

UMA METODOLOGIA PARA DEDUPLICAÇÃO DE GRANDES BASES DE DADOS COM UM REDUZIDO ESFORÇO DO USUÁRIO

Guilherme Dal Bianco, Renata Galante, Carlos A. Heuser

Bolsista PRH-PB 217, gbianco@inf.ufrgs.br, Instituto de informática, Universidade Federal do Rio Grande do Sul- UFRGS

R E S U M O

MOTIVAÇÃO: A deduplicação tem como objetivo identificar quais entidades correspondem a um mesmo objeto no mundo real. O processo de identificação de duplicatas é fundamental para a integração de sistemas, redução do custo de armazenamento e processamento. Entretanto, um primeiro desafio da deduplicação é a necessidade da varredura completa da base de dados para a identificação de pares duplicados. Tal análise torna-se extremamente complexa quando o volume de dados é acentuado, especialmente quando se deseja obter uma alta eficácia com uma reduzida intervenção do usuário. Um segundo desafio, é a necessidade de intervenção de um usuário especialista para a configuração do sistema. Tal configuração é realizada através da rotulação manual de um conjunto de pares (representando duplicatas e não duplicatas) ou através da definição de limiares. A criação do conjunto de pares é uma tarefa extremamente dispendiosa que deve ser minimizada o tanto quanto possível. Da mesma forma, a escolha de bons limiares é extremamente dependente da qualidade do conjunto de dados. Por exemplo, bases de dados mais “suja” ou com mais pares duplicatas podem demandar um limiar mais baixo, já em outras bases o limiar poderá ser mais agressivo para evitar a seleção de falsos pares duplicados.

OBJETIVO: Neste trabalho, está sendo desenvolvida uma metodologia para deduplicação de grandes bases de dados cujo intuito é reduzir a intervenção do usuário sem a depreciação da eficácia. Mais especificamente, a metodologia proposta visa identificar a configuração ideal para cada base de dados requisitando somente a rotulação de uma amostra mínima de pares, ou seja, o usuário deve intervir somente através da rotulação de um conjunto reduzido de pares, selecionados de forma automática. Desta forma, o usuário é liberado da tarefa de conhecer a características do conjunto de dados e as configurações do método de deduplicação para a calibração dos diferentes limiares.

APLICAÇÃO NA INDÚSTRIA DO PETRÓLEO: A agilidade para acessar os dados é vital para a tomada de decisões na indústria de petróleo. Assim, a manutenção de bases de dados “limpas” e “livres de duplicatas” é uma tarefa essencial para a redução de gargalos em uma variada gama de sistemas. A metodologia proposta tem como objetivo fornecer o arcabouço para o desenvolvimento de uma ferramenta para a remoção de dados duplicados em grandes bases de dados.

RESULTADOS OBTIDOS: Os resultados mais expressivos alcançados referem-se à redução do esforço do usuário mantendo uma qualidade equivalente ao resultado idealmente configurado manualmente (através de limiares). A metodologia foi avaliada, com sucesso, em bases reais com mais de 3 milhões de registros. O trabalho encontra-se posicionado no estado-da-arte da deduplicação com publicações junto a renomadas conferências na área de banco de dados.