Escalabilidade e Eficiência na Descoberta do Conhecimento em Grandes Volumes de Dados

Renato Ferreira *Dorgival Guedes* *Wagner Meira Jr.*

DCC-UFMG

Cenário

- Acúmulo crescente de grandes volumes de dados
  - Redução do custo por byte armazenado
  - Avanço da tecnologia, incluindo oportunidades crescentes de paralelismo e distribuição
- Demanda premente por mecanismos para organizar, encontrar, resumir e avaliar dados
- Exemplos:
  - email (spam), comércio eletrônico, prontuários médicos, governança eletrônica, bibliotecas digitais e bioinformática

Por que não temos soluções escaláveis e eficientes?

- Não apenas o volume de dados é crescente, mas o conhecimento associado evolue
- Técnicas correntes de descoberta do conhecimento são limitadas, em particular com relação aos padrões descobertos, que são simples e homogêneos
- Escalabilidade e eficiência computacional já são insuficientes para os algoritmos atuais, os quais se caracterizam pela irregularidade, iteratividade, intensidade computacional e de E/S

Alguns desafios de pesquisa

- Paralelizações escaláveis e eficientes devem explorar não apenas paralelismo de dados e tarefas, mas permitir o máximo de concorrência em termos dos componentes do sistema
- Todos os mecanismos afetados pelo grande volume de dados devem ser aperfeiçoados
- Algoritmos devem levar em consideração a semântica associada aos dados (p.ex., seqüências, hierarquias e relações de cardinalidade)

Sumário

- A descoberta automatizada de conhecimento é um problema não resolvido e em expansão, trazendo desafios para uma parcela significativa da comunidade acadêmica de CC: arquitetura e redes de computadores, processamento paralelo e distribuído, recuperação de informação, banco de dados, inteligência artificial, algoritmos e teoria.
- Além de ações de fomento, sugerimos a criação de um conjunto de *benchmarks* para essas tarefas de forma a podermos estabelecer metas e avaliar resultados.