Encontro de Saberes 2017 - II Mostra da Pós-Graduação

Descoberta de Características e Padrões Relacionados à Seleção de Termos no Processo de Blocagem para Resolução de Entidades

Laís Soares Caldeira (Autor), Anderson Almeida Ferreira (Orientador)

Instituição de Ensino - Universidade Federal de Ouro Preto

Palavras Chaves:

Recuperação da Informação, Integração de Dados, Resolução de Entidades, Técnicas de Blocagem, Técnicas de Processamento de Blocos.

Resumo:

A extensa variedade de informação disponível na Web motivou o desenvolvimento de aplicações que integram dados heterogêneos vindos de diferentes fontes. Na tarefa de integração de dados, pares de registros devem ser frequentemente comparados para identificar aqueles que pertencem à mesma entidade do mundo real. O processo de encontrar registros que referenciam a mesma entidade é conhecido como Resolução de Entidades, que é uma tarefa computacionalmente custosa para grandes conjuntos de dados, onde todos os pares de registros devem ser comparados. Nesse contexto, métodos de blocagem são utilizados para criar blocos que contêm os registros que são propensos a corresponder à mesma entidade no mundo real, de modo que a tarefa de Resolução de Entidades possa ser aplicada apenas a estes blocos. Técnicas de processamento de blocos aumentam ainda mais a eficiência, descartando comparações ou mesmo blocos inteiros que envolvam registros não correspondentes. Este trabalho investiga e propõe um método que avalia as características dos termos presentes em cada registro no conjunto de dados, denominado BTCI (Blocagem por Termos com Características Importantes), de modo a construir blocos que possibilitem encontrar o maior número de registros correspondentes, minimizando a quantidade de comparações desnecessárias. O intuito do trabalho é encontrar termos nos registros que possibilitem obter blocos com essas características. Além disso, pretende-se com este trabalho definir uma técnica de processamento de blocos. Tal técnica reestrutura os blocos gerados pelos métodos de blocagem baseados em redundância, removendo comparações supérfluas, reduzindo significativamente o custo computacional. Experimentos preliminares em coleções de dados da Web mostram que o BTCI tem potencial para alcançar resultados relevantes.

Publicado em:

- Evento:Encontro de Saberes 2017
- Área:CIÊNCIAS EXATAS E DA TERRA
- Subárea:CIÊNCIA DA COMPUTAÇÃO