O oráculo bibliográfico: sonhos de um pesquisador

Imre Simon
Universidade de São Paulo
05508-900 São Paulo, SP, Brasil
`is@ime.usp.br`
`http://www.ime.usp.br/~is/`

28 de outubro de 2002

> *Quando se proclamou que a Biblioteca abarcava todos os livros, a primeira impressão foi de extravagante felicidade. Todos os homens sentiram-se senhores de um tesouro intacto e secreto. Não havia problema pessoal ou mundial cuja eloqüente solução não existisse: em algum hexágono.*
>
> *[...]*
>
> *À desmedida esperança sucedeu, como é natural, uma depressão excessiva. A certeza de que alguma prateleira em algum hexágono encerrava livros preciosos e de que esses livros preciosos eram inacessíveis afigurou-se quase intolerável.*
>
> *Jorge Luis Borges, A Biblioteca de Babel, 1941. [1].*

1 Um sonho: o oráculo bibliográfico

De uma forma ou de outra, venho atuando há 40 anos na área científica da computação. Com certeza, os cenários mudaram muito neste período e, com certeza também, mudarão muito mais ainda!

Durante estes anos todos venho sonhando, e cada vez mais intensamente, com um assistente virtual que pudesse ser o meu oráculo bibliográfico. Tento me explicar melhor. Imagino que este assistente conheça toda a literatura científica, com todos os pormenores. Ademais, ele está à minha disposição a qualquer hora, em qualquer lugar, respondendo as minhas infindáveis perguntas sobre a literatura científica. Felizmente, o meu assistente imaginário adora compartilhar o seu conhecimento e orgulha-se da imensidão e da precisão do seu saber, a que sempre recorre com rigorosa neutralidade.

Sou forçado a sonhar com este assistente miraculoso porque sinto-me cada vez mais incapaz para acompanhar a literatura científica e sinto também que preciso de ajuda. Não é para menos. Existem hoje 20 mil publicações científicas periódicas, com arbitragem, que publicam dois milhões de artigos a cada ano. O volume total desses artigos deve ser da ordem de 20 ou 30 milhões de páginas por ano. Encadernados em papel bíblia, tal volume ocuparia uma estante de um quilômetro de extensão linear! Só mesmo tendo um assistente para ler (e decorar) isto tudo, a uma velocidade de três metros de estante por dia, ano após ano.

Como não tenho sucesso em encontrar o meu assistente imaginado, eu, como quase todos os outros pesquisadores, entrincheiro-me atrás de uma especialização rígida. Com isto, o volume de literatura a ser lida diminui muito e, estreitando bastante a abertura da especialização, ele se adapta às disponibilidades de cada um. Mas, será que esta é uma boa idéia? Tenho as minhas dúvidas, em especial nesta época de enorme efervescência da pesquisa, principalmente nas áreas multi-disciplinares. A especialização muito rígida certamente dificulta a comunicação inteligente até mesmo entre especialidades vizinhas. O que dizer, então, de disciplinas distintas?

Nos meus sonhos, dá-me grande prazer também imaginar que eu poderia contribuir, ainda que modestamente, para o conhecimento enciclopédico deste assistente. Faria isto disponibilizando para ele os resultados das minhas pesquisas, para que ele possa incorporá-los ao seu conhecimento e usá-los ao ajudar outros pesquisadores. É uma forma singela que encontro para retribuir os imensos benefícios que tiro do meu assistente desinteressado.

2 A boa notícia: o oráculo está a caminho

O convívio com a Internet nos últimos dez anos permite concluir que o oráculo bibliográfico que procuro é tecnicamente viável. De fato, nesta seção procurarei convencer o leitor de que esta façanha é perfeitamente possível e que ela já está realizada, é operacional e livremente acessível em segmentos restritos da literatura científica. Ademais, tais segmentos manipulam uma quantidade tal de documentos que é perfeitamente possível concluir que não há empecilhos tecnológicos para a realização do meu sonho.

A notícia mais importante é que os métodos sintáticos de indexação e os de ordenação por relevância progrediram fantasticamente nos últimos anos. Faremos inicialmente um rápido passeio por estes mecanismos de uso geral.

A ponta mais visível destas tecnologias são os motores de busca. O primeiro grande impacto foi o advento do motor da AltaVista [14, 13], em fins de 1995. Ele conseguiu indexar o conteúdo integral da teia mundial pública e implementar um mecanismo extraordinariamente eficiente de consultas dos seus índices através da rede.

Surgiu, em seguida, o Google [19] que, mantendo todos os valores da AltaVista, conseguiu evoluir substancialmente introduzindo uma ordenação por relevância das páginas selecionadas. Infelizmente, não temos condições de elaborar aqui sobre as tecnologias usadas, até porque muitos dos seus detalhes são mantidos em segredo. Mas, quem usa o Google regularmente sabe que ele recompensa uma pergunta bem feita com uma pontaria certeira para o objeto procurado. Experimente buscar "livro verde" lá e confira que a primeira indicação vai para o sítio do "Livro Verde" do programa da Sociedade da Informação no Brasil [21].

Vale a pena observar que o Google indexa, hoje, dois e meio bilhões de páginas. Esta montanha de informações é processada de forma extraordinariamente eficiente. Uma consulta típica leva um décimo de segundo. A consulta "Linux", por exemplo, demora meio segundo para retornar 50 milhões de páginas, cuidadosamente ordenadas de acordo com a sua reputação.

Recentemente, o Google lançou mais um serviço de características inacreditáveis, investindo sempre nos aspectos sintáticos do texto. Estamos nos referindo à sua página de notícias, a news.google.com [20]. Este serviço monitora continuamente quatro mil sítios de notícias, entre eles o New York Times, o Washington Post, a BBC, o CNN, etc. De forma inteiramente automática, isto é, sem a intervenção de editores humanos, ele digere e classifica as novidades e a partir delas prepara um resumo com as notícias mais relevantes. Este resumo é constantemente atualizado. O resultado é um portal de notícias muito dinâmico, feito sem a intervenção de editores humanos. De fato, o portal resulta dos cálculos de algoritmos, cuidadosamente ajustados, em que estão codificados todos os critérios usados para o agrupamento das notícias (evitando repetições) e a ordenação da sua importância. A nossa ênfase aqui está em apontar que os métodos sintáticos tornaram-se poderosos a ponto de possibilitar um serviço tão complexo como este, inimaginável poucos anos atrás.

E na área da literatura científica, existe algo comparável? Sim! Os laboratórios da NEC investiram, nos últimos anos, em um serviço que responde pelos nomes de ResearchIndex ou CiteSeer [16, 6, 7]. Este serviço vasculha constantemente a Internet procurando identificar artigos científicos na área de Ciência da Computação. Encontrando um artigo, o CiteSeer desmonta o seu texto e indexa todas as palavras que nele ocorrem. Mais ainda, ele anota todas as referências bibliográficas citadas no artigo e usa estes dados para estabelecer um índice de relevância para os artigos catalogados. Tal índice é usado na ordenação das respostas a todas as consultas. As referências são usadas também para localizar os trabalhos que se basearam num dado artigo. Além de permitir a medida do impacto do artigo em questão, este aspecto permite navegar prospectivamente na literatura científica, explorando as conseqüências de uma dada idéia.

O CiteSeer é restrito, ainda, para a área da Ciência da Computação. Ainda assim, ele já indexa mais de 500 mil documentos. Embora nem todos tenham os seus textos completos eletronicamente acessíveis, o serviço existente já demonstra que é tecnologicamente viável estabelecer uma biblioteca digital completa, ricamente indexada, abrangendo toda a literatura científica.

Apontamos, finalmente, que é possível combinar o conhecimento e os critérios dos dois serviços: o Google e o CiteSeer. Basta fazer uma consulta ao Google, restringindo o espaço de busca para as páginas alojadas no CiteSeer. A busca Google "search engine site:citeseer.nj.nec.com" retorna as páginas do CiteSeer que contém as palavras "search" e "engine", ordenadas de acordo com os critérios de reputação do Google. O resultado é quase sempre muito útil para localizar uma informação procurada.

Enfim, espero ter convencido o leitor de que o oráculo bibliográfico dos meus sonhos é perfeitamente viável. O uso contínuo do CiteSeer vem reforçando o meu apetite por um serviço universal deste tipo. Universal quanto aos artigos acessíveis e universal, também, na inclusão de todas as áreas científicas. Infelizmente, existem ainda algumas dificuldades antes que possamos chegar ao meu oráculo sonhado, como veremos na próxima seção.

3 A má notícia: é preciso superar algumas dificuldades

A má notícia é que o avanço tecnológico apontado não garante, por si só, a realização do sonhado oráculo bibliográfico. Há várias dificuldades a serem enfrentadas antes de termos um repositório completo e livremente acessível de toda a literatura científica.

De forma simplificada, as dificuldades tem suas raízes num mecanismo complexo que foi montado, predominantemente nos últimos sessenta anos, para o financiamento e registro da pesquisa científica. É um mecanismo com múltiplos atores e segmentos, cada um exercendo suas funções específicas num sistema que adquiriu alta credibilidade e estabilidade. Ademais, este mecanismo é essencial para o progresso de todas as ciências e é amplamente utilizado na distribuição dos recursos disponíveis dentro do sistema. As possibilidades abertas pelo registro e disseminação da literatura científica através da Internet questionam ou desafiam muitos dos valores aceitos no sistema vigente e os segmentos atingidos procuram se mobilizar e se defender de eventuais prejuízos que temem. É importante observar, porém, que não há nenhuma contradição essencial, de nenhum valor básico, entre o sistema estabelecido de ciência e tecnologia e os novos paradigmas propostos para o registro e disseminação da literatura científica.

Um dos mecanismos mais importantes que intervém decisivamente no sistema vigente é a instituição da propriedade intelectual. Estabeleceu-se uma prática onde o autor transfere os seus direitos autorais para a entidade que publica o seu trabalho. A partir deste momento, o editor passa a ter uma enorme influência sobre a disseminação do trabalho. Por outro lado, o editor acaba detendo uma extraordinária concentração de direitos autorais que usa de acordo com os seus interesses e suas próprias percepções. Esta situação é ilustrada numa entrevista recente que Derk Haank, Presidente da Reed Elsevier, a maior editora científica atual, deu à Information Today [5].

Não é nosso propósito entrar aqui nos detalhes desta situação, que são complexos. Remetemos o leitor para artigos muito lúcidos de Andrew Odlyzko [11] e de Peter Lyman [10] sobre o tema. Apontamos, também, uma discussão muito rica e abrangente que foi patrocinada pela revista Nature, em 2001, sobre o acesso eletrônico à literatura científica [2]. Foram ouvidos todos os segmentos envolvidos na discussão e procurou-se dar a oportunidade de manifestação para todos os pontos de vista.

Ao mesmo tempo, estão surgindo vários movimentos, quase sempre iniciados por pesquisadores, no sentido de estabelecer um repositório completo e livremente acessível da literatura científica. Acreditamos que uma das maiores inspirações para estes movimentos seja o sucesso retumbante do movimento de software livre, iniciado 18 anos atrás por Richard Stallman, e que hoje, inquestionavelmente, é um fator determinante no mercado de software [12], com importantes desdobramentos, também, na formulação dos impactos sociais da Internet [8, 9]. Gostaríamos de elencar algumas iniciativas no sentido apontado.

Talvez a iniciativa mais antiga seja um movimento lançado por Stevan Harnad propondo que cada autor arquive eletronicamente o texto completo dos seus próprios trabalhos [3, 4]. Denomina-se a isto de "self-archiving". Existem cada vez mais sistemas de software (livre) que facilitam este arquivamento, entre eles citamos o eprints [18], idealizado pelo próprio Harnad.

Uma evolução importante da proposta do Harnad é a "Open Archives Initiative" [22], que sistematiza os protocolos de troca de informações bibliográficas e organiza tanto a disponibilização local dos dados (que podem ou não conter o texto completo) quanto a sua integração global, por assim chamados provedores de serviços. Os protocolos são abertos e espera-se que este mecanismo leve à realização do oráculo bibliográfico. Isto, porém, passa pela adesão maciça de autores e de instituições acadêmicas a estas práticas e a estes protocolos. Há grandes avanços nesta direção, mas o progresso é um pouco lento.

A terceira iniciativa que mencionamos é o movimento "Public Library of Science" [23], lançado no ano passado por pesquisadores ilustres das áreas biológicas. Foi um movimento que atingiu grande adesão e que motivou o debate mencionado em Nature. O movimento não conseguiu, porém, atingir o seu objetivo, que revelou ser ambicioso demais.

O movimento mais recente para o estabelecimento de um repositório completo e livremente acessível da literatura científica é a "Budapest Open Access Initiative", BOAI [15]. Este movimento é mais abrangente, menos ambicioso em termos de resultados imediatos e mais prático do que o "Public Library of Science". Ele foi lançado no início de 2002 e desde lá vem conseguindo adesões importantes. Seu principal mérito, ao nosso ver, é incentivar a disseminação da prática do "self-archiving" descrito anteriormente. A adesão da comunidade científica, porém, é um tanto lenta, infelizmente, e caso não se acelere, a realização do repositório que procuramos sofrerá atrasos.

4 À guisa de conclusão

Este artigo visa realçar dois objetivos. De um lado, queríamos chamar a atenção do leitor para o imenso valor que um repositório completo e livremente acessível da literatura científica teria como uma ferramenta de apoio à pesquisa. Além da disponibilização dos trabalhos propriamente ditos, este repositório permitiria também o processamento desta massa de textos por computadores. Isto, por sua vez, colocaria um assistente virtual, com pleno conhecimento de toda a literatura científica, à disposição de cada pesquisador, abrindo possibilidades que são difíceis de serem avaliadas com precisão, em função do poder insuspeito embutido nestes mecanismos.

O segundo objetivo era chamar a atenção para o fato de que a realização desta façanha depende, antes de mais nada, da atitude individual de cada um de nós, os pesquisadores. Somos uma corporação tradicionalmente dedicada à troca generosa de informação científica. Munidos deste espírito fomos os pioneiros no uso da Internet e estabelecemos, assim, um novo paradigma de comunicação que incorpora, na sua própria arquitetura, os nossos valores liberais de troca de informações. Não há porque supor que não nos empenhemos na realização deste novo objetivo, tão nobre, fascinante e útil quanto o anterior.

O caminho proposto para estabelecer o repositório universal da literatura científica, de livre acesso, está claro também, e é fácil de ser trilhado. Basta que cada grupo de pesquisa ou departamento instale e mantenha atualizado um servidor das informações bibliográficas e dos textos completos de seus próprios artigos científicos. A instalação dos servidores é simples e os recursos de software necessários estão livremente disponíveis. Quanto aos direitos autorais, está se espalhando cada vez mais entre os pesquisadores a atitude recomendada pelo movimento de "self-archiving" de transferir os direitos autorais da publicação, sujeito à retenção do direito de o autor arquivar o próprio trabalho segundo os padrões da "Open Archives Inititive" (OAI). Veja os detalhes em [4].

No Departamento de Ciência da Computação do Instituto de Matemática e Estatística da USP está sendo montando um servidor com as características acima [17], onde serão mantidos apontadores atualizados para a tecnologia usada e para os recursos que facilitam a instalação de servidores semelhantes.

Acreditamos que a realização cooperativa do repositório universal da literatura científica, de livre acesso, seja mais uma oportunidade para evidenciarmos o enorme potencial da Internet para uma finalidade nobre e até mesmo fascinante. Cabe a cada um fazer a sua decisão, tomar a sua atitude, com os seus próprios trabalhos. E quanto mais cedo for, melhor!

Acreditamos que a comunidade acadêmica está na iminência de realizar esta façanha, que terá conseqüências marcantes sobre o aprimoramento da qualidade e sobre a disseminação da pesquisa científica. Ademais, esta façanha aliviaria, também, a depressão justamente temida por Jorge Luis Borges, mencionada na citação inicial. Deixar passar esta oportunidade seria algo imperdoável!

Referências[1]

[1] J. L. Borges. *Ficções*. Globo, 2001. 3a. ed.

[2] D. Butler and P. Campbell. Future e-access to the primary literature. *Nature WebDebates*, 2001. `http://www.nature.com/nature/-debates/e-access/`.

[3] S. Harnad. A subversive proposal. In A. Okerson and J. O'Donnell, editors, *Scholarly Journals at the Crossroads*. Association of Research Libraries, 1995. `http://www.arl.org/scomm/subversive/toc.html`.

[4] S. Harnad. For Whom the Gate Tolls? How and Why to Free the Refereed Research Literature Online Through Author/Institution Self-Archiving, Now., 2002. `http://cogprints.ecs.soton.ac.uk/-archive/00001639/` ou `http://www.cogsci.soton.ac.uk/ harnad/-Tp/resolution.htm`.

[5] D. Kaser. Ghost in a Bottle, Elsevier Science chairman Derk Haank responds to the Public Library of Science initiative. *Information Today*, 19(2), 2002. `http://www.infotoday.com/it/feb02/kaser.htm`.

[6] S. Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001. `http://citeseer.nj.nec.com/lawrence01online.html/`.

[7] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999. `http://-citeseer.nj.nec.com/lawrence99digital.html`.

[1] Todas as URLs mencionadas foram visitadas em 28 de outubro de 2002.

[8] L. Lessig. *CODE and other laws of cyberspace.* Basic Books, 1999.

[9] L. Lessig. *The future of ideas, the fate of the Commons in a connected world.* Random House, 2001.

[10] P. Lyman. Digital documents and the future of the academic community. Scholarly Communication and Technology, April 1997. Keynote address. See <http://arl.cni.org/scomm/scat/lyman.html>.

[11] A. Odlyzko. Tragic loss or good riddance? the impending demise of traditional scholarly journals. *Intern. J. Human-Computer Studies*, 42:71–122, 1995. http://www.dtc.umn.edu/ odlyzko/doc/eworld.html.

[12] E. S. Raymond. *The Cathedral & the Bazaar. Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly & Associates, 1999.

[13] R. Seltzer, E. J. Ray, and D. S. Ray. *The AltaVista Search Revolution.* Osborne McGraw-Hill, Berkeley, 1997.

[14] AltaVista - The Search Company. http://www.altavista.com/.

[15] Budapest Open Access Initiative. http://www.soros.org/-openaccess/.

[16] CiteSeer: The NEC Research Institute Scientific Literature Digital Library. http://citeseer.nj.nec.com/.

[17] Coruja: servidor de "eprints" do Departamento de Ciência da Computação do IME-USP. http://coruja.arca.ime.usp.br/.

[18] EPrints.org - Self-Archiving and Open Archives. http://www.-eprints.org/.

[19] Google. http://www.google.com/.

[20] Google News. http://news.google.com/.

[21] Sociedade da Informação no Brasil - Livro Verde, 2000. http://www.-socinfo.org.br/livro_verde/download.htm.

[22] Open Archives Initiative. http://www.openarchives.org/.

[23] Public Library of Science. http://www.publiclibraryofscience.-org/.