Resumo

A World Wide Web (Web) tem crescido em número de utilizadores, taxa de utilização e complexidade. É também cada vez mais frequente a adopção deste meio como interface de acesso aos sistemas de informação existentes nas organizações, dando origem aos designados Sistemas de Informação na Web (SI Web). À medida que aumenta a complexidade dos sítios web e a experiência dos utilizadores, tornam-se mais evidentes as mais valias de conhecer o comportamento dos utilizadores e ir ao encontro das necessidades destes.

Tendo em conta que a utilização de SI Web nem sempre corresponde ao idealizado, um estudo do comportamento efectivo dos utilizadores no sistema poderá contribuir na produção de novas versões do SI Web. Por outro lado, permitirá analisar a necessidade de introdução e alteração de funcionalidades e até detectar evoluções no domínio da aplicação.

Uma das alternativas para um estudo deste tipo, baseia-se no registo e análise de fluxos de cliques, constituídos por informação que caracteriza a interacção do utilizador com os vários elementos que integram um sítio web. Os ficheiros de log dos servidores web são uma das principais formas de obtenção desta informação. Em SI Web, a quantidade de dados a analisar atinge frequentemente proporções gigantescas, o que sugere a utilização de técnicas de data warehouse para o seu tratamento.

O armazenamento de fluxos de cliques, juntamente com outros dados contextuais, numa data warehouse, tendo em vista uma compreensão do comportamento dos utilizadores no sistema web, através de análises feitas via Web, constitui o que Kimball e Merz designaram por data webhouse. Esta é uma das várias alternativas descritas para a monitorização da utilização de SI Web. É também a opção seleccionada para estudar o comportamento dos utilizadores do SI Web instalado na maioria das instituições da Universidade do Porto.

Tendo em consideração a natureza do SI e o contexto académico onde este se insere, foi definido o modelo dimensional para a data webhouse a desenvolver. Juntamente com a descrição do modelo (dimensões e tabelas de facto), é descrita a metodologia utilizada na definição do modelo e é apresentado o processo de escolha da granularidade.

Os processos envolvidos na extracção, transformação e carregamento (ETC) dos dados na data webhouse são, depois, descritos. Aqui são também descritas as várias fontes de dados e a arquitectura definida para o processo de ETC no sistema de monitorização implementado.

Após a apresentação do processo de ETC, são descritas técnicas de análise de dados de uma data warehouse. É também definido um esquema de analise dos dados segundo vários critérios, onde se inclui a segmentação de utilizadores feita previamente. No final deste capítulo, são apresentados os resultados obtidos na análise efectuada.

Finalmente, são apresentadas linhas de trabalho futuro e conclusões sobre o trabalho desenvolvido.

Abstract

The World Wide Web (Web) is growing in number of users, usage rate and complexity. The use of this medium, as an access interface to organizational information systems (IS) and its applications, is also increasing, giving rise to what is called Web Information Systems (Web IS). With the growth of Web sites complexity and user's experience, the need to know user's behavior and meet user's demands becomes clearer.

As Web IS usage isn't always as foreseen, a study of the effective behaviour of system users, may contribute to the production of Web IS new versions. On the other side, it will allow to detect the need of new functions, the improvement of old functions and even to detect possible developments.

In the case of Web based IS this is usually done analyzing clickstreams, which have information about user's interactions with Web sites elements. The main way to get this information is through web servers log files. On Web IS, there is usually a huge amount of data, which calls for data warehouse techniques.

The storage of clickstreams and other contextual data in a data warehouse, to understand user behavior, through web based analysis, is what Kimball and Merz called a data webhouse. This is one of the several alternatives described to Web IS monitorization. It's also the one selected to study user's behavior in the Web IS used by the majority of the Universidade do Porto institutions.

The dimensional model for the data webhouse was defined having in mind the nature of the Web IS and the academic context in which it is inserted. The description of the model (dimensions and fact tables), the methodology followed to define the model and the process to choose the grain of the model are then presented.

Extraction, Transformation and Loading (ETL) processes are then described. Here are also described the several data sources and the ETL architecture of the developed data warehouse.

After the presentation of the ETL process, are described data analysis techniques for data webhouses. It's also defined an analysis outline according to several criterions, including the users segmentation done previously. In the end of tgis chapter, the analysis main results are presented.

Finally, are presented some ideas for future work and conclusions about the project.