

Ocupando o Próprio Território:
Quem nos guia pela nossa Web?

Cristina Duarte Murta
Departamento de Computação
CEFET-MG, Brasil

A quantidade de informação armazenada de forma eletrônica cresce diariamente a taxas jamais imaginadas. Estimativas indicam que o Google processa atualmente mais de vinte petabytes por dia [1] e responde a bilhões de consultas por mês [2]. As tarefas relacionadas aos processos de agregar, organizar, processar e distribuir informação eletrônica constituem atualmente um enorme desafio para a computação mundial. As soluções conhecidas, embora complexas e revolucionárias, revelam-se ainda frustrantes e insuficientes após pouco tempo de uso. A diversidade das consultas e as expectativas dos usuários são crescentes, assim como a insatisfação com as respostas [3]. O negócio bilionário das máquinas de busca guia a relevância das respostas, com espaço restrito para o aspecto regional.

Agregar e tratar a informação de um país ou região é uma tarefa demasiadamente estratégica para ser realizada apenas por estrangeiros. O caráter estratégico se apresenta de várias formas e agrega ao desafio uma importância maior. Em primeiro lugar, os fluxos de informação referem-se majoritariamente a temas de interesse local ou regional. Para as tarefas do dia a dia, é necessário casar, bit a bit, a demanda da população com os serviços oferecidos pela própria população. Quem vai nos informar o telefone da mercearia da esquina, da farmácia mais próxima, ou onde conferir os principais lances do jogo de futebol ocorrido na noite anterior? Apesar da globalização, a população é mais afetada no dia a dia por fatos e aspectos locais e regionais, que influenciam diretamente atividades essenciais da vida humana, como alimentação, trânsito, clima, educação, decisões governamentais, dentre outras.

Os fluxos de informação gerados pelas consultas dos usuários são uma espécie de grande banco de dados com informações sobre as necessidades, intenções, preocupações e desejos da população. A quem confiaremos a tarefa de armazenar nossos dados, nossos arquivos, nosso banco de intenções? A quem confiaremos a tarefa de nos mostrar o que há na Web? Que regras são utilizadas para definir o que encontraremos na Web brasileira ou latino-americana? A que governos e a que tipo de legislação respondem as empresas que agregam nossas necessidades e intenções e direcionam nosso caminho pela Web? É necessário captar e processar estes fluxos, tratando-os segundo aspectos culturais próprios e oferecendo respostas relevantes e específicas considerando o contexto de cada população. Além disso, captar e processar fluxos de informação é um negócio que gera riqueza no território físico onde os fluxos são processados e agregados.

O Desafio de Pesquisa

O grande desafio proposto neste texto é gerar conhecimento e tecnologia para construção de uma máquina de indexação e busca, na América Latina, para indexar prioritariamente a Web local e regional. Este desafio representa uma pesquisa computacional complexa, com potencial para beneficiar diretamente a população de cada país envolvido e agregar competências tecnológicas específicas em várias áreas da computação. O objetivo é concreto, claro e bem definido. O tema define uma pesquisa de longo prazo, considerando a complexidade do problema e a quantidade crescente de informação a ser agregada. O tema apresenta potencial para produzir significativos avanços científicos e tecnológicos, além de oportunidade ímpar para fomentar o domínio da tecnologia de busca na Web, considerando o estágio nascente de desenvolvimento das máquinas de busca atuais.

O problema é complexo e multidisciplinar, envolvendo áreas diversas da computação como recuperação de informação, banco de dados, sistemas distribuídos, sistemas operacionais, redes de computadores, dentre outras. A pesquisa deverá tratar de questões fundamentais da computação atual com o objetivo de projetar e construir sistemas escaláveis, confiáveis e com ótimo desempenho. Alguns desafios específicos são:

- Sistemas escaláveis: para tratar quantidades crescentes de dados, é necessário que as várias etapas do processo sejam escaláveis com a Web, desde a coleta de dados, o armazenamento, a indexação, até o tratamento das consultas; este requisito perpassa todo o sistema e requer a construção de algoritmos e sistemas de hardware e de software com esta característica.
- Respostas relevantes, específicas, reproduzíveis: embora a busca na Web global seja mais interessante para o público da academia, a população em geral busca com maior frequência informações locais ou regionais; assim, indexação e busca no contexto regional deve ser priorizada, e ir além da simples informação de domínio Internet; adicionalmente, deve haver mecanismos que permitam localizar novamente e recuperar um resultado encontrado anteriormente, o que muitas vezes não conseguimos atualmente.
- Desempenho e disponibilidade: a grande quantidade de dados e a busca em tempo real requer que os sistemas computacionais e as redes apresentem excelente desempenho e disponibilidade, inclusive em dispositivos móveis e sem fio.
- Acesso ao sistema em novos dispositivos: é razoável prever que, no futuro, grande parte das consultas será feita a partir de dispositivos eletrônicos de mão, como celulares, e a partir de computadores de bordo instalados em veículos em movimento. Nestes casos, a localização geográfica do usuário poderá ser um parâmetro importante para a resposta, e os desafios são relacionados à comunicação via rede sem fio entre o sistema de busca e o dispositivo em movimento.

Este desafio é uma pesquisa de longo prazo e requer o envolvimento de grupos grandes de pesquisadores. Há grupos no continente que desenvolvem pesquisa nestes temas, porém é necessário interagir e disseminar o conhecimento para dominar esta tecnologia estratégica.

Aspecto Estratégico do Desafio para a América Latina

O conceito moderno de território inclui não apenas o lugar e a área em si, com seu meio-ambiente natural, mas também os acréscimos históricos materiais realizados pelo homem e seu uso social, técnico e político [4]. O domínio sobre um território aumenta à medida que empresas e governos locais operam, gerenciam e provêm suporte aos fluxos que trafegam no território, sejam fluxos de pessoas, de bens ou de informação. Teorias geopolíticas mais atuais ensinam que o verdadeiro dono de um território é quem o ocupa, o gerencia, e coordena os fluxos de pessoas, bens e informação que trafegam por ele. A ocupação do território latino-americano, particularmente das áreas menos densamente povoadas, como a Amazônia, é questão estratégica discutida diariamente no continente.

A América Latina precisa ocupar seu território real, bem como o virtual. Não há fronteiras para a coleta de informação na Internet e na Web, mas a riqueza gerada pelos processos de agregar, tratar e servir informação certamente não é diluída ou distribuída, a exemplo de sua origem. Os fluxos de informação são fluxos de riqueza, valem mais do que petróleo, basta comparar o valor respectivo das empresas. Coletados no nosso quintal e direcionados para fora, geram riqueza no exterior, como petróleo coletado de graça e escoado pelos caminhos da Internet.

Há vários aspectos importantes e estratégicos nesta proposta. A ênfase da busca deve ser em conteúdo regional, no idioma local, de maior interesse para a maioria da população, principalmente a população que ainda terá acesso à Internet. Há oportunidade de negócios: captar, organizar, processar e distribuir fluxos de informação é um grande negócio, que agrega inegável valor ao lugar ou região onde as empresas estão instaladas. A riqueza gerada ficará concentrada na região. Os governos locais poderão se beneficiar da tecnologia gerada, indexando suas informações para facilitar o acesso da população aos serviços e facilidades digitais ofertadas.

O modelo de computação atual, computação distribuída em centenas ou milhares de máquinas, nasceu há pouco. O desenvolvimento está por vir. Não podemos nos inibir frente a progressos iniciais e abortar iniciativas que, embora complexas, são estratégicas para o continente.

Referências bibliográficas

- [1] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", *Communications of the ACM*, 51(1), pp. 107-113, January 2008.
- [2] ComScore, www.comscore.com/press/release.asp?press=2337, June 2008
- [3] Udi Manber, "Why Search is a Hard Problem", Talk at Supernova Conference 2007, http://www.readwriteweb.com/archives/udi_manber_search_is_a_hard_problem.php
- [4] Milton Santos, *A Natureza do Espaço*, 4a. edição, Editora EDUSP, 2006.

Sobre a autora

Cristina Duarte Murta é doutora em ciência da computação pela UFMG (1999) e professora do Departamento de Computação do CEFET-MG. Foi pesquisadora visitante na University of California, Davis, no período de 2005 a 2006. Trabalha nas áreas de análise e modelagem de desempenho de sistemas computacionais, com ênfase em Web e Internet. O currículo Lattes da autora pode ser acessado em <http://lattes.cnpq.br/5249973626810280>. Email: cristina@decom.cefetmg.br.