

Escalabilidade e eficiência em descoberta do conhecimento em grandes volumes de dados

*Renato Ferreira, Dorgival Guedes, Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
{renato, dorgival, meira}@dcc.ufmg.br*

Dado não é informação, informação não é conhecimento, conhecimento não é sabedoria.

Além de dar nome à nossa era, informação ocupa uma posição de destaque no atual estado de desenvolvimento humano. Estamos numa época em que controlamos energia e matéria em vastas quantidades e em escalas variadas. Informação, no entanto, permanece importante nesse contexto: como deve ser organizada, e a inteligência necessária para manuseá-la. No cenário atual, vivemos um problema crescente de sobrecarga de informação: temos cada vez mais informação que pode ser relevante, e muito pouco tempo, ou capacidade computacional para avaliar adequadamente. Encontrar novos e poderosos mecanismos para organizar, encontrar e avaliar essa informação é fundamental para enfrentar esse problema. Habilitamos, assim, o próximo passo: a transmissão e armazenamento de conhecimento.

Os avanços da tecnologia, assim como a redução do custo por byte armazenado resultaram em um acúmulo sem precedentes de dados. Um exemplo tradicional de tais dados são transações eletrônicas, como compras governamentais, guias ambulatoriais e chamadas telefônicas. Outros cenários típicos onde temos observado um armazenamento crescente são mensagens eletrônicas, genômica e proteômica, bibliotecas digitais e processos judiciais. Em síntese, é um acervo gigantesco e inexplorado, onde técnicas automatizadas para extração de informação relevante e significativa poderiam causar impacto nos respectivos setores e na sociedade em geral. Por exemplo, a detecção de spams se mostra como um desafio crescente, ameaçando a evolução da mensageria eletrônica. A detecção de fraudes é outro problema crescente, e bastante coerente dentro do atual cenário nacional.

Por outro lado, notamos uma tendência de evolução das plataformas computacionais para os grandes sistemas distribuídos. Sejam eles no nível de máquinas independentes interligadas por redes de altíssima velocidade, até às arquiteturas *multi-core* atuais, com diversas unidades funcionais independentes dentro de um mesmo processador. Do ponto de vista de software temos, pois, uma tendência forte para processamento paralelo. Os modelos de programação paralela e os respectivos ambientes de programação se mostram bastante inadequados para garantir a escalabilidade e a eficiência de implementações paralelas de muitas das aplicações atuais. Aplicações essas que chamamos de irregulares, já que a demanda computacional é fortemente dependente dos dados de entrada, podendo variar significativamente em função deles. É importante ressaltar a importância das operações de entrada e saída nesse cenário de processamento de grandes volumes de dados.

Do ponto de vista das aplicações propriamente, podemos distinguir alguns fatores que explicam a inexistência de soluções eficientes:

- Não apenas o volume de dados é crescente, como os dados “evoluem”, ou seja, estratégias de extração de informação perdem a sua efetividade se não forem atualizadas. Novamente, mensagens indesejáveis e fraudes são um caso típico. Essas características demandam o

desenvolvimento de técnicas que também sejam evolutivas.

- As técnicas correntes ainda são muito limitadas, em particular com relação à complexidade dos padrões de informação que elas determinam. Em geral os padrões são simples e homogêneos, não permitindo a composição de padrões. Por exemplo, se considerarmos um leilão eletrônico, temos claramente três tipos de padrões que se entrelaçam: as relações entre os participantes, a série temporal em termos de quando os lances foram dados, e a série de lances oferecidos. Não há hoje técnica que permita extrair informações que congreguem esses três tipos de padrões.
- A paralelização dos algoritmos de extração de informação é uma necessidade, tanto pelas máquinas atuais já não serem suficientes, quanto pela tendência da colocação de vários processadores em um mesmo chip. A dificuldade nesse caso é que os algoritmos que determinam esses padrões são complexos, seu comportamento depende da entrada e os problemas são intrinsecamente irregulares, ou seja, obter paralelizações escaláveis e eficientes é um desafio.

Do ponto de vista de ambientes e modelos de programação, também conseguimos distinguir alguns desafios:

- Nossa experiência mostra que temos que explorar não apenas as duas estratégias tradicionais de paralelismo (dados e tarefa) como ainda possibilitar o máximo de sobreposição entre os diversos componentes de um sistema de computação: processamento, comunicação e entrada/saída.
- Os dados de entrada são muito grandes, muitas vezes maiores do que os dados de saída. As operações de entrada e saída se tornam componentes importantes nesse cenário, e devem ser levados em consideração.
- Um outro aspecto que demanda reavaliação é a semântica associada aos dados. Em geral, dados são agrupados em dois grandes grupos: numéricos e categóricos, ou seja, considera-se apenas a sua natureza e relações de cardinalidade e hierárquicas não são consideradas. O uso extensivo de linguagens como XML provê uma semântica mais rica durante a extração de dados, mas o modelo de programação não contempla tal riqueza, restringindo os algoritmos e mesmo a sua aplicação, assim como sobrecarregando o desenvolvedor que incorpora manualmente no código a semântica dos dados ou o usuário que tem que realizar laborioso processo de engenharia de dados.

Desta forma, consideramos que a extração automatizada de informações se mostra como um desafio para uma significativa parcela da comunidade de ciência da computação, abrangendo desde as arquiteturas dos computadores e das redes que os conectam até a sua aplicação a um contexto específico, passando por novos algoritmos e a sua paralelização. É importante ressaltar que as limitações correntes já tem restringido o escopo e aplicabilidade das técnicas, fazendo por agravar cada vez mais o abismo entre dados e informações.

No sentido de fomentar pesquisas nos tópicos relacionados, sugerimos a criação de um conjunto de benchmarks, a serem atualizados periodicamente, que permitam à comunidade avaliar o seu progresso. Temos vários exemplos de situações onde esses benchmarks foram fundamentais para o progresso da área, como a coleção TREC na área de recuperação de informação. No nosso caso, há um significativo número de bases, inclusive públicas, que poderiam ser utilizadas para esse fim.

Breve currículo dos autores

Renato Ferreira é professor adjunto do Departamento de Ciência da Computação da UFMG, é Ph.D. pela University of Maryland, College Park em Ciência da Computação em 2001. Realizou pós-doutorado no Department of Biomedical Informatics da Ohio State University. Coordenador de Recursos Computacionais do DCC, atua como pesquisador no suporte ao desenvolvimento e execução de aplicações de alto desempenho em plataformas paralelas e distribuídas. Pesquisador em produtividade do CNPq, publica regularmente e fóruns nacionais e internacionais e atua como consultor ad-hoc para o CNPq, CAPES e FAPEMIG. É um dos coordenadores do Projeto Tamanduá.

Dorgival Guedes é professor adjunto do Departamento de Ciência da Computação da UFMG. Engenheiro Eletricista e Mestre em Ciência da Computação pela UFMG, Ph.D. em Ciência da Computação pela University of Arizona, Tucson, sob orientação do Professor Larry Peterson. Suas áreas de trabalho principais são Redes de Computadores, Sistemas Operacionais e Sistemas Distribuídos. Nessas áreas, seus interesses incluem o desenvolvimento de protocolos e novas arquiteturas de aplicação (tais como redes overlay e sistemas peer-to-peer), escalabilidade de sistemas e análise de desempenho. O Prof. Dorgival publica regularmente em veículos nacionais e internacionais de prestígio de sua área, além de já ter publicado um livro. É um dos coordenadores do Projeto Tamanduá (FINEP) e responsável pelo nó da rede PlanetLab instalado na UFMG.

Wagner Meira Jr. é professor adjunto do Departamento de Ciência da Computação da UFMG, tendo obtido o seu doutoramento na University of Rochester em Ciência da Computação em 1997. Suas áreas de interesse são processamento paralelo e distribuído e mineração de dados, tendo publicado mais de cem artigos em fóruns internacionais e nacionais, além de dois livros. Pesquisador em produtividade do CNPq, recebeu vários prêmios como orientador de mestrado e iniciação científica, além de ter sido contemplado com um IBM Faculty Award em 2004. Wagner é um dos coordenadores do Projeto Tamanduá, que tem por objetivo construir uma plataforma de serviços de mineração de dados escalável e eficiente, a qual vem sendo utilizada para análise de compras governamentais, procedimentos de alto custo do SUS e ocorrências policiais, entre outras.