Abstract

Data Webhousing Systems: Analysis, Design, Implementation and Operation of Real Systems

The Web is becoming one of the most appealing environments for the many organisations as a means of promoting its businesses and activities as well as a commercialisation channel. However, a Web user can easily leave one organisation's Web site for its competitors if he doesn't find what he is looking for or if he finds something unpleasant on one organisation's site. To know the site's users and making sure that the products, services or information the site is providing is what the users want is nowadays a must. That is why many organisations have started to study how their web site users browse the site, where are they leaving the site and why, how frequently do their users return, what products and services are most appealing and, in general terms, everything that may be used to improve the Web site and attract new users.

Every user moves may be tracked by retaining the clicks selections they do on the different Web pages during their visit. This flow of clicks is now called clickstream. It is the data logged by the Web server on the user's selections that will enable the organisation to study their moves and behaviour. However, the Web server log only keeps the bare bones of the user's activity. This data will have to be enriched with data collected by other systems designed to provide the Web site with contents or additional functionalities. Traditionally, the gathering and integration of data from heterogeneous data sources is done inside a Data Warehouse. By adding clickstream data to it we are creating a Data Webhouse. However, Web technology, the data volume, its heterogeneity and incompleteness will create difficulties in the process of extracting, transforming and loading data into the Data Webhouse.

In this document we present a dimensional model for a Data Webhouse whose purpose is to analyse a commercial Web site. Several data sources are presented and analised in detail. Some of the techniques used to eliminate or reduce clickstream data problems are also described. The Data Webhouse extraction, cleaning, transformation and loading process is described and special attention is paid to clickstream processing tasks such as user and robot identification and user session reconstruction. A new decision support system prototype, named Webuts - Web Usage Tracking Statistics, is presented. This system's purpose is to track and analyse a Web site users' moves and actitivities as well as generate some statistical data on the Web site operation. Its operation is based on a Data Webhouse and its development incorporated some of the elements, techniques and best practices studied and described.

Keywords: Data Webhouse, Data Warehouse, Clickstream, Web, Web server logs, HTTP, user identification, session identification, data heterogeneity, dimensional modelling