P-Canopy uma proposta para blocagem paralela

Guilherme Dal Bianco[1], Renata Galante[1], Mirella M. Moro[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
[2]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) – 31.290-901 – Belo Horizonte – MG – Brasil

{gbianco,galante}@inf.ufrgs.br, mirella@dcc.ufmg.br

1. Introdução

A deduplicação tem por objetivo identificar e combinar registros que representam a mesma entidade no mundo real, melhorando, assim, a qualidade dos dados e a integração. No entanto, esta tarefa se torna muito difícil quando as bases de dados são extremamente grandes e complexas. Para reduzir o custo computacional e de recursos humanos diversos trabalhos estão sendo realizados com objetivo de estender técnicas de deduplicação para arquitetura paralela [Santos et al. 2007][Kim and Lee 2007]. Desse modo, tarefas que antes eram realizadas em dias podem ser executas em algumas horas, com alta eficiência.

Entretanto, o tratamento de algoritmos paralelos implica em algumas questões que precisam ser solucionadas. A perda de desempenho acarretada pela baixa velocidade da comunicação é um dos principais desafios, podendo determinar o bom funcionamento da técnica. Neste contexto, uma solução para reduzir este problema é a blocagem, na qual registros semelhantes são inseridos em um mesmo grupo, restringindo o número de possíveis réplicas para elementos de um mesmo bloco. A blocagem possibilita, portanto, que a base de dados seja fragmentada em blocos independentes e distribuída entre processos, sem a necessidade de uma exaustiva troca de mensagens.

Sendo proposto, por este artigo, uma nova técnica de blocagem paralela chamada *P-Canopy*, como uma extensão da técnica de blocagem *Canopy* [McCallum et al. 2000]. O objetivo principal é melhorar a interdependência e, conseqüentemente, aperfeiçoar o desempenho das técnicas de deduplicação paralela.

2. *Canopy Clustering*

Inúmeras abordagens são discutidas para implementar técnicas de blocagem, mas poucas são eficientes para grandes bases de dados [Baxter et al. 2003] e, são desconhecidos trabalhos que abordem a blocagem paralela. Esta seção apresenta a técnica de blocagem *Canopy,* cujos conceitos foram utilizados para definição da proposta de blocagem paralela, descrita na próxima seção.

A técnica de blocagem Canopy é utilizada para grandes e multidimensionais conjuntos de dados devido à obtenção de bons resultados com custo computacional reduzido [Baxter et al. 2003]. A melhora de eficiência é obtida através da divisão em duas fases: agrupamento dos conjuntos utilizando um método de medida de similaridade aproximada, com baixo custo computacional; e a utilização de uma técnica precisa com alto custo computacional sobre os subconjuntos gerados. Os subconjuntos criados na primeira fase do método são chamados de *Canopies*. Um *Canopy* é simplesmente um agrupamento de elementos formado de acordo com uma medida de distância aproximada, dentro de um limiar mínimo estabelecido, como um índice invertido. Cada elemento pode

pertencer a mais de um bloco. Os *Canopies* seguem a propriedade de que pontos que não estão em um mesmo *Canopy* são tão distantes que não podem ser possível estar em um mesmo bloco. Como a distância usada para criar os grupos é aproximada, não se tem como garantir esta propriedade. Mas como os *Canopies* podem se sobrepostos através da escolha de um limiar suficientemente grande, na grande maioria dos casos, esta propriedade pode ser garantida.

3. *P-Canopy Clustering*

A técnica *Canopy* é utilizada de forma seqüencial. Este artigo propõe a extensão para o modelo paralelo, chamada de *P-Canopy*. Usufruindo das características de baixo custo computacional, independência na geração dos blocos e eficiência, a técnica original é adaptada para obter um bom desempenho com a extensão para o ambiente multiprocessado.

Seguindo o modelo seqüencial a primeira fase é responsável pela divisão da base de dados. Como o maior esforço computacional está concentrando na segunda fase, com o cálculo da similaridade entre cada par de registros, a primeira etapa será executada de forma centralizada. A etapa tem como saída blocos independentes com uma precisão aproximada. Na segunda fase, cada bloco é enviado para um processo para a execução de uma técnica mais precisa, como por exemplo *K-means* ou *Greedy Aglomerative Clustering* (GAC) [McCallum et al. 2000], de forma paralela. Como cada processo que executa a segunda etapa possui uma porção independente de dados não é necessário uma interação entre os blocos para troca de registros, restringindo a comunicação apenas para troca de informações entre a fase centralizada e distribuída.

4. Considerações Finais

Este trabalho propõe uma nova abordagem para o algoritmo de blocagem *Canopy*. Originalmente tem como característica a divisão da base de dados em duas fases, possibilitando englobar duas características primordiais em técnicas de blocagem que são: eficiência e precisão. A técnica *P-Canopy* é proposta visando estender estas características para uma arquitetura paralela. Objetivando melhorar a independência de dados, minimizando o *overhead* da comunicação que são fatores de vital influência para degradação de sistemas paralelos.

Como trabalho futuro será analisada em detalhes a proposta de blocagem paralela proposta por este artigo, para uma futura validação. Além disso, serão implementados alguns algoritmos de precisão, como *K-means* e GAC, para comparar seus desempenhos. Posteriormente, pretende-se trabalhar no desenvolvimento de outras abordagens para blocagem paralela.

Referências

Baxter R., Christen P., and Churches T. (2003) "A comparison of fast blocking methods for record linkage". In *ACM SIGKDD workshop on Data Cleaning, Record Linkage and Object Consolidation*, pages 25–27, Washington DC.

McCallum A., Nigam K., and Ungar L. (2000) "Efficient clustering of high-dimensional data sets with application to reference matching". In Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.

Santos W., Teixeira T., Machado C., Meira W., Silva A., Ferreira R., Guedes D. (2007) "A Scalable Parallel Deduplication Algorithm". In *International Symposium on Computer Architecture and High Performance Computing*.

Kim H., Lee D. (2007) "Parallel Linkage". Technical report, The Pennsylvania State University.