INTERNET LATENT CORPUS JOURNAL
VOL. 1 N. 1
(2010) ISSN 1647-7308
http://portal.doc.ua.pt/journals/index.php/ilcj/index

Editorial

Internet: Florestas de Dados Ainda por Explorar

Ж

A internet tem expandido de forma exponencial não somente em número de utilizadores, mas também em diversidade de interacções, que acarretam uma maior produção de conteúdos e informações diversificadas. Por exemplo, o Facebook em 2010 encerrou o ano com mais de 400 milhões de utilizadores activos e com mais de 5 bilhões de pedaços de conteúdos (fotos, comentários, vídeos, links web etc.). Todas essas pessoas interligadas têm poderes de influenciar e de serem influenciadas, para o bem e para o mal, que nunca foi disponível antes na história da humanidade.

A taxa de penetração da internet é grande nos EUA, Austrália e Europa, mas no resto do mundo não se verifica (é assim). Apesar de existir aproximadamente um bilhão de utilizadores de internet no mundo, a taxa de penetração global é de apenas 15.4% segundo o relatório AberCom/SAPO (Cardoso & Espanha, 2009), mas com tendência de crescimento rápido. Um indicador desta realidade é o esgotamento de endereços do tipo IPv4 segundo a ICANN (Internet Corporation for Assigned Names and Numbers).

Apesar da riqueza e expansão, a internet não é uma fonte de dados e informações que represente todas as camadas sociais, nem tão pouco todos os países do mundo. No entanto, o poder da internet, mesmo que seja para uma parte menor da população em muitos países periféricos, tem mostrado a sua influência na sociedade em geral, em termos políticos, económicos e sociais.

Valdir Leme, gerente de marketing da rede social Orkut, ao falar sobre o triunfo desta rede social no Brasil em relação ao Facebook, comenta: “atingimos o mesmo público que o da ‘novela das nove’. Todas as classes sociais estão representadas no Orkut. Você encontra todos os seus amigos lá, independentemente de ser da classe A, B, C, D ou E” (Ikeda, 2011).

Embora toda analogia seja redutora, podemos comparar a internet com a floresta amazónica, considerando a grande “biodiversidade” existente em termos de dados e potencialidade de criação de contextos para o surgimento de novos dados. Assim como a biodiversidade amazónica, a internet tem dados em latência ainda por explorar nas ciências sociais e humanas, especialmente na educação. Battelle (2005) concorda com esta visão quando relata que “… no interior da rica base de dados do Google está um potencial campo de trabalho para milhares de doutoramentos em antropologia cultural, psicologia, história e sociologia” (p. 15).

Naturalmente, nenhum biólogo tentaria estudar as girafas ou elefantes amazónicos, este teria que se deslocar para a savana africana para estudar estes animais nos seus *habitats* naturais. Veja que os dados disponíveis na internet não têm potencial infinito para responder a todas as questões de investigações, nas diferentes áreas mencionadas. Contudo, podemos ter a certeza que actualmente os dados com potencial latente na internet são subestimados como fonte de dados para as investigações nas ciências humanas e sociais (Neri de Souza & Almeida, 2009).

Para compreender as potencialidades e limitações do “Corpus de Dados Latente na Internet” deveríamos questionar as limitações tradicionais na recolha de dados para a investigação. Em termos de acesso

aos dados, poderemos fazer investigações em todas as dimensões, áreas e assuntos em ciências humanas e sociais? A resposta sensata deve ser não. Por exemplo, é impossível desenvolver em tempo útil uma tese de doutoramento que responda à questões de investigação que necessitem de recorrer a dados secretos, que não seja ético ou que seja necessário muito tempo para os obter. Assim, podemos concluir que mesmo com a recolha de dados na "forma tradicional" há limitações em termos de quantidade e qualidade dos dados.

O que estamos a sugerir é que existe um precioso corpus latente de dados na internet disponível. Estes dados também têm limitações, muitas destas são semelhantes às limitações que ocorrem quando se procura produzir ou recolher dados para questões de investigação específicas.

Existe naturalmente, um longo percurso para construir ferramentas, técnicas e procedimentos metodológicos para lidar com dados produzidos na e da internet. Neri de Souza & Almeida (2009) propõe uma classificação geral para a investigação com a internet (Ver Figura 1).

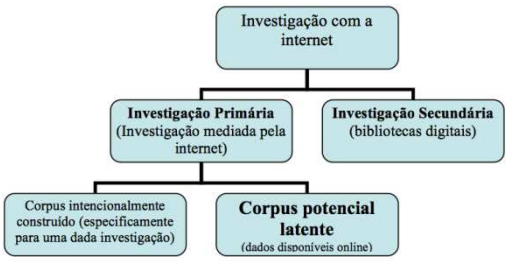

Figura 1 Classificação de investigações com a internet (Neri de Souza & Almeida, 2009)

Embora o *Internet Latent Corpus Journal* (*ILCJ*) ou Revista de Corpus Latente na Internet (RCLI) possa conter investigações cujo corpus de dados tenha sido intencionalmente construído para uma investigação específica, esta revista está claramente vocacionada para investigações que foram produzidas com base em dados que já estavam disponíveis na internet, ou que foram produzidos sem a intencionalidade de responder a um conjunto de questões de investigação específicas. No entanto, as questões de investigação que estes dados respondem são questões específicas.

Estudar dados que foram produzidos sem intencionalidade investigativa pode produzir constrangimentos na caracterização dos sujeitos e na construção dos contextos em que os dados surgiram. No entanto, a riqueza destes dados podem ser surpreendentemente diversificados, permitindo abordagens e análises para além do quantitativo. Por exemplo, podemos não saber o sexo, nem a idade de um grupo de pessoas que interagem num fórum de discussão online, mas podemos analisar os padrões de interacções na argumentação e questionamento, as dificuldades, os temas mais abordados, a dimensão semântica e linguística das interacções etc. Também podemos analisar outros indicadores numéricos, como o tempo e periodicidade dos acessos, países de acesso, *feedbacks* e outros indicadores de preferências e gostos que muitos destes sites disponibilizam através de votações.

Os artigos da primeira edição do ILCJ foram escritos por estudantes de mestrado, num período de menos de três meses, no contexto de uma unidade curricular na Universidade de Aveiro. Naturalmente, apenas alguns artigos de todos os que foram submetidos, tiveram condições de ser seleccionados para serem rescritos e revistos por três avaliadores diferentes, processo que durou mais de um ano.

Assim como os estudantes, tivemos que aprender a lidar com este tipo de dados, sendo também um desafio à sua orientação. As vantagens desta estratégia de ensino, que procura ensinar técnicas, instrumentos e metodologias de investigação, foram notórias na aprendizagem destes estudantes (Neri de Souza, Almeida, & Neri de Souza, 2010). Assim, os estudantes tiveram a oportunidade de aprender a fazer ciência, passando por todo o processo de construção e publicação da investigação.

Na Tabela 1, apresentamos um resumo geral do corpus de dados e do tipo de análise que os artigos desta primeira edição desenvolveram.

Tabela 1 Resumo dos artigos desta edição

Artigo	Corpus de dados	Tipo de Análise
1.Tutoriais online e fóruns	Conteúdos do site Pixe2life	Análise qualitativa
2.Liberdade na Internet: confessionários online?	Confissões em três site online	Análise qualitativa
3.Vídeos promocionais das Universidades no YouTube	Vídeos de 5 universidades	Análise mista
4.Utilização de ferramentas Web e teletrabalho colaborativo	Inquérito online por questionário	Análise mista
5.Utilização do Twitter pelos meios de comunicação portugueses	Conteúdos dos Twitter de 9 jornais	Análise qualitativa

A escolhemos destes artigos deve-se ao valor das questões de investigação e aos resultados alcançados, mas também porque acabam por representar uma diversidade de dados de corpus latente na internet de diferentes áreas académicas.

Apesar do claro valor dos dados numéricos, foi nos dados não-numéricos e não-estruturados (textos, vídeos, som e imagens) que incidiram na maioria das análises destes artigos. Este é um indicador metodológico importante, que mostra que os dados de corpus latente na internet podem ter sua riqueza plenamente explorada numa análise mista (Qualitativa e Quantitativa).

Esperamos que esta edição seja inspiradora para novas abordagens metodológicas para com os dados provenientes da internet e aprofundamentos em diversas áreas de investigação do conhecimento.

Universidade de Aveiro, Dezembro 2010

Francislê Neri de Souza

Bibliografia

Battelle, J. (2005). *The Search: Como o Google Mudou as Regras do Negócio e Revolucionou a Cultura* (1ª ed.). Lisboa: Casa das Letras.

Cardoso, G., & Espanha, R. (2009). *A Internet em Portugal 2009* Lisboa: Relatório OberCom - SAPO.

Ikeda, A. (2011). Líder no Brasil, Orkut completa 7 anos sem temer crescimento 'monstruoso' do Facebook. *UOL Tecnologia* Retrieved 15 Março, 2011, from http://tecnologia.uol.com.br/ultimas-noticias/redacao/2011/01/28/lider-no-brasil-orkut-completa-sete-anos-sem-temer-crescimento-monstruoso-do-facebook.jhtm

Neri de Souza, F., & Almeida, P. (2009). *Investigação em Educação em Ciência baseada em dados provenientes da internet*. Paper presented at the XIII Encontro Nacional de Educação em Ciências.

Neri de Souza, F., Almeida, P., & Neri de Souza, D. (2010, 13-14 May). *University students' scientific writing as a bridge between teaching and research.* Paper presented at the 8th International Conference The London Scholarship of Teaching and Learning (SoTL) London.