo enunciado concreto é um elo na cadeia da comunicação discursiva de um determinado campo" (BAKHTIN, 2003, p. 296), meu enunciado de hoje é também uma resposta a um conjunto de enunciados de outrem, com quem dialoguei ontem. Ao controlar as vozes com quem sou capaz de dialogar na Internet, os dispositivos de busca são capazes de controlar também o meu próprio discurso. Partindo dessa premissa, o presente artigo discute o critério de relevância adotado pelos dispositivos para selecionar o resultado das buscas, em particular o critério do Google, responsável por metade das buscas realizadas na Internet.

PALAVRAS-CHAVE
Internet. Dispositivos de busca. Controle do discurso.

ABSTRACT
"Any concrete statement is a link in the communication chain of a particular human activity field" (BAKHTIN, 2003, p. 296). If it is true, my discourse is now a response to another discourse with which I interacted yesterday. If search

*Sobre o autor, ver página 116.

engines can control the voices of whom I am able to talk on the Internet, they are also able to control my own discourse. From this point of view, this article discusses the criteria of relevance adopted by search engines to select theirs outputs, especially Google's criteria, responsible for half of the Internet searches.

KEYWORDS
Internet. Search engines. Discourse control.

1 Introdução

Neste artigo, partimos da premissa que é a arquitetura tecnológica da Internet, ao constituir uma rede ex-cêntrica, que a torna um espaço de (quase) irrestrita liberdade de expressão, de tal modo que Lévy, em seu prefácio à obra de Lemos, se refere a ela como "um espaço de comunicação propriamente surrealista, do qual 'nada é excluído', nem o bem, nem o mal, nem suas múltiplas definições, nem a discussão que tende a separá-los sem jamais conseguir" (LEMOS, 2002, p. 14). Deve-se notar apenas que esta não é uma característica acidental, mas o *objetivo primário* da rede. Não é necessário nos alongarmos aqui na história da rede (ver, por exemplo, CASTELLS, 2003); basta-nos assinalar que no seu berço está também um projeto militar destinado a encontrar alternativas de comunicação para a eventual destruição dos centros de comando dos Estados Unidos por um hipotético ataque nuclear. Ora, a alternativa encontrada foi uma rede cujo centro de comando não pudesse ser destruído simplesmente porque inexistia... Em outras palavras, a rede nasceu com características tais que visavam impedir o seu controle por potências hostis. Como decorrência inevitável desse objetivo, ela se tornou avessa a seu controle por *qualquer* entidade.

Essa característica foi reforçada pela sua evolução posterior, levada a cabo por uma comunidade acadêmica fortemente influenciada pelos movimentos de base típicos da sua época, fazendo com que "a cultura da liberdade individual que floresceu nos *campi* universitários nas décadas de 1960 e 1970 [usassem] a interconexão dos computadores para seus próprios fins" (CASTELLS, 2003, p. 25). Por conta disso, o mecanismo de agregação de novas tecnologias e padrões à rede sempre foi fortemente participativo, aberto à discussão e à intervenção de *toda* a comunidade técnica interessada,

mecanismo utilizado até hoje. Visto em retrospecto, era, portanto, inevitável que essa evolução fosse acompanhada pela inclusão de tecnologias que facilitassem a *participação* e a *interação* ativa dos usuários da rede. Ao contrário da maioria das novas tecnologias computacionais, originadas nas pesquisas da indústria e dos governos, as tecnologias da Internet se originaram a partir da própria comunidade de usuários, colaborando em projetos tecnológicos. Todos os grandes padrões e tecnologias utilizados na Internet nasceram desse modo: a *web*, o *chat*, o *e-mail*, etc. – a lista abrange *todos* os recursos de *software* utilizados na rede. A indústria simplesmente seguiu a comunidade de usuários, bem depois das tecnologias serem desenvolvidas.

Essas tecnologias e padrões se originaram de projetos colaborativos de iniciativa espontânea das comunidades tecnológicas. Era natural que dessa interação surgisse "um conjunto de valores e crenças", isto é, uma cultura. E é essa cultura que Castells e Lévy chamam de *cultura hacker*, um subconjunto da cibercultura, que "emergiu das redes de programadores de computador que interagiam on-line em torno de sua colaboração em projetos autonomamente definidos de programação criativa" (CASTELLS, 2003, p. 38). Essa cultura autônoma permanece ativa e forte até hoje, mesmo com a "invasão" pela indústria dos mecanismos colaborativos clássicos da Internet (como o IETF – *Internet Engeneering Task Force*, responsável por todos os padrões da Internet). Pensemos na – imensa – comunidade de *software* livre que continua a florescer.

Trata-se de um movimento de base na Internet iniciado por Richard Stallman, pesquisador do MIT, na segunda metade da década de 1980. Esse pesquisador e agora as centenas de milhares de participantes dos mais de 180 mil projetos colaborativos de *software* disponíveis (veja-se *site* Source Forge – http://www.sourceforge.net) concebem o *software* como conhecimento humano materializado e, como tal, não deveria ser patenteado e comercializado, mas *compartilhado* por todos. Ele se propunha a produzir *software* em colaboração com quaisquer voluntários, compartilhando com eles seus conhecimentos e divulgando-os livremente para todos os usuários interessados. Assim, fundou a *Free Software Foundation*, mais conhecida como Projeto GNU, cujo lema principal é "*Free as in Freedom*". Para essas iniciativas – tão bem sucedidas que a própria indústria aderiu – o "*software* livre se refere à liberdade dos usuários executarem, copiarem, distribuírem, estudarem, modificarem e melhorarem

o *software*" (THE FREE, 2006). O sucesso é tal que mais de 80% do tráfego Internet se baseia em aplicativos desenvolvidos sob o regime de *software* livre, pela própria comunidade especializada de usuários.

Portanto, colaboração e interação na busca e no exercício da liberdade – tais são as palavras-chave da cultura *hacker*, expressões que acabam por se materializar nos produtos desenvolvidos por essa comunidade. É a encarnação desses princípios no *software* produzido por essa tribo que permite que eles possam migrar para *toda* a comunidade, simplesmente através do seu exercício. Talvez o melhor exemplo da migração desses princípios para além da tribo tecnológica seja a Wikipedia (http://www.wikipedia.org). Trata-se de uma "enciclopédia livre", um projeto colaborativo para o registro em regime de compartilhamento e liberdade dos conhecimentos disponíveis, contendo hoje mais de 3 milhões de verbetes em 10 idiomas, incluindo o Português. Naturalmente, o *site* e os aplicativos que suportam a iniciativa foram produzidos em regime de *software* livre; porém, o que nos importa é o fato de essa enciclopédia ser produzida por *qualquer* usuário da rede que queira participar voluntariamente. Os colaboradores simplesmente acessam funções públicas do *site* e editam os verbetes existentes, produzem artigos novos, criticam o material disponível – enfim, interagem colaborativamente para registrar o conhecimento disponível na comunidade. A participação é inteiramente livre e a avaliação dos produtos é igualmente livre, baseada no escrutínio dos próprios usuários.

É provável que se registre ali muita tolice; aliás, os próprios mantenedores são os primeiros a reconhecê-lo e convidam os usuários a formularem suas críticas e proporem correções. Não é certo, porém, que a iniciativa privada equivalente mais consagrada – a Enciclopedia Britannica – esteja totalmente livre de tolices; afinal, a variedade e a especialidade dos conhecimentos ali registrados por pessoal remunerado (e, portanto, necessariamente limitado tanto em abrangência quanto em qualidade) devem ser fortes dificultadores para assegurar a precisão do material. Aliás, qualquer especialista em qualquer assunto pode contribuir com a *Wikipedia* (e não poucos o fazem), o que não é verdadeiro para a Enciclopedia Britannica. Do ponto de vista do conhecimento ali registrado, porém, é mais importante o fato de se tratar do *conhecimento da comunidade* ali comunitariamente partilhado. Mais do que a "sociedade do conhecimento" o que a *Wikipedia* anuncia é a

comunidade do conhecimento.

Queremos assinalar com estes exemplos o fato de a arquitetura tecnológica da rede, por suas características históricas, ter fornecido a infra-estrutura material necessária para que a atividade humana ali realizada se baseasse fortemente na interação e na colaboração, livremente definidas e adotadas entre os agentes. Interação e colaboração em massa – creio serem estes os subprodutos mais relevantes da arquitetura tecnológica da Internet. E é este subproduto que torna possível à rede ser um palco para o exercício da liberdade (pelo menos por seus usuários). E é esta infra-estrutura material que permite que a atividade humana na rede acabe por produzir um conjunto de valores e crenças específico – a cibercultura. É certo que, tomados isoladamente, esses valores e crenças estão presentes em outras "culturas"; afinal, a Internet não é uma nova sociedade, mas uma extensão das sociedades humanas atuais. No entanto, é a velocidade com que tais valores se disseminam, a liberdade com que podem ser exercidos em conjunto, a facilidade com que podem ser partilhados, a impossibilidade de se estabelecerem fronteiras políticas e nacionais que tornam o ciberespaço um novo campo de atividade humana, capaz de produzir uma cultura própria.

2 Proibir, vigiar, punir

A impossibilidade de controle ligada à própria arquitetura tecnológica em torno da qual a rede foi fundada pode ser facilmente ilustrada pelos fracassos seguidos do governo chinês em restringir a navegação dos seus cidadãos. Para isso, aquele governo elaborou diversas estratégias, que têm sido objeto de reportagens na imprensa ocidental. As mais relevantes (com maior – embora ínfima – chance de sucesso) envolvem a tentativa de criar centros (parciais) na rede. Se não pode controlar que computadores estão conectados à rede, um país autocrático pode, no entanto, controlar (e restringir) as conexões da sua rede nacional à rede internacional. Na maioria dos países, qualquer empresa ou instituição com recursos suficientes pode estabelecer seus próprios canais de acesso à rede internacional (*backbones*), utilizando-os privadamente ou liberando-os (mediante remuneração) ao público através dos provedores de acesso. São esses os canais que ligam as

redes nacionais à rede internacional. É o caso, por exemplo, do Brasil, que conta com vários canais de acesso. Na China, no entanto, apenas o estado tem esse direito. Como esses canais de acesso são necessariamente limitados (pelo seu alto custo), é possível estabelecer neles mecanismos de controle. É o que fez (ou tentou fazer) a criatividade governamental chinesa.

O controle envolveu a utilização não convencional de um número limitado de *firewalls*, introduzindo artificialmente um centro numa rede ex-cêntrica. Os *firewalls* são dispositivos computacionais projetados para analisar (e eventualmente bloquear) o tráfego de rede. Tais dispositivos foram construídos para localizar tráfego que possa representar ameaças aos usuários (ataques de *crackers*, invasões à privacidade, vírus, etc.) com base em padrões já razoavelmente conhecidos e bloqueá-lo. No entanto, como qualquer arma, essa também aponta para dois lados. Como a maior parte do tráfego na Internet é legível por seres humanos, a maior parte da informação trafegando em língua natural, é possível utilizar os *firewalls* para localizar requisições a sites específicos e/ou contendo palavras "suspeitas", bloqueando-as. Que palavras podem ser essas deixo à imaginação e às convicções do leitor. O preço a pagar, no entanto, é elevado: o processamento computacional envolvido torna *todo* o tráfego de rede proibitivamente lento, não apenas aquele considerado suspeito. Isso prejudica mesmo o uso "legítimo" da rede, incluindo o próprio uso governamental. A solução encontrada envolveu a restauração da ex-centricidade da rede: o *firewall* apenas analisaria o tráfego, que seria bloqueado por outros recursos computacionais. Os detalhes técnicos não nos importam aqui; importa o fato de que o controle, ao envolver a restauração da ex-centricidade da rede, poderia ser vencido. Assim é que pesquisadores (ocidentais, naturalmente) demonstraram rapidamente *como* o controle poderia ser vencido, publicando seus resultados na Internet e condenando ao fracasso a iniciativa chinesa.

Uma variante igualmente insidiosa desse controle parece estar sendo patrocinada pelos órgãos de informação dos Estados Unidos. Recentemente, um ex-funcionário de uma das grandes companhias de telecomunicações responsável pela manutenção de um dos *backbones* americanos denunciou na imprensa que o tráfego por aquele nó da rede estaria sendo monitorado, armazenado e eventualmente repassado à National Security Agency. O

objetivo aqui é menos bloquear e mais vigiar (e eventualmente punir) tráfego considerado de risco para a segurança nacional daquele país. Esse controle *a posteriori* da informação é impossível de evitar, caso seja tentado, simplesmente porque, pelas características tecnológicas da rede, não temos como determinar por que nossa informação eventualmente trafegará até chegar a seu destino. Além disso, salvo por denúncia dos envolvidos no processo, não temos como saber se estamos sendo monitorados. Não queremos insistir nesse ponto, porém. Basta-nos assinalar que, na Internet, onde quer que se instale algum "centro" parcial para a informação que nela trafega, aí o controle pode ser estabelecido.

Sabemos que a arquitetura tecnológica da Internet impede a existência de um centro único, fazendo com que qualquer nó da rede possa ser tanto receptor quanto fonte de informação. Essa característica torna possível a qualquer usuário da rede tornar-se também ele mesmo um centro de informação. Como os custos globais da rede são compartilhados por todos os – milhões de – usuários, qualquer pessoa digitalmente incluída pode publicar e manter seu próprio *site* a um custo desprezível, normalmente agregando aos custos de acesso menos de 20% em países como o Brasil. Isso torna possível que todo o conhecimento registrado em meio magnético no planeta possa vir a estar disponível em algum lugar da rede. Na medida em que a Internet se globaliza, esta possibilidade está se tornando cada vez menos remota. No entanto, essa mesma arquitetura tecnológica cria um outro problema: *como* encontrar o conhecimento? Ele está acessível, sim – mas somente se conhecermos sua localização! E como a rede foi projetada sem centros, esses "endereços" não estão disponíveis num único repositório. Na Internet, toda informação pode estar *disponível*, mas não é facilmente *acessível*. A solução encontrada pela própria comunidade tecnológica foram os dispositivos de busca.

Os *search engines* são dispositivos computacionais projetados para localizar e indexar qualquer documento publicado na rede, através de autômatos que varrem sistematicamente todos os seus nós. Havendo poder computacional bastante e dado um tempo suficientemente longo é pelo menos teoricamente possível a esses dispositivos catalogar *todos* os documentos jamais publicados na Internet. Assim, os *search engines* podem funcionar como imensos "catálogos telefônicos" da Internet, fornecendo – pelo

menos teoricamente – a localização de qualquer documento disponível na rede. Essa necessidade dos usuários é de tal ordem que o *site* Search Engine Watch estima que sejam realizadas na rede 213 milhões de buscas (ver SULLIVAN, 2004) por dia em todo o planeta nas poucas centenas de dispositivos existentes! Como são relativamente poucos, esses dispositivos funcionam como "centros de informação" numa rede sem centro – e, portanto, podem ser controlados de algum modo. Se tivermos o "controle" desses dispositivos, podemos, por exemplo, fornecer, como resultado de uma busca, determinados *sites* de interesse nosso e não quaisquer outros, estes se tornando inacessíveis na prática. Podemos também escolher *não* responder a determinadas buscas.

Esse controle, uma vez mais, ainda que possível, é limitado na prática e novamente pela arquitetura tecnológica da rede: afinal, qualquer um com relativamente poucos recursos pode instalar o seu próprio dispositivo de busca e funcionar como um centro de informação. Na verdade, existem centenas de dispositivos funcionais já catalogados. Na prática, porém, os dispositivos realmente utilizados pelo público são muito poucos. A Nielsen//Net Ratings, em pesquisa permanentemente, disponível no Search Engine Watch, estima que apenas 3 *sites* (o Google, o Yahoo e o MSN) são responsáveis por 81,1% de todas as buscas realizadas na Internet, distribuídos da seguinte forma: 49.2%, 23.8% e 9.6%, respectivamente. Isso torna viável o controle – e foi dessa possibilidade que se aproveitou o governo chinês na sua eterna ânsia de controle.

Para permitir que as empresas responsáveis por aqueles *sites* realizassem negócios na China, o governo daquele país exigiu (e foi atendido) que seus dispositivos não respondessem a buscas realizadas utilizando-se de expressões consideradas "suspeitas", nem que retornassem determinados *sites* independentemente do tipo de busca realizada. Esse controle é pelo menos temporariamente eficaz, dada a barreira da língua. Além disso, mesmo para aqueles poucos chineses com conhecimento de inglês as versões americanas dos dispositivos não são acessíveis, já que a estrutura centralizada da rede chinesa permite que eles sejam facilmente bloqueados. Embora qualquer cidadão chinês possa acessar qualquer *site* em qualquer lugar do planeta, dada as características técnicas da rede, e maioria não será capaz de fazê-lo simplesmente porque não sabe onde estão localizados.

Não nos alongaremos demais no problema. Queremos apenas assinalar que os dispositivos de busca, por se tratarem de poucos centros de informação

indispensáveis a uma rede ex-cêntrica, representam (por enquanto) um flanco acessível ao controle dessa caixa de Pandora chamada Internet.

3 O controle das profecias do oráculo

O controle exercido nos dispositivos de busca é um mecanismo de controle do discurso que procura censurar a enunciação por meios indiretos, controlando a palavra do outro: ao restringir o acesso a documentos que incorporam temas "proibidos", o que se quer é restringir a vigência no discurso de múltiplas vozes, perigosas porque diferentes, substituindo-as pelo discurso *monótono*, obliterando-as para o território do não-dito (porque não-lido). É essa uniformidade de um discurso plano o que se quer alcançar com a censura. No entanto, os mecanismos que abordamos sucintamente sempre se revelaram de curto alcance a longo prazo, sendo, portanto, ineficazes. Muito mais eficaz que a censura é a auto-censura, sobretudo quando assumida como expressão da liberdade para a palavra do outro...

A partir de agora abordaremos o funcionamento desses mecanismos de auto-censura incorporados aos dispositivos de busca, assumindo que eles têm como efeito não apenas o controle do próprio discurso, como também o do discurso do outro – o usuário que busca – ao delimitar (e, implicitamente, *limitar*) o universo de discursos a que temos acesso no ciberespaço. O universo alvo dessas buscas é muito grande: em novembro de 2004, o Google anunciava que já havia cerca de 8 bilhões de documentos indexados em suas bases de dados (SULLIVAN, 2004). Mesmo supondo que os índices nas bases de dados de todos os demais dispositivos estejam contidos neste conjunto, é fácil notar que sem o auxílio dos *search engines* seria impraticável o acesso a qualquer parcela de documentos disponíveis na *web*. Portanto, do ponto de vista prático, o acesso à informação na Internet é controlado por esses três dispositivos. O que sustentamos é que tais dispositivos controlam não apenas o *acesso* à informação, mas também, em grande medida, sua própria *produção*. Isso decorre da concepção de linguagem com a qual privilegiadamente trabalhamos – a de Bakhtin.

Na tradição bakhtiniana todo discurso é orientado à audiência, da conversa cotidiana ao poema lírico: "o traço essencial (constitutivo) do enunciado é o seu *direcionamento* a alguém, o seu *endereçamento*" (BAKHTIN, 2003, p. 301). Isto significa que determinadas suposições sobre o interlocutor

afetam as escolhas disponíveis para o locutor, determinando seu próprio enunciado. Este seria um traço de toda interação comunicativa. Ora, se supusermos que a natureza artificial da inteligência dos dispositivos de busca não afeta a interação comunicativa com eles, que teriam as mesmas características básicas de interação do mesmo tipo entre seres humanos, estaremos admitindo também que tanto o enunciado que elaboramos numa busca quanto a resposta dada pelo dispositivo são afetados por certas suposições sobre as partes envolvidas. Na resposta a uma indagação, o dispositivo humano elabora uma hipótese sobre o que o locutor quer de fato saber e responde com base nela; na elaboração dessa hipótese é levado em consideração tanto o conhecimento de mundo disponível quanto uma concepção prévia do próprio enunciador da pergunta (a audiência da resposta). Assim, um dispositivo computacional que simule em alguma medida o interlocutor humano, ao elaborar sua resposta, deve levar em consideração uma concepção prévia do que o enunciador da pergunta deseja saber.

Essa hipótese tem pelo menos uma decorrência que pode ser facilmente testada por experimentos: diferentes dispositivos de busca, se programados com base em diferentes concepções da audiência, devem responder com resultados distintos às mesmas buscas, independente dos "conhecimentos" disponíveis (a base de dados consultada). Isso significa que a resposta fornecida pelo dispositivo não é "neutra": ele não responde com o que queremos saber, mas com aquilo que ele mesmo entende que queremos saber.

Se "todo enunciado concreto é um elo na cadeia da comunicação discursiva de um determinado campo" (BAKHTIN, 2003, p. 296), meu enunciado de hoje é também uma resposta a um conjunto de enunciados de outrem, com quem dialoguei ontem. Ora, ao controlar em grande medida as vozes com quem sou capaz de dialogar na Internet, os dispositivos de busca, a longo prazo, são capazes de controlar também, em alguma medida, o meu próprio discurso. Se o ciberespaço fosse o único espaço social de interação discursiva disponível, os resultados das buscas nos *search engines* abarcariam praticamente *todos* os enunciados com os quais sou capaz de dialogar... Cremos que essa é uma razão muito boa para tornar essencial o estudo dos dispositivos de busca do ponto de vista discursivo.

4 Diferentes concepções dos dispositivos de busca

De uma forma geral, todos os dispositivos de busca utilizam a mesma estratégia para construir suas bases de dados (seu "conhecimento de mundo"): eles incorporam um robô que varre sistematicamente os endereços de rede da Internet em busca de servidores que disponibilizem documentos mediante requisição, tais como servidores *web*, *newsgroups*, servidores de arquivos, etc. Esses robôs compõem o chamado módulo de *crawling* do dispositivo. Isso significa que dispositivos de busca com a mesma capacidade computacional à disposição dos seus robôs, dado um período de tempo suficientemente longo, tenderão a dispor das mesmas bases de dados para suas consultas. Uma implicação disso é que é de se esperar que o resultado das buscas em dispositivos equivalentes não seja afetado pela base de dados disponível para suas consultas; esse resultado deve ser somente uma função dos critérios utilizados para sua classificação (indexação) e ordenação – e são esses critérios que darão conta da concepção de audiência programada naqueles dispositivos.

Qualquer um que já tenha utilizado um *search engine* saberá que o crítico nesses dispositivos não é o acervo, mas o critério de ordenação e apresentação dos resultados, isto é, a classificação da *relevância* das saídas. Uma simples busca pode envolver milhares de documentos como saída – e obviamente ninguém é capaz de inspecionar todos. A maioria das buscas é concluída após uma fração desprezível das suas saídas ter sido inspecionada, o que significa que a maioria delas deve se resumir às primeiras páginas do resultado. Em pesquisa recente, a iProspect (IPROSPECT, 2006) concluiu que 88% dos usuários abandonam a busca após consultar apenas a terceira página de resultados e 41% deles a abandonam já após a primeira página!

Isso implica que o conteúdo do acervo dos dispositivos é irrelevante para o resultado concreto da interação de busca não apenas porque diferentes dispositivos deverão ter o mesmo acervo à disposição ao longo do tempo, mas principalmente porque a quantidade de documentos coerente com a expressão buscada é sempre muito maior do que a nossa capacidade de inspecioná-los. Desse modo, o fator isolado mais importante na busca deve ser o critério de classificação do acervo e de ordenação das saídas para apresentação. É

aqui que os projetistas de sistemas devem fazer algumas suposições críticas sobre o que os usuários querem obter quando interrogam o dispositivo em suas buscas no ciberespaço.

A idéia na origem dos dispositivos de busca é coletar, armazenar e eventualmente localizar o que se convenciona chamar "documentos relevantes", isto é, aqueles documentos disponíveis na Internet nos quais o usuário da busca está interessado. Essa é uma tarefa impraticável, ainda que teoricamente possível: a quantidade de documentos disponíveis na Internet é muito grande e cresce continuamente. Além disso, uma fração não desprezível desses documentos é continuamente removida e atualizada. Esses fatores aliados à limitada capacidade computacional dos dispositivos fazem com que o conjunto de respostas de uma busca deva ser diferente em algum grau daquele conjunto (teórico) de documentos relevantes.

Seja R o conjunto de documentos relevantes no universo da Internet e |R| o número de elementos em R; seja D o conjunto de documentos obtidos por um dispositivo de busca e |D| o número de elementos em D. O alvo do processo de busca deve ser, então, obter o conjunto intersecção $R_d = R \cap D$. Chama-se *cobertura* a fração C de documentos relevantes recuperados na busca, isto é, a razão entre $|R_d|$ e |R|; chama-se *precisão* a fração P de documentos recuperados que é relevante, isto é, a razão entre $|R_d|$ e |D|. Como não é praticável obter C = 1, o objetivo dos projetistas desses dispositivos é aumentar o valor da precisão P para valores tão próximos da unidade quanto possível. Duas estratégias são aqui utilizadas. A primeira delas é de natureza estritamente tecnológica: a indexação dos documentos deve ser tal que assegure ao máximo possível que os diferentes assuntos abordados num documento que eventualmente possam ser relevantes numa busca estejam acessíveis durante o processo. Como resultado, os dispositivos refinam continuamente sua capacidade de indexação, não apenas de modo a abarcar *todo* o conteúdo de um documento, independente do seu formato, como também de modo a catalogar elementos do "contexto" do documento, incluindo coisas como o diretório virtual onde reside, o nome do documento, documentos vizinhos, etc. A maior parte das pesquisas nesse domínio acadêmico visa elevar e acelerar essa capacidade de indexação e a localização dos documentos catalogados.

Uma outra estratégia, porém, pode ser também adotada: trata-se

de elevar a capacidade do dispositivo de determinar *o que* o usuário está buscando. Essa estratégia permite que os dispositivos selecionem do acervo de indexação somente aquelas entradas *relevantes* para o usuário. Quanto maior a capacidade de o dispositivo interpretar essa *intenção* do usuário, mais preciso será o resultado de P. Mesmo isso pode ainda resultar num valor muito elevado para D, mesmo que a razão R_d seja ótima (igual à unidade). Desse modo, os dispositivos devem igualmente classificar o *grau* de relevância de um documento e ordenar sua saída de acordo.

Esta estratégia é, certamente, mais significativa para uma pesquisa em Lingüística, já que é neste ponto que os projetistas dos dispositivos devem elaborar uma *hipótese interpretativa* sobre o *sentido* e não sobre o *significado* da pergunta elaborada pelo usuário no diálogo de busca, isto é, o seu sentido. Isso significa que o dispositivo não precisa resolver o problema do significado da expressão de busca, isto é, ele não precisa interpretar o que o usuário está tentando *dizer*. Ela precisa, na realidade, resolver o problema de determinar o que o usuário está tentando *saber* com aquela expressão de busca. É com base nessa hipótese interpretativa que os dispositivos avaliam as entradas encontradas nos seus índices e ordenam o resultado das buscas.

O critério clássico de relevância se baseia, grosso modo, no seguinte raciocínio. Suponhamos que o documento *a* qualquer aborde *principalmente* o tema *x* e *marginalmente* o tema *y*, enquanto que o documento *b* faça o inverso. Esse conhecimento deve ser expresso de algum modo no dispositivo, de tal modo que um usuário que esteja pesquisando *x* receba *a* antes de receber *b*, já que, para o assunto da busca, aquele documento deve ser (supostamente) *mais relevante* que este. No caso da busca *y* deve ocorrer precisamente o inverso.

O critério clássico para classificação e ordenação dos resultados das buscas parte, então, da premissa de que a incidência da expressão de busca num documento é indício do interesse do usuário. Assim, a taxa de incidência num documento e a exatidão da expressão são os critérios principais utilizados no ordenamento. Quanto mais ocorrências da expressão buscada (e quanto maior sua exatidão) forem encontradas num documento, maior será a probabilidade de ele ser ordenado nas primeiras posições apresentadas na saída. Esse critério elementar é refinado continuamente pelos dispositivos

dedicados à busca, incluindo elementos "contextuais" e heurísticas complexas para a tomada de decisão classificatória; no entanto, a idéia geral é (quase) sempre esta. Duas são as hipóteses (relacionadas) utilizadas para dar conta da "intenção" do usuário:

> i. Quando interrogamos um bibliotecário (o equivalente pré-ciberespaço do dispositivo de busca) a propósito de um livro, nós geralmente nos referimos ao "assunto" principal abordado no texto;
> ii. Há boa probabilidade de que o tema de um documento se exprima através da incidência do mesmo vocabulário no corpo do texto.

Por exemplo, este capítulo é dedicado aos dispositivos de busca; até este ponto, a palavra-tema se repetiu treze vezes, enquanto que os temas relacionados (ciberespaço e Internet) foram repetidos cinco e vinte e seis vezes, respectivamente. Naturalmente, a hipótese (ii) nem sempre é boa. Por exemplo, o tema de um livro que contenha somente receitas de cozinha pode ser expresso por vocábulos como "culinária", "gastronomia", "receitas", etc. e, ainda assim, não conter em todo o documento uma única ocorrência dessas expressões! Por conta disso, a medida que ganham experiência, os próprios usuários percebem que os dispositivos de busca respondem segundo a hipótese (ii) e se adaptam isso. Eles passam também a perguntar não por documentos relacionados ao "tema", mas por documentos que *contenham* uma determinada expressão de busca.

Um critério simples de classificar a relevância, elaborado com base nessa hipótese interpretativa, poderia ser, por exemplo, determinar a razão entre a expressão tema e a quantidade total de palavras no documento. No nosso caso (até o parágrafo anterior contamos 4.416 palavras), a razão da relevância dos temas catalogados – dispositivos de busca, ciberespaço e Internet – seria, respectivamente, 0.00294, 0.00113 e 0.005887, o que ainda não parece consistente com a estrutura temática deste fragmento de documento. Poderíamos, por exemplo, supor que existe boa probabilidade do tema do documento ser expresso na primeira linha (o título) e ponderar o cálculo de acordo; se arbitrarmos peso cinco para essa condição, o valor calculado de relevância naquela seção do documento da expressão "dispositivos de busca" saltaria para 0.00384, o que melhora consideravelmente o nosso "cálculo"

para o(s) tema(s) do artigo. Esse critério pode ser refinado continuamente através de novas hipóteses sobre como os produtores de texto exprimem do ponto de vista vocabular o assunto dos seus documentos, agregando características típicas da Internet. Por exemplo, é altamente provável que um *site* cujo nome de domínio seja http://www.searchengine watch.com se refira a dispositivos de busca! Isso pode se refletir como valor mais alto na ponderação das nossas entradas de índice.

Esse critério tem um problema: ele *não* reage como um bibliotecário humano culto e bem informado. Suponhamos que estamos utilizando um dispositivo de busca construído com base no critério descrito no parágrafo anterior para o cálculo da relevância dos documentos. Para uma busca pela expressão "ciberespaço" ele devolveria o presente texto antes do artigo *A emergência do ciberespaço e as mutações culturais* (ver http://empresa. portoweb.com.br/ pierrelevy/aemergen.html), de Pierre Lévy, cujo cálculo de relevância para essa expressão de busca teria sido de 0.00110. É de se presumir que um bibliotecário humano, culto e bem informado, não respondesse desse modo, mas devolvesse o texto de Lévy em primeiro lugar. Afinal, este bibliotecário hipotético provavelmente imaginaria que estamos em busca de idéias de pesquisadores importantes e seu conhecimento de mundo o forçaria a se perguntar: mas afinal, quem é mais importante, um obscuro pesquisador júnior de uma universidade brasileira ou o filósofo francês Pierre Lévy?

Foi raciocinando deste modo que alguns pesquisadores de Stanford (Larry Page, Serguei Brin, Rajeev Motwani e Terry Winograd – os dois primeiros os fundadores do Google) elaboraram há alguns anos um critério distinto, batizado como *PageRank*. A idéia central daqueles pesquisadores é buscar um meio de determinar a "importância" de um documento na *web* consistente com o raciocínio daquele bibliotecário hipotético. Para isso, eles partiram da premissa de que um dos critérios acadêmicos de avaliação dos pesquisadores, a contagem de citações recebidas, seria aplicável para o cálculo da "importância" de um documento na *web*, tirando proveito da própria estrutura do hipertexto, baseada em *links* entre páginas. Para eles, "páginas [*web*] para as quais muitos *links* apontam são, em geral, mais 'importantes' que páginas com poucos *links*" apontando para elas – do mesmo modo que "a simples contagem de citações tem sido utilizada para especular sobre os

futuros ganhadores do Prêmio Nobel" (PAGE et al, 1998, p.3). Pelo critério, a importância de uma página se propaga do mesmo modo que uma citação elogiosa a outro feita por um pesquisador muito importante. À época, esses pesquisadores determinaram, por exemplo, que a página principal do Yahoo era (então) o documento para o qual mais *links* apontavam. Pelo critério adotado, eles admitem então que "se uma página tem um *link* saindo da *home page* do Yahoo, ele pode ser apenas um *link*, mas trata-se de um muito importante", já que "esta página deveria receber um *ranking* mais alto que muitas páginas com mais *links* [apontando para elas], mas oriundos de lugares mais obscuros" (PAGE et al, 1998, p. 3).

Algumas implicações desse critério de relevância serão discutidas mais adiante. Por ora, basta assinalar que, embora não seja dito no artigo citado, parte-se da premissa de que o usuário está interessado primariamente nos *sites* mais populares que atendam ao critério de busca. Neste caso, as saídas passam a ser classificadas e ordenadas segundo a taxa de referência ao documento feita por outros documentos. Quanto mais *links* forem encontrados em *sites* diferentes apontando para um determinado documento, maior a probabilidade de este último aparecer nas primeiras saídas do resultado da busca. Trata-se aqui de determinar a audiência de um *site* a partir das referências feitas a eles por outros *sites* e devolver ao usuário aqueles documentos de maior audiência que atendam à expressão de busca utilizada.

5 Em busca da palavra do mesmo

É preciso afirmar inequivocamente que tudo indica que o critério de relevância do Google é o que melhor atende às expectativas dos usuários, não tanto pelas suas qualidades intrínsecas, mas pela palavra final dos usuários das buscas: a cada duas buscas realizadas no mundo, uma é feita neste dispositivo, a outra ficando por conta de pelo menos treze dispositivos, para contarmos somente os mais populares. Mesmo sem conhecer a natureza dessa diferença, os usuários devem ter escolhido o Google por força da consistência do produto das buscas realizadas no dispositivo com seus próprios interesses. Note-se que não é possível atribuir essa predominância simplesmente a um maciço investimento de capital: o Google nasceu pequeno, como "mera" pesquisa acadêmica – e só recebeu

investimentos de capital *depois* de se tornar um sucesso de público. Apesar dessa maciça adesão, precisamos considerar as características intrínsecas do seu critério de relevância e indagar o que resulta dele: quais suas implicações para a *prática social* em torno de um dos eventos comunicativos mais comuns no ciberespaço e para os hábitos de navegação dos usuários da Internet? A nosso ver, as premissas assumidas por Page e colegas implicam em alguns problemas importantes.

Em primeiro lugar (e esse ponto é essencial para o restante do raciocínio), é preciso considerar que o cálculo da relevância relativa dos vários documentos que atendam a uma expressão de busca, ainda que imaginado somente como critério de *ordenação* dos resultados, é, na prática, um critério de *seleção* de que subconjunto daquele conjunto D de documentos obtidos será, de fato, exibido ao usuário. Já dissemos que o produto da maioria das buscas, por envolver geralmente milhares de documentos, é impossível de ser processado pelos seres humanos. E os projetistas dos dispositivos sabem disso! Por exemplo, a busca pela expressão "theory of games" no Google resultou, em 29 de outubro de 2006, em 404 mil documentos; no entanto, apenas 710 foram exibidos pelo dispositivo. A mesma expressão no Yahoo, resultou em 120 mil documentos encontrados, mas apenas mil exibidos! Isso é consistente com o comportamento de fato dos usuários. Como já assinalamos, apenas 12% dos usuários insistem na busca além da terceira página de resultados (IPROSPECT, 2006, p. 5). E mesmo que essa minoria de usuários pacientes insista na sua pesquisa, os dispositivos não lhe permitirão seguir além de um certo ponto arbitrário. Portanto, o cálculo da relevância relativa (não importando segundo qual hipótese foi elaborado) é utilizado de fato para a *seleção* do que será exibido ao usuário e não meramente para sua ordenação.

Ora, como vimos, a hipótese do Google é baseada no critério acadêmico de contagem de citações recebidas para julgamento da importância de um pesquisador. Admitamos, por hipótese, a validade desse critério no domínio para o qual elaborado. Trata-se de um julgamento pelos próprios pares do pesquisador. Embora esse julgamento implícito nem sempre possa ser "justo", ao longo do tempo as melhores pesquisas acabam por ser reconhecidas pelo menos por parte da comunidade acadêmica. Pensemos no caso de Galileu, quase condenado à fogueira por seus pares, um destino

que (pelo menos metaforicamente) ameaça todas as idéias revolucionárias: se hoje fizer minhas as críticas dos seus detratores, provavelmente serei eu o condenado à fogueira... Ocorre que no domínio acadêmico, a contagem de citações é apenas um critério de "ordenamento", podendo servir para determinar o próximo ganhador do Prêmio Nobel, mas não sendo utilizada para decidir que pesquisas serão ou não submetidas à comunidade acadêmica. Esta dispõe de inúmeros mecanismos para assegurar a circulação de novas idéias, mesmo que produzidas por pesquisadores obscuros, oriundos de instituições igualmente obscuras, em qualquer parte do mundo. Além disso, essa comunidade é suficientemente pequena para que tais mecanismos sejam viáveis: sempre posso saber com pouco esforço o que outros pesquisadores em qualquer parte do mundo no meu domínio de interesses estão publicando, já que são relativamente poucas e facilmente acessíveis as fontes de informação a consultar.

Isso não ocorre na Internet, quando a consideramos isoladamente. Aquilo que a torna avessa aos controles e à censura – o fato de ser uma rede sem centro virtualmente ilimitada – é também aquilo que restringe o *acesso* à informação publicada: *tudo* pode estar *disponível*, mas *pouco* está efetivamente *acessível*. A circulação de idéias em grande escala na Internet depende consideravelmente dos dispositivos de busca – e se estes *selecionam* mais que ordenam o que estará acessível aos usuários, funcionam na prática como mecanismos de *bloqueio* à circulação de algumas informações, *restringindo* o que os usuários podem conhecer. Por si só (e independente de qualquer ação institucional), os dispositivos controlam, *censuram* a circulação da informação na Internet – e *o que* eles censuram está, como vimos, diretamente relacionado às hipóteses interpretativas adotadas para elaborar o critério de cálculo da relevância relativa dos documentos buscados.

Como vimos, o critério do Google é baseado na *audiência* dos *sites*: ele valoriza as páginas na proporção direta da quantidade de *links* que apontam para ela, propagando esse valor para as páginas "citadas". Isto significa que uma busca no Google retorna *somente* os *sites* mais populares para uma dada expressão de busca. Ora, idéias novas e revolucionárias não podem ser populares simplesmente porque ainda são novas e revolucionárias... Essas são as idéias *censuradas* pelo Google! Acrescente-se a isso aquelas idéias não tão novas e reacionárias, mas impopulares (admitamos que elas existam),

sobre um determinado assunto e concluiremos que o Google é avesso aos direitos de expressão das minorias. A "missão" do Google – "organizar a informação mundial e torná-la universalmente acessível e útil" (GOOGLE, 2006) – precisa ser vista sob esta ótica menos otimista. Nosso argumento é que a primeira parte da missão é verdadeira, já que "organizar" implica em *controlar*. No entanto, há que se questionar a autenticidade do restante da proposição, em particular seu advérbio...

6 Radiografia do google

É fácil imaginar um experimento visando determinar se estamos ou não sofrendo de uma Síndrome de Cassandra ideológica. Não é necessário imaginarmos idéias novas e revolucionárias ou simplesmente impopulares, pesquisando-as no Google. Aquela empresa disponibiliza uma versão *desktop* do seu dispositivo, o Google Desktop (ver http://desktop.google.com), destinada a uso em computadores pessoais, para indexação e recuperação de arquivos ali armazenados. Embora inúmeras características de implementação sejam, obviamente, diferentes do dispositivo Internet, a que nos interessa – o critério de relevância – é igualmente implementada nesta versão. Isso nos permite realizar experimentos laboratoriais inteiramente controlados, experimentos impraticáveis nas condições reais de uso da Internet, em especial numa base de oito bilhões de documentos.

Suponhamos que estamos na metade do Século XVI e que um certo Nicolau Copérnico, obscuro astrônomo polonês, acaba de publicar um curioso estudo intitulado *De revolutionibus orbium coelestium* no hipotético *website* http://www.frombork.edu na não menos hipotética Internet da Renascença. Imaginemos também que isso é tudo o que, naquele momento, os astrônomos têm à disposição para divulgar suas idéias. Como se trata de um novo estudo de um pesquisador obscuro de uma universidade obscura com idéias excessivamente diferentes daquelas admitidas não apenas pelos seus pares, mas por todos os usuários da Internet Renascentista, nenhum outro *website* se deu ao trabalho de indicá-lo com um *link* de qualquer tipo. Estamos supondo também, para simplificar o problema, que o *website* http://www.malleusmaleficarum.org, àquele tempo, não publicasse listas exaustivas de candidatos à fogueira, mas apenas daqueles efetivamente

incinerados para edificação dos fiéis; não fosse assim e Copérnico teria sido candidato ao prêmio de popularidade da Inquisição. Como não existe nenhum *link* apontando para ele, o dispositivo de busca http://www.googlorum.com classifica o documento com a menor relevância possível. O único modo de alguém interessado nos movimentos celestes localizar esse documento no Googlorum é digitar uma combinação de palavras que só exista naquele texto e em nenhum outro com relevância maior, o que não é plausível, a menos que o usuário tenha dons de clarividência, caso em que não teria necessidade de utilizar o dispositivo de busca. Num conjunto universo de pelo menos oito bilhões de documentos e milhões de *websites*, a probabilidade de um usuário, navegando ao acaso, localizar um documento específico sem a utilização de um dispositivo de busca é, para todos os efeitos práticos, desprezível. Nesse cenário, o documento de Copérnico está irremediavelmente inacessível – e nós condenados a continuar acreditando ser a Terra o centro do universo, circundada por uma esfera de estrelas fixas.

Com o Google Desktop, a plausibilidade desse cenário é facilmente verificável. Para isso, construímos uma rede de *websites* com documentos no formato HTML armazenados em disco, com as seguintes características primárias:

> i. Cinco deles, denominados fanzine_?.htm, são documentos extraídos de um *site* dedicado a um jogo para computadores bastante popular que utiliza personagens históricos, dentre eles um certo Nicolau Copérnico.
> ii. Três deles, denominados note_?.htm, são notas biográficas sobre o astrônomo Nicolau Copérnico, extraídas da Wikipedia.
> iii. Um deles, cujo arquivo recebeu o nome de wikipedia_whois.htm, contém um outro fragmento da mesma nota biográfica encontrada na Wikipedia.
> iv. Um deles, de nome wikipedia_ideas.htm, é um verbete sobre a obra *De revolutionibus*, de Copérnico, também extraído da Wikipédia.
> v. O último é um fragmento do texto original de Copérnico, encontrado na Internet, que recebeu o nome de arquivo de_revolutionibus.htm.

Os títulos dos documentos foram escolhidos ao acaso e visam tão somente melhor visualização dos resultados das buscas. Além disso, todos os documentos contêm a expressão a ser utilizada na busca, "Nicolaus Copernicus". É de se notar também que estes documentos foram

confeccionados em computador diferente daquele que seria utilizado no experimento, visando eliminar quaisquer "contaminações" do ambiente onde seria realizado o experimento. Finalmente, construímos uma rede de *links* ligando os vários hiperdocumentos, construindo um hipertexto e deixando o documento com o texto original de Copérnico fora dele, conforme *graphos* exibido na Figura 1.

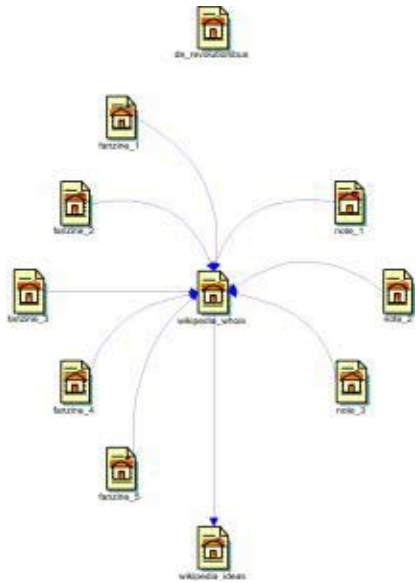

Figura 1: A estrutura de *links* dos hiperdocumentos do experimento Copérnico.

O cálculo da relevância definido para o Google indica que o documento wikipedia_whois.htm deveria receber o mais alto valor de relevância relativa, simplesmente porque é o mais citado; já o documento wikepedia_ideas.htm, deveria receber um valor intermediário porque, apesar de receber uma só citação, ela é feita pelo documento de maior relevância. Todos os demais documentos devem receber valor de relevância relativa muito próximos entre si, senão iguais.

O que se espera de um experimento que tente simular o pesquisador júnior Galileo Galilei em sua busca intelectual no ramo da astronomia utilizando nosso hipotético dispositivo http://www.googlorum.com é que, para qualquer histórico de buscas com uma mesma expressão, os documentos wikepedia_whois.htm e wikipedia_ideas.htm são sempre

devolvidos primeiro e sempre nesta ordem. Como o documento de_revolutionibus.htm recebeu o menor valor possível para a relevância e o conjunto universo de documentos obtidos numa busca qualquer é muito grande (uma busca no Google pela expressão "Nicolaus Copernicus" retornou, em 29 de outubro de 2006, 437 mil documentos), ele *nunca* será retornado em qualquer histórico de buscas.

Em vista do ambiente estar (quase que) inteiramente sob controle do observador, o protocolo da experiência pode ser bastante simples e se inicia após a instalação do Google Desktop e a completa indexação dos documentos armazenados no computador de simulação da Internet. Antes de qualquer outro passo precisamos executar a busca (pela expressão "Nicolaus Copernicus"), visando determinar o estado corrente do computador onde será realizado o experimento. Todos os documentos obtidos nesta busca serão, naturalmente, retornados nas buscas subseqüentes e deverão ser removidos dos resultados, já que não afetam o experimento. No nosso experimento, foram encontrados oito documentos nessa busca, que foram ignorados em todas as avaliações posteriores. Porém, é de se notar que, em todas as buscas realizadas, tais documentos foram sempre devolvidos ao final da lista de resultados, o que corrobora nossa decisão de ignorá-los na avaliação do experimento. Somente após este passo, os documentos do experimento foram instalados, cada um deles num diretório de arquivos diferentes, visando melhor assegurar a simulação da Internet. Após esse passo, aguardamos tempo suficiente para que os arquivos instalados fossem corretamente indexados pelo Google. Após esses procedimentos de controle, executamos a busca alvo do experimento, ilustrada no Quadro 1.

	Título da página	Localização do documento
1	Nicolaus Copernicus	C:\Temp\web_11\wikipedia_whois.htm
2	De revolutionibus orbium coelestium	C:\Temp\web_10\wikipedia_ideas.htm
3	About Copernicus	C:\Temp\web_09\note_3.htm
4	About Nicolaus Copernicus	C:\Temp\web_08\note_2.htm
5	Nicolaus Copernicus	C:\Temp\web_07\note_1.htm
6	De Revolutionibus	C:\Temp\web_01\de_revolutionibus.htm
7	Game Spot	C:\Temp\web_06\fanzine_5.htm
8	Game Center	C:\Temp\web_05\fanzine_4.htm
9	Civilization The Game	C:\Temp\web_04\fanzine_3.htm
10	Civilization IV	C:\Temp\web_03\fanzine_2.htm
11	Civilization IV Fanatics' Center	C:\Temp\web_02\fanzine_1.htm

Quadro 1: Resultado da primeira busca do experimento Copérnico.

O resultado está rigorosamente dentro do ordenamento esperado. A posição do documento de_revolutionibus.htm após os vários documentos note_?.htm é irrelevante, já que podemos atribuir esse fato a fatores acidentais. Afinal, as saídas de mesmo valor de relevância relativa precisam, ainda assim, ser ordenadas de alguma forma. Uma outra explicação plausível é a utilização subsidiária dos critérios clássicos de relevância. Isso, porém, não afeta as conclusões que queremos ter para o experimento: num universo de 437 mil documentos, o texto original de Copérnico estaria inacessível.

O passo seguinte do experimento foi abrir diretamente o arquivo fanzine_1.htm, o de menor relevância na busca. Isso foi feito cinco vezes e visava determinar o efeito dos *hits* realizados diretamente num documento independente dos resultados das buscas. Com isso, pretendíamos determinar se a *audiência* independente de um *website* afetaria a relevância de um documento, caso pudesse ser captada pelo dispositivo de busca. O resultado é visto no Quadro 2. Note-se que a relevância do documento aumentou.

	Título da página	**Localização do documento**
1	Nicolaus Copernicus	C:\Temp\web_11\wikipedia_whois.htm
2	De revolutionibus orbium coelestium	C:\Temp\web_10\wikipedia_ideas.htm
3	About Copernicus	C:\Temp\web_09\note_3.htm
4	About Nicolaus Copernicus	C:\Temp\web_08\note_2.htm
5	Nicolaus Copernicus	C:\Temp\web_07\note_1.htm
6	results.doc	E:\results.doc
7	De Revolutionibus	C:\Temp\web_01\de_revolutionibus.htm
8	Civilization IV	C:\Temp\web_03\fanzine_2.htm
9	Civilization IV Fanatics' Center	C:\Temp\web_02\fanzine_1.htm
10	Game Spot	C:\Temp\web_06\fanzine_5.htm
11	Game Center	C:\Temp\web_05\fanzine_4.htm
12	Civilization The Game	C:\Temp\web_04\fanzine_3.htm

Quadro 2: Resultado de uma busca captando audiência independente.

Este resultado deve ser visto com cautela e considerado válido tão somente para a versão *desktop* do dispositivo. Isto porque para obter o mesmo efeito na Internet, um dispositivo de busca deveria ser capaz de analisar tráfego de rede independente dos documentos indexados. Isto pode ocorrer (o Google nada nos informa), mas não nos parece plausível,

em vista do gigantismo da tarefa. O que queremos ressaltar com esse passo do experimento é sua consistência com o critério primário de cálculo da relevância relativa: o rato do nosso laboratório se comporta sempre como esperado... Mais uma observação precisa, porém, ser feita. Note-se a sexta saída da busca. Trata-se do documento que criamos para registrar cada um dos resultados do experimento que, obviamente, foi indexado pelo Google e, como continha inúmeras referências a Nicolau Copérnico, foi devolvido nas buscas. É um caso de Princípio de Incerteza de Heisenberg aplicado à nossa pesquisa: não podemos realizar qualquer experimento sem que os eventos que queremos medir sejam afetados de alguma forma pelos meios utilizados como medição no próprio experimento.

O próximo passo do protocolo foi navegar, a partir da página de resultado da busca anterior, para o documento note_1.htm, o de menor relevância no grupo, visando simular os *hits* num documento qualquer a partir das buscas realizadas no dispositivo. Esse passo, em conjunto com o seguinte, descrito no próximo parágrafo, visa determinar se buscas bem sucedidas no dispositivo afetam o resultado das buscas subseqüentes, isto é, se o dispositivo ignora ou não suas próprias páginas, realimentando o processo. A partir da mesma página de busca, efetuamos cinco *hits* no documento alvo. A seguir, efetuamos nova busca, que obteve rigorosamente o mesmo resultado exibido no Quadro 2, referente à busca anterior.

A seguir, executamos cinco vezes a mesma busca, continuando o experimento anterior, para avaliar mais diretamente a influência do próprio processo de busca no cálculo da relevância. Ora, o produto das buscas é armazenado em *caches* temporários, visando acelerar o processo de devolução dos documentos na Internet. Esses *caches* são compostos por hiperdocumentos válidos e são considerados pelo mecanismo de indexação, em especial porque contêm *links* – e *links* são a pedra de toque do dispositivo. É esperado que isso afete o experimento, como de fato ocorreu e pode ser comprovado pelos dados listados no Quadro 3. Como se pode notar, o documento note_1.htm saltou para a primeira posição da relevância (se ignorarmos o fruto espúrio do nosso próprio experimento), num processo de realimentação positiva. Novamente, isso é coerente com a filosofia de projeto do dispositivo: documentos muito encontrados em buscas pela comunidade de usuários constituem um índice de audiência que o buscador pode captar facilmente. Isso significa que o próprio dispositivo contribui para a popularidade dos *websites* mais populares, realimentando o processo.

	Título da página	Localização do documento
1	Results.doc	E:\results.doc
2	Nicolaus Copernicus	C:\Temp\web_07\note_1.htm
3	Nicolaus Copernicus	C:\Temp\web_11\wikipedia_whois.htm
4	De revolutionibus orbium coelestium	C:\Temp\web_10\wikipedia_ideas.htm
5	About Copernicus	C:\Temp\web_09\note_3.htm
6	About Nicolaus Copernicus	C:\Temp\web_08\note_2.htm
7	De Revolutionibus	C:\Temp\web_01\de_revolutionibus.htm
8	Civilization IV	C:\Temp\web_03\fanzine_2.htm
9	Civilization IV Fanatics' Center	C:\Temp\web_02\fanzine_1.htm
10	Game Spot	C:\Temp\web_06\fanzine_5.htm
11	Game Center	C:\Temp\web_05\fanzine_4.htm
12	Civilization The Game	C:\Temp\web_04\fanzine_3.htm

Quadro 3: Resultado de busca com realimentação positiva.

É preciso assinalar que os resultados rigorosamente dentro do previsto neste experimento não significam que acrescentamos algum conhecimento ao problema dos dispositivos de busca: eles decorrem claramente da concepção do Google. O experimento Copérnico apenas descreve um cenário plausível para uma das implicações da hipótese interpretativa sobre os interesses dos usuários no diálogo de busca. São essas implicações que importam. Se as premissas com as quais trabalhamos são corretas, a censura exercida pelo dispositivo de busca mais utilizado no mundo implica em dirigir a produção do discurso para um coro de múltiplas vozes que, no entanto, cantam em uníssono, sem tons discordantes, sem contrapontos nem vozes dissonantes. O que a concepção do Google acaba por produzir é o discurso do Mesmo no discurso do Outro – as idéias da multidão, nunca das minorias, repetidas por todos.

Esta é, certamente, uma situação limite, até porque a Internet fornece um conjunto de outros mecanismos de acesso à informação que permitem a circulação de idéias novas, embora não na mesma escala e com a mesma abrangência dos dispositivos de busca. No entanto, o que queremos assinalar é essa possibilidade – assustadora – implícita naquela ferramenta que muitos parecem considerar o oráculo dos novos tempos. Se assim for, a sacerdotisa só interpreta uma única profecia – a das vozes em uníssono. Infelizmente, nada temos a contrapor a esta concepção para o cálculo da relevância, em particular quando todos parecem satisfeitos com ela...

Referências bibliográficas

BAKHTIN, M. **Estética da criação verbal**. Tradução de Paulo Bezerra. 4. ed. São Paulo: Martins Fontes, 2003.

CASTELLS, M. **A galáxia da Internet**: reflexões sobre a Internet, os negócios e a sociedade. Tradução do inglês de Maria Luiza X. De A. Borges. Rio de Janeiro: Jorge Zahar, 2003.

GOOGLE corporate information: company overview. **Google**. Disponível em: <http:// www.google.com/intl/en/corporate/index.html>. Acesso em: 29 outubro 2006.

IPROSPECT. **Search engine user behavior study**. 2006. Disponível em: <http://www. iprospect.com/premiumPDFs/WhitePaper_2006_ SearchEngineUserBehavior.pdf>. Acesso em: 27 Agosto 2006.

LEMOS, A. **Cibercultura**: tecnologia e vida social na cultura. Porto Alegre: Sulina, 2002.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking**: bringing order to the web. Stanford University, 1998. Disponível em: <http://dbpubs.stanford.edu:8090/pub/1999-66>. Acesso em: 5 outubro 2005.

SULLIVAN, D. Search Engine Size Wars V erupts. **Search Engine Watch**, 11 Novembro 2004, Search Ratings & Stats. Disponível em: <http://blog.searchengine watch.com/blog/041111-084221>. Acesso em: 28 Outubro 2006.

_____. Nielsen NetRatings search engine ratings. **Search Engine Watch**, 22 Agosto 2006, Search Ratings & Stats. Disponível em: <http:// searchenginewatch.com/show Pagehtml?page=2156451>. Acesso em: 28 Outubro 2006.

THE FREE software definition. **GNU project** – the free software foundation. Disponível em: <http://www.gnu.org/philosophy/free-sw. html>. Acesso em: 25 julho 2006.

Recebido em junho de 2009
Aprovado para publicação em outubro de 2009.

SOBRE O AUTOR

Marco Antônio Gutierrez é mestre em Letras pela UERJ (Lingüística) e professor do curso de Especialização em Análise de Sistemas da Unigranrio. Autor de treze livros e inúmeros artigos publicados em periódicos brasileiros. Áreas de atuação: Lingüística

Aplicada e Engenharia de Software. Temas de pesquisa: aplicação da Teoria dos Jogos à correspondência eletrônica e pesquisa sobre critérios de relevância alternativos para os dispositivos de busca na Internet.
E-mail: magut@ism.com.br