Anais do SEPE – Seminário de Ensino, Pesquisa e Extensão
Vol. VIII (2018) – ISSN 2317-7489

DEDUPLICAÇÃO DE GRANDES BASES DE DADOS ONLINE: UM ESTUDO COMPARATIVO ENTRE TÉCNICAS DE CLASSIFICAÇÃO

Jhuan Marco Dondoerfer Zamprogna (apresentador)[1]
Guilherme Dal Bianco (orientador)[2]

Resumo: Com a crescente quantidade de dados produzidos dia após dia, novas técnicas de processamento estão sendo propostas para um melhor processamento e entendimento sobre os dados coletados. Nesse cenário, uma das importantes tarefas da área é identificar entidades que se relacionem entre uma ou mais bases de dados. A deduplicação de dados online em tempo real tem como objetivo encontrar registros duplicados que se referem à mesma entidade, em um fluxo contínuo de dados. Por exemplo, em uma base de dados clínica com informações sobre pacientes, a deduplicação tem a tarefa de encontrar pacientes inseridos de forma duplicada nessa mesma base de dados. A deduplicação é baseada em três principais etapas: indexação, comparação e classificação. A indexação tem como objetivo agrupar registros com informações semelhantes e gerar os pares de dados a serem comparados. Durante a comparação, os pares de dados gerados pela indexação são comparados campo a campo por uma ou mais funções de similaridade, para que no último estágio do processo sejam classificadas como duplicatas ou não. O presente trabalho tem como objetivo aprimorar os resultados da ferramenta RedBlock, uma ferramenta capaz de processar e deduplicar grandes volumes de dados de forma online e em tempo real. O trabalho tem como enfoque a etapa de classificação, com propósito de avaliar alguns dos principais modelos de classificação baseados em aprendizado de máquina supervisionado, dado tal contexto do fluxo contínuo de grandes bases de dados. Nos resultados provisórios obtidos, um dos métodos propostos tem evidenciado ser superior em relação ao baseline, com ganhos em torno de 95% dos cenários aplicados. Em execuções com menores quantidades de ruídos, o Redblock se mantêm com bom desempenho, porém conforme os ruídos aumentam os dados se tornam menos consistentes, caindo bruscamente de desempenho. Já o método proposto permanece de forma constante independente dos cenários, caindo pouco seu rendimento conforme a quantidade de ruídos aumenta.

Palavras-chave: Integração de Dados, Deduplicação Online, Técnicas de Classificação.

1 Acadêmico do curso de Ciência da Computação, Universidade Federal da Fronteira Sul (UFFS), Campus Chapecó, Contato: jhuanmarco@gmail.com
2 Professor Doutor em Ciência da Computação, Universidade Federal da Fronteira Sul (UFFS), Campus Chapecó, Contato: guilhermedb@gmail.com

Anais do SEPE – Seminário de Ensino, Pesquisa e Extensão
Vol. VIII (2018) – ISSN 2317-7489

Categoria: Pesquisa
Área do Conhecimento: Ciências Exatas e da Terra
Formato: Comunicação Oral