

Resumo

Técnicas de Mineração Incremental

A crescente necessidade de exploração e análise dos dados, na procura de novo conhecimento sobre o negócio de uma organização nos seus sistemas operacionais, tem feito o estudo das técnicas de mineração de dados ganhar um grande impulso. Este pode ser notado claramente no domínio do comércio electrónico, no qual a análise do comportamento passado dos clientes é extremamente valiosa e pode, eventualmente, fazer emergir novos elementos de trabalho, bastante válidos, para a determinação do seu comportamento no futuro. Desta forma, é possível prever aquilo que um visitante de um sítio Web pode andar à procura e, então, preparar esse sítio para atender melhor as suas necessidades. Desta forma, consegue-se fazer com que o visitante permaneça mais tempo a navegar por esse sítio o que aumenta naturalmente a possibilidade dele ser atraído por novos produtos e proceder, eventualmente, à sua aquisição. Para que este objectivo possa ser alcançado, a adaptação do sítio tem de ser suficientemente rápida para que possa acompanhar a navegação do visitante, ao mesmo tempo que assegura os mais recentes padrões de comportamento de navegação dos visitantes. Isto requer um algoritmo de mineração de dados com um nível de desempenho suficientemente bom para que se possa actualizar os padrões frequentemente.

Com as constantes mudanças que ocorrem ao longo do tempo nas bases de dados, invalidando ou introduzindo novos padrões, as técnicas de mineração de dados convencionais provaram ser ineficientes, uma vez que necessitam de ser reexecutadas a fim de actualizar os resultados do processo de mineração com os dados subjacentes às modificações ocorridas na base de dados. As técnicas de mineração incremental surgiram com o intuito de evitar essa reexecução do algoritmo

para actualizar os resultados da mineração quando novos dados (incrementais) são adicionados ou dados antigos são removidos. Assim, consegue-se assegurar uma maior eficiência aos processos de mineração de dados.

Neste trabalho, analisamos algumas das diferentes estratégias e modelos para a mineração incremental de dados, dando-se particular ênfase à sua aplicação em sítios Web, visando desenvolver modelos para a descoberta de padrões de comportamento dos visitantes desses sítios e gerar automaticamente recomendações para a sua reestruturação em tempo útil. Para atingir esse objectivo projectámos e implementámos o sistema Spottrigger, que cobre todo o ciclo de vida do processo de reestruturação de um sítio Web. Este ciclo é composto, basicamente, por tarefas especialmente orientadas para a extracção de dados “crus” armazenados nos servidores Web, passar estes dados por fases intermédias de limpeza e preparação, executar uma técnica de mineração incremental para extrair padrões de navegação dos utilizadores e, finalmente, reestruturar o sítio Web de acordo com os padrões de navegação encontrados e com o perfil do próprio utilizador. Além disso, o sistema Spottrigger foi aplicado no nosso estudo de caso, o qual é baseado em dados reais provenientes de um jornal *online*. Nosso principal objectivo foi colectar, em tempo útil, alguma informação sobre o perfil dos utilizadores que num dado momento estão a consultar o sítio e, assim, fazer a reestruturação do sítio num período de tempo tão curto quanto o possível, exibindo os anúncios desejáveis, activados de acordo com o perfil do utilizador. Os anúncios do sistema estão classificados por níveis. Os sítios são reestruturados para que os anúncios de nível mais elevado sejam lançados nas páginas com maior probabilidade de serem visitadas. Nesse sentido, foi definida uma classificação das páginas para o utilizador, baseada nos padrões frequentes adquiridos através do processo de mineração incremental. Visto que o comportamento de navegação dos visitantes pode mudar ao longo do tempo, o algoritmo de mineração incremental será também responsável por capturar essas mudanças de comportamento e rapidamente actualizar os padrões.

Palavras-chave: Clickstreams; Algoritmos e Estratégias para a Mineração de Dados Incremental; Reestruturação de Sítios Web.