Recuperação de Informação em Grandes Volumes de Dados Multimídia Distribuídos

Nivio Ziviani
DCC/UFMG

Edleno S. de Moura
DCC/UFAM

A convergência digital, o surgimento de grandes redes de comunicação ponto a ponto e o aprimoramento de novas tecnologias, tais como sistemas embarcados e telefonia móvel, trarão consigo um dos grandes desafios para a Ciência da Computação nos próximos anos: a necessidade de se desenvolver novos algoritmos, paradigmas e tecnologias que permitam a recuperação de informação eficiente e eficaz nestes novos ambientes.

Com a convergência digital, todo tipo de aparelho eletrônico usado no dia-a-dia das pessoas poderá tanto produzir como ter acesso a informação, podendo também ser integrado a outros dispositivos em rede. Espera-se sem dúvida um crescimento cada vez maior do volume de dados multimídia neste ambiente, com a inclusão de fotos, som, imagem, vídeo, texto e outros tipos de informação que usuários podem desejar tornar disponíveis. O desenvolvimento de soluções computacionais para ajudar na busca por informação relevante nestes novos ambientes será fundamental para permitir que usuários possam tirar proveito do imenso volume de informação a ser disponibilizado.

O objetivo desta proposta de desafio de pesquisa em Ciência da Computação é desenvolver soluções para a recuperação de informação relevante em volumes exponecialmente crescentes de dados multimídia. Para obter efetividade e eficiência é fundamental o desenvolvimento de soluções escaláveis que possam responder às necessidades das futuras gerações de grandes volumes de dados distribuídos. A tendência atual é que quase tudo que vemos, lemos, ouvimos, escrevemos e medimos fique disponível em sistemas de informação computacionais.

A busca tradicional tem sido aplicada a dados estruturados na forma de registros com atributos que casam exatamente com a consulta. Com a diversificação crescente dos tipos de dados digitalizados, cobrindo praticamente todos as formas de representação de fatos, os sistemas computacionais têm que levar em conta a noção de relevância da informação. Um tipo de busca mais moderno, que permita a recuperação de informação baseada no conteúdo de tipos de dados complexos tais como imagem, video, música, texto, séries temporais ou sequências de DNA, tem que considerar a busca por relevância. A busca por relevância é baseada em uma ordenação gradual dos resultados pela importância estimada da informação recuperada para os usuários.

As máquinas de busca atuais na Web utilizam mecanismos para recuperar documentos que são baseados em índices de texto e índices de *links*. Desde que menos de 1% dos dados na Web são do tipo texto e o restante é do tipo multimídia, a nova geração de máquinas de busca têm que contemplar esses tipos heterogênos de dados. Nesses casos consultas exatas não atendem às necessidades dos usuários, seja na Web aberta ou em *intranets* das grandes corporações. Assim, o processamento tradicional de consultas tem que ser suportado por uma plataforma poderosa de computação distribuída para capacitar a próxima geração de métodos de busca com ordenação por relevância de dados digitais baseados na noção de similaridade.

A situação atual da busca na Web mostra as necessidades de soluções inovadoras:

- As máquinas de busca não *indexam conteúdo multimídia* (apenas texto e *links*).
- O modelo de busca centralizado não é capaz de lidar com as necessidades de espaço e desempenho da busca de conteúdo multimídia (necessidade de novas *estruturas de busca que sejam escaláveis e distribuídas*).
- O conteúdo multimídia muda frequentemente (necessidade de *estruturas de busca dinâmicas*).
- Os algoritmos tradicionais de coleta *pull*, em que os objetos são copiados, são inadequados para conteúdos dinâmicos.

O objetivo desa proposta é discutir novas soluções para uma infraestrutura tecnológica para a nova geração de *máquinas de busca multimídia*. O esforço de pesquisa deve contribuir para a criação de uma arquitetura *peer-to-peer* de máquina de busca distribuída, ao contrário das máquinas de busca com dados centralizados. Os principais componentes de pesquisa e suas características principais para atender aos requisitos apontados são:

- Extração automática de características e classificação de conteúdo de objetos multimídia para permitir o processo de busca distribuído.
- Estruturas de indexação dinâmicas e distribuídas (do tipo P2P) para suportar busca por similaridade.
- Inclusão da noção de relevância em sistemas de banco de dados.
- Integração de diferentes tipo de informação em sistemas de busca por relevância.
- Técnicas dinâmicas escaláveis de *caching* para permitir bom desempenho em um ambiente P2P.
- Suporte ao complexo processo de busca em ambiente P2P usando o paradigma de similaridade no processamento de consultas para permitir combinar fontes de evidência de diferentes índices multimídia, juntamente com a busca Web tradicional.
- Suporte à coleta baseada no conceito *push*, onde os provedores de informação de qualquer formato e dimensão são solicitados a publicar e "empurrar" informação para os coletores de objetos. Esta função deve ser combinada com a coleta tradicional do tipo *pull*.
- Busca baseada em contexto (considerando localização do usuário, atividade, interesses, etc.) devem ser integrados no processo de busca, explorando a arquitetura colaborativa P2P.
- Os aspectos de segurança e confiabilidade são importantes em um ambiente colaborativo P2P. A coleta do tipo *push* pode aumentar a efetividade, mas pode também facilitar o *spamming*. Além disso, parte do conteúdo multimídia (notícias, música, imagens, etc.) têm proteção de propriedade intelectual e a indexação e a busca podem ser incentivadas, mas a difusão tem que ser controlada.

Esses itens contribuem para uma arquitetura inovadora para a próxima geração P2P de máquinas de busca que resolvam as principais limitações da tecnologia atual. No caso de uma aprovação da nossa proposta pretendemos discutir em maiores detalhes os objetivos da pesquisa apresentada.

Breve Currículo dos Autores

Nivio Ziviani é Ph.D. em Ciência da Computação pela Universidade de Waterloo, Canadá. É Professor Titular do Departamento de Ciência da Computação da UFMG, onde coordena o Laboratório para Tratamento da Informação (LATIN). É Professor Emérito do Instituto de Ciências Exatas da UFMG. É co-fundador da Miner Technology Group, vendida para o Grupo Folha / UOL em junho de 1999, e da Akwan Information Technologies, vendida para a Google Inc. em julho de 2005. É autor do livro Projeto de Algoritmos e co-autor de mais de cem artigos técnicos nas áreas de algoritmos, recuperação de informação, compressão de textos e áreas relacionas. Foi co-criador da conferência SPIRE, foi General Chair da conferência ACM SIGIR 2005, e participou de dezenas de comitê de programas das principais conferências da área de recuperação de informação.

Edleno Silva de Moura é Doutor em Ciência da Computação pela Universidade Federal de Minas Gerais e Professor Adjunto do Departamento de Ciência da Computação da Universidade Federal do Amazonas (UFAM). Foi Diretor de Tecnologia da Akwan Information Technologies e atualmente é coordenador do Programa de Pós-graduação da UFAM. Tem atuado na área de recuperação de informação nos últimos dez anos, sendo autor de dezenas de artigos publicados nas principais conferências da área. É também membro do comitê de programa de diversas conferências da área de recuperação de informação, tais como ACM SIGIR, ACM CIKM e SPIRE.