Resumo

Sistemas de *Data Webhousing*: Análise, Desenho, Implementação e Exploração de Sistemas Reais

A *Web* tem-se tornado um dos espaços mais apelativos para as organizações como forma de divulgação das suas actividades, promoção dos seus produtos e serviços e desenvolvimento de actividades comerciais. Todavia, os visitantes de um sítio *Web* podem facilmente saltar para um sítio da concorrência caso não encontrem rapidamente aquilo que procuram, ou se tiverem qualquer outro motivo que não seja do seu agrado. Conhecer os visitantes e garantir que os produtos, serviços ou informação são aqueles que eles procuram é imperativo. É por isso que as organizações têm tentado analisar vários tipos de questões relacionadas, por exemplo, com a forma como os clientes procuram os produtos, onde abandonam o sítio e porquê, qual a frequência de visitas dos seus clientes, quais os produtos ou serviços que mais interesse despertaram nos visitantes, enfim tudo o que possa contribuir para a melhoria do sítio e para manter ou atrair novos clientes.

Todos os movimentos e selecções dos utilizadores de um sítio *Web* podem ser acompanhados através dos "cliques" que vão fazendo ao longo do seu processo de interacção com as diversas páginas *Web*. A esta sequência de "cliques" dá-se o nome de *clickstream*. Será a partir dos dados registados pelo servidor *Web* sobre as selecções do utilizador que se poderá iniciar o estudo das suas iterações e comportamento. Contudo, o registo mantido pelos servidores *Web* forma apenas um esqueleto que terá de ser enriquecido com os registos dos vários componentes e sistemas que suportam o seu funcionamento. Este tipo de integração e conciliação de dados num único repositório é, tradicionalmente, feito no seio de um *Data Warehouse* que, pelo acréscimo dos dados de *clickstream*, se torna num *Data Webhouse*. Todo o processo de extracção, transformação e integração no *Data Webhouse* é, no entanto, dificultado pelo volume, incomplitude e heterogeneidade dos dados e pela própria tecnologia utilizada no ambiente *Web*.

Nesta dissertação, é apresentado e descrito um modelo dimensional para um *Data Webhouse* para análise de um sítio *Web* comercial. São estudadas e apresentadas algumas das suas fontes de dados bem como técnicas que podem ser utilizadas para eliminar ou reduzir os problemas existentes nos dados de *clickstream*. É descrito todo o desenvolvimento e implementação do processo de extracção, limpeza, transformação e integração de dados no *Data Webhouse* com especial relevo para as tarefas de *clickstream* - a identificação de utilizadores e agentes automáticos e a reconstrução de sessões. É apresentado o *Webuts – Web Usage Tracking Statistics*, um protótipo de um sistema de apoio à decisão para acompanhamento e análise estatística das actividades dos utilizadores de um sítio *Web* e onde se incorporam alguns dos elementos, técnicas, princípios e práticas descritas.

Palavras Chave: *Data Webhouse*, *Data Warehouse*, *Clickstream*, *Web*, logs de servidor Web, HTTP, identificação de utilizadores, identificação de sessões, heterogeneidade de dados, modelação dimensional