A internet e o mito da visibilidade universal

Joaquim Paulo Serra*

Índice

1 Introdução

"O grande paradoxo do artista é ter de tornar invisível a visibilidade do artifício com que torna visível esse invisível."
Vergílio Ferreira, *Pensar*

A existência dos homens como "seres vivos políticos" (*zoon politikon*) pressupõe, antes de mais, a visibilidade de uns perante os outros no quadro de um espaço comum. Nas sociedades modernas, com as suas cidades, os seus estados nacionais e as suas organizações supranacionais, esse espaço tornou-se, cada vez mais, um espaço virtual, assegurado nos e pelos media. Neste espaço virtual - como, aliás, no espaço "real" que o antecedeu e com ele coexiste - a regra tem sido a particularidade e a desigualdade em termos daquela visibilidade; uma situação que, ainda que a propósito da "ordem do discurso", foi oportunamente tematizada por Foucault. A Internet, e em particular a world wide web, foi antevista, pelos seus fundadores, como um "espaço" que, dada a sua infinidade virtual, derivada da sua virtualidade infinita, permitiria, finalmente, assegurar a universalidade e a igualdade em termos de visibilidade. Pretendemos, na nossa comunicação, demonstrar que o funcionamento dos sistemas automáticos de busca de informação, mais concretamente dos motores de busca, assenta em critérios de relevância que impedem, desde logo, a efectivação de tal universalidade e tal igualdade; que, no fundo, a Internet não representa, neste aspecto, senão a velha política por novos meios.

2 A importância da visibilidade

Há uma tradição filosófica que, inspirando-se numa certa leitura de Parménides e da sua distinção entre a "via da verdade" (*aletheia*) e a "via da opinião" (*doxa*), se obstinou em opor a "realidade" à "aparência" e desvalorizar totalmente a segunda em relação à pri-

*Universidade da Beira Interior

meira[1]. No entanto, como observa Hannah Arendt, e pelo menos no que à realidade humana se refere, a "aparência" é constitutiva da própria "realidade"[2]. Mas, para ser efectiva, esta "aparência" ou visibilidade exige o "espaço público": um espaço que, mais do que um espaço "em si", físico, literal, que os homens se limitariam a ocupar e a tornar comum, é antes um espaço "virtual", "simbólico", criado mediante a acção (*praxis*) e o discurso (*legein*)[3] de cada um perante todos os outros[4]. Quiçá nenhum episódio ilustre tão bem - seja pelo tom trágico, seja pelo contraste que envolve - esta importância da visibilidade de uns perante os outros na definição da "condição humana" quanto o episódio em que, conhecida a terrível verdade acerca do assassínio do seu pai Laio e do casamento com sua mãe Jocasta, Édipo, optando por uma expiação que contraria a determinação dos deuses, que previa o exílio ou a morte, decide vazar os olhos[5]. Pese embora todo o peso da interpretação freudiana, que vê o acto de Édipo como símbolo da "castração", preferimos ater-nos aqui às palavras do próprio herói e das quais ressalta, como sua motivação fundamental, a impossibilidade de encarar - olhos nos olhos, como se diz - no Hades, os seus pais, e, na Cidade, os seus filhos e os seus concidadãos em geral. A cegueira que Édipo inflinge a si próprio representa, assim, mais do que a óbvia recusa de ver, a recusa de se ver a ser visto: o exílio voluntário, em si próprio, na Cidade que outrora o aclamou como herói e à visão da qual ele não quer, de forma alguma, eximir-se através da morte.[6]

3 Os modos grego e moderno da visibilidade

Reconhecer a importância da visibilidade não equivale, no entanto, a afirmar a identidade das suas modalidades. É possível, nomeadamente, distinguir entre uma modalidade própria do "espaço público dos gregos", centrado na ideia de polis, e uma modalidade própria do "espaço público burguês", centrado na categoria de "publicidade".[7]

A primeira pode ser caracterizada, de forma

[1] Cf. Jean Brun, *Os Pré-Socráticos*, Lisboa, edições 70, s/d, pp. 61-67

[2] "Para nós, a aparência - alguma coisa que está a ser vista e ouvida tanto pelos outros como por nós - constitui a realidade." Hannah Arendt, *The Human Condition*, The University of Chicago Press, 1989, p. 50. Como uma das melhores ilustrações deste "papel decisivo da mera aparência, de nos distinguirmos e sermos conspícuos no domínio dos negócios humanos", H. Arendt dá o exemplo dos trabalhadores que, quando "fizeram a sua entrada na cena da história, sentiram a necessidade de adoptar um vestuário próprio, o sans-cullote, do qual, no decorrer da Revolução francesa, derivou mesmo o seu nome." *Ibidem*, p. 218. Note-se que Hannah Arendt utiliza o termo appearence que pode, neste contexto, traduzir-se quer por aparência quer por aparecer; no que se segue utilizamos intencionalmente o primeiro termo, jogando com a ambiguidade semântica que ele comporta.

[3] A linguagem desempenha um papel tão essencial neste processo que, como é sabido, ao definir o homem como "ser vivo político" (*zoon politikon*) Aristóteles define-o, também, como "ser vivo capaz de discurso" (*zoon logon ekhon*). Cf. Aristóteles, *Política*, Livro I, 1253a5-15, Lisboa, Vega, 1998, p. 55.

[4] Hannah Arendt, *op. cit.*, pp. 198-199.

[5] Cf. Sófocles, *Rei Édipo*, verso 1330, Lisboa, Edições 70, 1999, p. 142.

[6] *Cf. ibidem*, versos 1441-1443, p. 145.

[7] Seguimos aqui a distinção de Jean-Marc Ferry, "*Las transformaciones de la publicidad politica*, in Jean-Marc Ferry, Dominique Wolton y Otros (org.), *El Nuevo Espacio Publico*, Barcelona, Gedisa, 1995, pp. 13 ss.

sumária, como presencial - assenta na presença física de cada cidadão perante todos os outros no quadro de um espaço comum, de que a agora é o símbolo por excelência -, igualitária - no sentido de uma igualdade agonística, constituindo os cidadãos uma comunidade de "iguais" (*homoioi*) de que cada um procura, pela sua acção e pelo seu discurso, atingir uma excelência (*aretê*) que lhe permita distinguir-se de todos os outros[8] - e exclusiva - ela está reservada aos cidadãos e à acção e ao discurso destes no "espaço público", excluindo todos aqueles - mulheres, crianças e escravos - e todas as actividades - biológicas, afectivas, produtivas - que só têm lugar no "espaço privado". Entendida desta forma - grega - a visibilidade confunde-se com a própria cidadania, definida por Aristóteles como a "capacidade de participar na administração da justiça e no governo".[9]

A visibilidade de cada um perante todos os outros que caracteriza a *polis* grega pressupõe obviamente, como condições fundamentais, um território e um número de cidadãos limitados.[10] Numa sociedade como a moderna, em que o território, o número de cidadãos e a complexidade da vida social aumentaram indefinidamente, conduzindo progressivamente de um homem "fixado ao solo", "localizado" e "enraizado" a um homem "móvel", "nómada" e animado pelo "ideal de ubiquidade"[11], a visibilidade torna-se uma visibilidade *in absentia*, que se efectiva num espaço - o "espaço público burguês" - cuja origem e existência é indissociável dos media, mais especificamente da imprensa. Enquanto tipo ideal, iluminista, este espaço aparece como um espaço em que todos os indivíduos, em condições de paridade, "fazem uso público da razão, com a publicitação das suas ideias e a defesa argumentativa das suas posições"[12]; em que, portanto, cada um tem direito à visibilidade perante todos os outros. Ora, cabe-nos hoje constatar que, desde o tempo em que foi construído, os factos não se têm cansado de contrariar tal tipo ideal. Com efeito, e como o mostra a "reconstrução" que Luhmann faz do conceito de "opinião pública"[13], o funcionamento dos media, mais especificamente da imprensa e do audiovisual, assenta em certas "formas" e "distinções"[14] que "determinam o que é visto e o que não é visto, o que é dito e o que não pode ser dito"[15], de um modo tal que a "evidência" do que é visto e dito - "os temas da opinião pública, as notícias e os comentários na imprensa e no audiovisual" - tem por função esconder e encobrir o que não é visto nem dito, que é apenas o "realmente importante".[16] O que esta "reconstrução" também

[8] Cf. Hannah Arendt, *op. cit.*, pp. 41, 48-49

[9] Aristóteles, *Política*, Lisboa, Vega, 1998, Livro III, 1275a20-25, p. 187. Como adiante esclarece Aristóteles, esta definição de cidadão "é sobretudo a do cidadão num regime democrático" (*ibidem*, 1275b5, p. 189).

[10] Cf. *ibidem*, Livro VII, 1326b10-20, p. 499; Hannah Arendt, op. cit., p. 43.

[11] Paul Valéry, "Notre destin et les lettres", in Ňuvres II, Paris, Gallimard, 1993, p. 1063.

[12] João Pissarra Esteves, *A Ética da Comunicação e os Media Modernos*, Lisboa, FCG-JNICT, 1998, p. 203.

[13] Cf. Niklas Luhmann, "Complexidade societal e opinião pública", in *A Improbabilidade da Comunicação*, Lisboa, Vega, 1993.

[14] Já que, como diz Luhmann, "as formas assentam sempre em distinções" (*ibidem*, p. 77). Luhmann refere-se, nomeadamente, às distinções de tempo - antes/depois (a novidade) -, de quantidade -mais/menos - e de posições de conflito - a favor/contra.

[15] *Ibidem*, p. 83.

[16] *Ibidem*, p. 85. Como observa Elisabeth Noelle-Neuman, ainda que a propósito de um outro texto

significa é que o chamado "espaço público mediático", longe de ser um espaço universal e igualitário, é um espaço em que só podem tornar-se visíveis, ser vistos e ouvidos - ser sujeitos e/ou objectos dos "temas", das "notícias" e dos "comentários" de que fala Luhmann -, os indivíduos que se enquadram em figuras ou categorias muito específicas. Utilizando uma linguagem mais ou menos metafórica, e apenas a título indicativo, diremos que essas figuras ou categorias giram à volta da distinção central entre estrelas - entendendo por tal os indivíduos que são, como se diz, "famosos", cuja visibilidade é um processo mais ou menos contínuo e cumulativo - e cometas - entendendo por tal aqueles que são, como também se diz, "ilustres desconhecidos", cuja visibilidade é descontínua e pontual. No primeiro termo da distinção incluem-se, nomeadamente, os mediadores - os próprios profissionais dos media que, tendo como função garantir a visibilidade a determinados indivíduos, a garantem em primeiro lugar a si próprios - e os notáveis - os indivíduos que se destacam em determinados campos da vida económica, política, social, cultural, desportiva, etc.[17] No segundo termo incluem-se, nomeadamente, os desviantes - os cidadãos comuns que são sujeitos ou objectos de acontecimentos que escapam à continuidade e normalidade das coisas, como no caso do homem que morde o cão[18], mas que não visam, em princípio, a visibilidade mediática -, e os provocadores - os indivíduos que desencadeiam acções que visam, em primeiro lugar, a obtenção de uma visibilidade mediática "forçada" ou "violenta", configurando aquilo a que Adriano Duarte Rodrigues chama os "meta-acontecimentos".[19] Note-se que estas figuras ou categorias não só não são mutuamente exclusivas - a mesma pessoa pode ser, simultaneamente, um notável e um desviante, como no caso do príncipe inglês, menor, que se embriaga - como o facto de um mesmo indivíduo figurar em mais do que uma figura ou categoria só o valoriza como centro de visibilidade; diríamos, aliás, que o máximo de visibilidade mediática - a "notícia explosiva", como por vezes se diz - existe sempre que uma "estrela" se torna também "cometa".

4 A Internet e os critérios de relevância dos motores de busca

A Internet está, desde os seus inícios - refiro-me aos académicos e científicos -, ligada à utopia iluminista de uma visibilidade universal e igualitária, ou, como diz António Fidalgo, de "uma rede sem centros nem peri-

de Luhmann, esta sua concepção de opinião pública aproxima-se dos resultados a que chegaram os investigadores americanos da comunicação, nomeadamente os ligados à "agenda-setting function". Cf. Elisabeth Noelle-Neuman, *La Espiral del Silencio*, Barcelona, Paidós, 1995, pp. 201-202.

[17] É a estes indivíduos que se refere, fundamentalmente, o conceito de "media events" cunhado por Daniel Dayan e Elhiu Katz. Cf. *A História em Directo. Os acontecimentos mediáticos na televisão*, Coimbra, Minerva, 1999.

[18] Cf. Adriano Duarte Rodrigues, "O acontecimento", in Nelson Traquina (org.), *Jornalismo: Questões, Teorias e "Estórias*, Lisboa, Vega, 1993. Acrescente-se que muito daquilo a que hoje se chama a "acção política", protagonizada quer pelo "governo" quer pelas "oposições" passa hoje, em grande medida, pela organização destas "provocações" - retomamos, propositadamente, esta designação da área dos serviços de informação e contra-informação - e pela visibilidade que elas conseguem nos media.

[19] *Cf. ibidem*.

ferias".[20] É certo que a Internet se distingue da imprensa e do audiovisual pelo facto de o acesso ao seu "espaço" não estar, em princípio, condicionado por quaisquer mecanismos prévios de filtragem da informação: qualquer um, em qualquer lugar, em qualquer tempo, pode publicar aí o que quiser. Mas publicar não é, obviamente, sinónimo de ser visto ou ouvido. O mesmo é dizer que também aqui existem determinados mecanismos de filtragem, de selecção e de exclusão - só que eles exercem-se a posteriori, sobre o "oceano" de informação que vai sendo acumulada. Recorrendo à imagem da "caixa negra", diremos que o que é condicionado, agora, são não as "entradas" - tudo e todos podem "entrar" - mas as "saídas"; e condicionadas em função de critérios muito específicos, como o demonstra o funcionamento dos motores de busca.

4.1 Os critérios de relevância dos motores de busca

Basicamente podemos reduzir a três as formas como pesquisamos a informação na Web, e que, não sendo incompatíveis umas com as outras podem mesmo ser vistas como complementares: a consulta de um sítio do qual conhecemos previamente o endereço, quer porque nos foi indicado por um "outro significativo", quer porque corresponde a uma instituição/organização reconhecida, quer ainda porque o encontrámos no decurso de uma pesquisa anterior, etc; a navegação sem destino certo através do "labirinto" das ligações hipertextuais, que nos vai levando de página para página, de documento para documento, muito ao estilo do *flâneur* de Baudelaire; a pesquisa através das directorias e dos motores de busca, orientada por uma palavra-chave ou uma expressão específicas.[21] Em relação às duas primeiras formas, a terceira, que é, segundo os dados disponíveis, a forma mais vulgarizada de pesquisa de informação na Web[22], coloca um problema especial: o da selecção das páginas Web relevantes de entre as centenas, os milhares e mesmo os milhões que podem ser obtidas como resposta à nossa pesquisa. É certo que podemos sempre, seja através de palavras-chave ou expressões mais especializadas, seja através dos operadores booleanos, quando utilizáveis, estreitar o âmbito da nossa pesquisa e, assim, diminuir a quantidade de páginas Web obtidas; mas um tal estreitamento e uma tal diminuição comporta sempre o risco de eliminarmos páginas Web que até poderiam vir a revelar-se como mais relevantes do que as seleccionadas. Este pro-

[20] Cf. António Fidalgo, *Metáfora e realidade ou cooperação e concorrência na rede*, 2001, disponível em www.bocc.ubi.pt. Atente-se, a propósito, na declaração do homem que, em 1989, inventou a www: "Eu tive (e ainda tenho) um sonho de que a Web podia ser menos um canal de televisão e mais um mar interactivo de conhecimento partilhado. Imagino-o imergindo-nos como um meio ambiente quente e amigável, feito de coisas que nós e os nossos amigos vimos, ouvimos, acreditámos ou imaginámos. Eu gostaria que ele tornasse os nossos amigos e colegas mais próximos, de forma a que, trabalhando neste conhecimento em conjunto, chegássemos a uma melhor compreensão." Tim Berners-Lee, *Hypertext and Our Collective Destiny*, 1995, http://www.w3.org/Talks/9510_Bush/Talk.html.

[21] As duas últimas formas costumam ser distinguidas através dos termos *browsing* e *searching*, respectivamente. A pesquisa orientada por uma palavra-chave, *keyword*, ou uma expressão, *phrase*, costuma ser designada *keyword searching*.

[22] Cf. Danny Sullivan, "GVU Survey Results" (1998), *Search Engine Watch*, http://searchenginewatch. com/reports/gvu.html.

blema da selecção, crucial quer para aqueles que colocam a informação na Web e almejam, portanto, a atenção de e a visibilidade perante cada um dos cibernautas, quer para aqueles que, por uma ou outra razão, por exemplo de investigação, fazem pesquisa de informação na Web, é tanto mais relevante quanto se sabe que, na sua maior parte, os pesquisadores da Web tendem a dar atenção apenas às dez ou vinte primeiras páginas Web seleccionadas pelos motores de busca. A questão que se coloca é, portanto, a seguinte: quais são os critérios que determinam que umas páginas sejam consideradas, pelos motores de busca, como mais "relevantes" do que outras e sejam, consequentemente, apresentadas em primeiro lugar?

Em relação a esta questão temos de fazer uma distinção entre os motores de busca ditos "da primeira geração", de que o *Lycos* e o *Altavista* são dois dos exemplos mais antigos e conhecidos, e os ditos "da segunda geração", de que o *Google* e o *Clever*[23] são dois dos exemplos mais importantes e a cujo funcionamento aqui dedicaremos uma especial atenção. Para a determinação da relevância das páginas Web, e apesar da diferença na forma como os aplicam - ou, como também se pode dizer, da diferença dos seus "algoritmos de ordenação"[24] -, os motores "da primeira geração" baseiam-se em critérios como os seguintes: a frequência absoluta ou relativa - tomando ou não em consideração o tamanho da página Web - da palavra-chave ou da expressão nas páginas Web e, eventualmente, o seu destaque mediante um tipo especial de letra; a posição da palavra-chave ou da expressão nas páginas Web, nomeadamente a sua colocação em lugares estratégicos como o título, o subtítulo, a secção inicial, as meta-etiquetas, as meta-descrições, etc.; o peso relativo de certos termos nas páginas Web que contêm as palavras-chave ou as expressões, tendo em consideração factores como a presença de termos não habituais ou incomuns, o desprezo das chamadas *stopwords*[25], etc.; a proximidade das palavras-chave ou das expressões em relação a certos termos que, por isso mesmo, serão também considerados relevantes. No entanto, a utilização destes critérios apresenta vários problemas, de entre os quais se destacam a sua grande permeabilidade em relação às diversas técnicas de *spam*[26], a sua dificuldade ou mesmo impossibilidade em lidarem com fenómenos típicos da linguagem

[23] Ainda que o *Clever* da IBM seja, ainda hoje, mais um projecto em experimentação do que um motor de busca em funcionamento efectivo, tem interesse analisar o conceito em que assenta - até por comparação com o do *Google*.

[24] Traduzimos deste modo a expressão *ranking algorithms*.

[25] *Stopwords* são palavras - como preposições, conjunções, artigos, etc. - que, por norma, se repetem em qualquer texto e que, precisamente por isso, podem ser desprezadas quando se trata de verificar e avaliar o conteúdo específico de um certo texto.

[26] No contexto dos motores de busca, *spam* designa o conjunto de processos, considerados "eticamente reprováveis", mediante os quais o criador de uma determinada página Web intenta forçar os motores de busca a seleccionarem essa página numa determinada pesquisa. Dois dos mais conhecidos e utilizados nos primeiros tempos dos motores de busca "da primeira geração" são: a repetição de uma certa palavra - supostamente, a que constituirá a palavra-chave de uma eventual busca - de forma a aumentar a sua frequência na página; a inserção de texto invisível à vista desarmada, recorrendo quer à eliminação do contraste figura-fundo quer à utilização de caracteres minúsculos. Actualmente a generalidade dos motores de busca utiliza processos que permitem contrariar, de forma mais ou menos efectiva, estes e outros processos de *spam*.

natural como a sinonímia, a homonímia ou a flexão das palavras[27], o carácter quase unilingue da Web - que é por enquanto, mais do que uma *World Wide Web*, uma *English Wide Web*, e isto apesar de alguns motores de busca já começarem ter versões em várias outras línguas.[28] Em consequência destes problemas, o resultado de um pesquisa nos motores de busca "da primeira geração" era, habitualmente, algumas páginas Web relevantes no meio de uma imensidão de páginas irrelevantes ou mesmo despropositadas em relação à busca.

Na tentativa de ultrapassarem a "cegueira quantitativa"[29] dos motores de busca "da primeira geração", os motores de busca "da segunda geração" utilizam critérios de relevância que permitem agrupá-los em duas grandes categorias: os que, como o Excite, o Northern Light, o Inference Find, o Oingo e o SimpliFind, determinam a relevância das páginas Web em função de um conceito ou campo semântico , de tal forma que são consideradas como relevantes todas as páginas circunscritas a tal conceito ou campo semântico[30]; os que determinam a relevância das páginas Web em função do comportamento dos utilizadores da mesma. Nesta segunda categoria há a considerar, por sua vez, duas subcategorias: os motores de busca que, como o *Google* e o *Clever*, têm em conta a estrutura de ligações hipertextuais que os utilizadores vão construindo, o que permite determinar quais as páginas Web que constituem quer "autoridades"[31] - páginas para que apontam ligações de páginas em grande quantidade ou de páginas que são, elas próprias, "autoridades"[32] - quer "centros" - páginas que apontam para páginas que são consideradas "autoridades"; os motores de busca que, como o DirectHit, ou "motor da popularidade", têm em conta as páginas que os utilizadores visitaram em pesquisas anteriores similares, considerando como mais "relevantes" as páginas mais visitadas.

O que de imediato ressalta, em ambos as categorias de motores de busca, e o que verdadeiramente marca a grande diferença dos motores "da segunda geração" em relação aos da primeira, é a importância crescente que tem vindo a assumir o "factor humano"[33] na determinação dos seus critérios de relevância; uma tendência que também se poderia caracterizar dizendo que, se nos motores de busca "da primeira geração" os critérios

[27] Assim, por exemplo, "films" pode não dar os resultados referentes a "movies" ou "cinema", "jaguar" tanto pode referir-se ao animal como à marca de automóvel, "car" e "cars" podem dar resultados totalmente diferentes.

[28] Estes problemas afectam também, e de forma decisiva, a indexação automática da informação - nomeadamente pelo facto de implicarem uma capacidade de computação que atrasa inexorável e crescentemente a indexação da Web em relação ao seu crescimento.

[29] Retomamos a expressão de Laura Cohen, *Second Generation Searching on the Web*, Feb. 2001, http://library.albany.edu/internet/second.html.

[30] A chamada *concept-based searching*.

[31] Ou páginas dotadas de *source authority*, no sentido em que uma página apontada pelo *Yahoo* - exemplo dos criadores do *Google* - terá mais "autoridade" do que se for apontada por uma página do sr. X.

[32] A principal diferença entre o *Google* e o *Clever* é que, enquanto o primeiro centra a determinação da relevância na utilização das "autoridades", o segundo pretende utilizar, de forma conjugada, "autoridades" e "centros" ou *hubs*; para além disso o *Google* utiliza, complementarmente, critérios como a proximidade, típicos dos motores de busca da "primeira geração".

[33] Aquilo a que, no texto atrás citado, Laura Cohen chama *the human element*.

de relevância eram essencialmente sintácticos, já nos "da segunda geração" eles são essencialmente semânticos e pragmáticos - o que não exclui, em muitos casos, alguns dos critérios sintácticos, e problemas, dos motores de busca "da primeira geração" -, levando em linha de conta a actividade humana de atribuição de "sentido".

4.2 Questionamento dos critérios de relevância dos motores de busca

O anterior não significa, no entanto, que os critérios de relevância dos motores de busca "da segunda geração" - referimo-nos, nomeadamente, à relevância por conceito ou campo semântico e à relevância por "popularidade" e por "autoridade" - não sejam problemáticos e/ou não possam ser questionados. Podemos distinguir, a este respeito, entre problemas gerais, comuns a todos os tipos de critérios de relevância e problemas específicos, que se referem a um ou a outro dos tipos de critérios de relevância.

Em relação aos problemas gerais, um problema que os motores de busca da "segunda geração" herdaram dos "da primeira geração" é o carácter globalmente relativo dos critérios de relevância, no sentido em que um mesmo documento d pode ser considerado como muito relevante pelo motor de busca X e pouco relevante pelo motor de busca Y; uma relatividade que parece apontar, à partida, para a necessidade de qualquer pesquisa utilizar mais do que um motor de busca - uma solução que, no entanto, acaba por agravar o problema que procura resolver.[34]

Em relação aos problemas específicos, o problema principal da pesquisa baseada em conceitos, na utilização dos conceitos ou campos semânticos como critérios de relevância, reside na dificuldade do estabelecimento preciso e objectivo, seja por meios estatísticos e mecânicos, seja por meios qualitativos e humanos[35], das relações semânticas entre os termos; além disso, alguns dos problemas de linguagem que afectam os motores "da primeira geração", nomeadamente a homonímia, não só não são resolvidos como acabam mesmo por se multiplicar neste tipo de pesquisa. Quanto aos critérios de relevância que assentam na "popularidade" ou na "autoridade", e apesar do sucesso que, sobretudo os segundos, têm vindo a ter[36], eles colo-

[34] Este é, também, um dos problemas que afectam os chamados "motores de meta-busca" (*meta-search engines*), a que adiante nos referiremos.

[35] Na abordagem estatística, o "conceito" é construído pelo motor de busca a partir dos termos que, de forma estatisticamente relevante, tendem a ocorrer simultaneamente com as palavras que orientam a busca; na abordagem qualitativa/humana, o "conceito" é construído a partir de uma base de conhecimento (*knowledge base*) ou *thesaurus*, dando conta das relações semânticas - sinonímia, homonímia, hiponímia-superordenação, relação parte-todo, etc. - entre os diversos termos de uma língua. O projecto WordNet, desenvolvido por George A. Miller e colegas na Universidade de Princeton, é um dos mais conhecidos exemplos desta segunda abordagem. Cf. George A. Miller, Richard Beckwith, Christane Fellbaum, Derek Gross, Katherine Miller, *Introduction to WordNet: An On-line Lexical Database* (Revised August 1993), ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf.

[36] Referimo-nos nomeadamente ao *Google*, considerado consecutivamente em 2000 e 2001 como o melhor motor de busca em aspectos essenciais como a quantidade de páginas web indexadas, a qualidade do serviço de busca da informação - em termos de

cam alguns problemas de fundo. O primeiro desses problemas é o seguinte: tais critérios não condenarão as novas páginas Web, que, como são novas, não podem ser nem "populares" nem "citadas", a uma invisibilidade inultrapassável, correndo-se assim o risco de excluir da Web informação que até poderia ser mais "relevante" do que a já existente e limitando, consequentemente, a própria riqueza da Web? O segundo desses problemas é o seguinte: o mais "popular" ou o mais "citado" será necessariamente o mais relevante? Quanto ao mais "popular", a resposta negativa parece óbvia - podendo mesmo afirmar-se que os motores de busca que assentam em tal critério mais não fazem do que desempenhar, na Web, o papel que os chamados "mass media" desempenham, há muito, fora da Web. Quanto ao mais "citado" - ao dotado de maior "autoridade", para utilizarmos um termo já referido -, o caso do Google é exemplar a este respeito e merece uma análise mais detalhada.[37]

A "coluna vertebral" do *Google* é o *PageRank*, "um método para avaliar as páginas Web objectiva e mecanicamente, medindo efectivamente o interesse e a atenção humanos a ela devotados".[38] Intuitivamente, o *PageRank* pode ser descrito dizendo que, no contexto global da Web, "uma página tem uma classificação alta se a soma das classificações das ligações que apontam para ela é alta"[39] - o que significa que a classificação da página depende tanto da quantidade das ligações que apontam para ela quanto da importância dessas mesmas ligações, sendo, portanto, completamente independente do conteúdo dessa mesma página.[40] A classificação de cada página permite definir a sua "autoridade" relativa, de um modo que se inspira de forma directa no "factor de impacto" teorizado por Eugene Garfield, o fundador do *Science Citation Index*, e com aplicação no domínio da citação científica[41] - considerando-

rapidez e relevância - e o carácter "amigável" do design. Cf. Danny Sullivan, "2001 Search Engine Watch Awards", SearchEngineWatch.com, Feb. 6, 2002, http://searchenginewatch.com/awards/2001-winners.html.

[37] Para uma descrição da arquitectura e dos princípios do *Google* pelos seus criadores, cf. Sergey Brin, Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 1998, http://www-db.stanford.edu/pub/papers/google.pdf

[38] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, 1998, http://citeseer.nj.nec.com/368196.html.

[39] *Ibidem*.

[40] No entanto, e como já referimos em nota anterior, o *Google* recorre também, a título complementar, a critérios mais "tradicionais" como o tipo de letra, a posição dos termos na página, a proximidade da página com outras páginas, etc., típicos dos motores de busca da "primeira geração".

[41] O "factor de impacto" ou impact factor obtém-se "dividindo o número de vezes que uma revista científica foi citada pelo número de artigos que publicou durante um período de tempo específico. O factor de impacto da revista reflectirá, portanto, um valor médio de citação por artigo publicado." Eugene Garfield, "Citation Analysis as a Tool in Journal Evaluation", *Essays on Information Scientist*, Vol. 1, pp. 527-544, 1962-73, reprinted from Science, (178): 471-479, 1972, p. 537. A formulação de um tal factor resulta da constatação objectiva de que, para além de factores como o mérito científico, a reputação do autor, o carácter controverso do assunto, a circulação da revista, etc., cujo peso relativo é difícil senão impossível determinar, quanto maior for o número de artigos publicados por uma revista maior é a possibilidade de tal revista ser citada - de tal modo que "a frequência de citação de uma revista científica é uma função não apenas do carácter significativo do material que ela publica (e de que a citação é um reflexo) como também da quantidade [de artigos] que ela publica". *Ibidem*. Para além do ensaio citado, cf. os seguintes

se, para o efeito, que uma ligação da página p para a página q equivale à citação de q por p e, *mutatis mutandis*, que a citação do trabalho científico t pelo trabalho científico s equivale a uma ligação de s para t. Contudo, e como reconhecem os próprios criadores do *Google*, há uma diferença abissal entre o que se passa no domínio da citação científica e o que se passa no domínio das ligações da Web: no caso do primeiro, os artigos citados são-no por membros de uma "comunidade de interpretação" que tem os seus mecanismos de selecção da informação bem definidos e os aplica de forma bastante rígida e formalizada[42] - e que, em termos gerais, impede que a publicação científica se transforme naquilo a que Georg Franck chama uma "feira de vaidades"[43]; já no caso do segundo a "citação" não obedece a quaisquer mecanismos de selecção, de tal modo que, em princípio, qualquer um pode criar as páginas que quiser, incluindo o tipo de informação que quiser e ligá-las a quaisquer outras - e não necessariamente pelas melhores razões. Deste modo, caberia aqui observar, com Tom Koch, que "o que a evolução do online não mudou é a necessidade de pesar as fontes e avaliar declarações à luz de algum critério externo".[44] É precisamente esse problema que, ao fazer a distinção entre "autoridades" e "centros", o projecto do Clever pretende ultrapassar, delineando os princípios de uma pesquisa focada em tópicos específicos e dando a perceber as "comunidades hiperligadas" a que tais tópicos correspondem.[45]

5 Conclusão

Se é verdade que, como refere Roland Barthes, e sendo o mito uma fala, "tudo o que é passível de um discurso pode ser um mito"[46], caracterizando-se este não pela

ensaios de Garfield: "Citation Indexes for Science: a New Dimension in Documentation through Association of Ideas", Science, Vol. 122, No 3159, pp. 108-111, July 15, 1955; "Citation Indexes - New Paths to Scientific Knowledge", *The Chemical Bulletin*, Chicago, 43(4): 11-12, April 1956; "Citation Analysis as a Tool in Journal Evaluation", *Essays on Information Scientist*, Vol. 1, pp. 527-544, 1962-73 (reprinted from Science, 178: 471-479, 1972).

[42] As obras epistemológicas de Thomas Kuhn, Karl Popper e Paul Feyerabend podem ser tomadas, no seu conjunto, como bons exemplos da análise - de que não está ausente um tom fortemente crítico - destas mesmas práticas.

[43] Cf. Georg Franck, "Scientific Communication - a Vanity Fair?", *Science Magazine*, Volume 286, Number 5437, Issue of 1 Oct. 1999, pp. 53-55, http://www.sciencemag.org/cgi/content/full/286/5437/53.

[44] Tom Koch, *The Message is the Medium*, Westport, Connecticut, London, Praeger, 1996, p. 188.

[45] Acerca do *Clever*, e mais especificamente acerca da relação intuitiva e algorítmica entre "autoridades" e "centros", cf.: J. Kleinberg. "Authoritative sources in a hyperlinked environment", *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998, Journal of the ACM, 46, 1999, http://www.cs.cornell.edu/home/kleinber/auth.pdf; S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Hypersearching the Web", Scientific American, June 1999, http://www.sciam.com/1999/0699issue/0699raghavan.html#link3; Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, Seyda Ertekin, *The Shape of the Web and Its Implications for Searching the Web* (2000), http://citeseer.nj.nec.com/efe00shape. html. Para uma comparação resumida entre o Google e o Clever, cf. Soumen Chakrabarti, H. Gurushyam, "Filtering Focused Information", *PC Quest*, November 11, 2000, http://www.pcquest.com/content/technology/100102901. asp.

[46] Roland Barthes, *Mitologias*, Lisboa, Edições 70, 1988, p. 181.

ocultação ou pela mentira mas pela "deformação" que produz[47], então podemos dizer que a Internet se tornou no nosso mito mais recente: no mito de que, sendo uma Rede, ela não é senão um conjunto de nós e ligações equivalentes que permitem que cada um se torne visível perante todos os outros. Ora, o que uma análise sumária do funcionamento dos motores de busca e dos seus critérios de relevância mostra é que, se a universalidade e a igualdade existem à partida, elas não existem já à chegada; também aí a particularidade e a desigualdade são a regra. Mas temos de ir mais longe e afirmar que a "deformação" reside, aqui, essencialmente no facto de que as segundas são a condição *sine qua non* das primeiras Com efeito, como poderiam constituir-se as categorias da "autoridade" ou da "popularidade" se não houvesse quem - idealmente toda a gente - acedesse ao sistema? Com a particularidade de os dotados de maior "autoridade" e "popularidade" não serem, na Internet, muito diferentes daqueles que o eram - o são - nos meios mais tradicionais como a imprensa e o audiovisual.

[47] *Ibidem*, pp. 192.