Towards a Model-Driven Framework for Web Usage Warehouse Development*

Paul Hernández, Octavio Glorio, Irene Garrigós, and Jose-Norberto Mazón

Lucentia – DLSI – University of Alicante, Spain
{phernandez,oglorio,igarrigos,jnmazon}@dlsi.ua.es

Abstract. Analyzing the usage of a website is a key issue for a company to improve decision making regarding the business processes related to the website, or the evolution of the own website. To study the Web usage we need advanced data analysis tools which require the development of a data warehouse to structure data in a multidimensional model. In this paper, we describe two possible scenarios that could arise and we claim that a model-driven approach would be useful for obtaining a multidimensional model in a comprehensive and structured way. This model will drive the development of a data warehouse in order to enhance the analysis of Web usage data: the Web usage warehouse.

Web usage analysis is the process of finding out what users are looking for on the Internet. This information is extremely valuable for understanding how a user "walks" through a website in, thus supporting decision making process.

Commercial tools for Web usage data analysis have some drawbacks: (i) significant limitations performing advanced analytical tasks, (ii) uselessness when trying to understand navigational patterns of users, (iii) inability to integrate and correlate information from different sources, or (iv) unawareness of the conceptual schema of the application. For example, one of the most known analysis tools is Google Analytics (`http://www.google.com/analytics`) which has emerged as a major solution for Web traffic analysis, but it has a limited drill-down capability and there is no way of storing data efficiently. Worse still, the user does not own the data, Google does.

There are several approaches [3,4] that define a multidimensional schema in order to analyze the Web usage by using the Web log data. With these approaches, once the data is structured, it is possible to use OLAP or data mining techniques to analyze the content of the Web logs, tackling the aforementioned problems. However, there is a lack of agreement about a methodological approach in order to detect which would be the most appropriate facts and dimensions: some of them let the analysts decide the required multidimensional elements, while others decide these elements by taking into consideration a specific Web log format. Therefore, the main problem is that the multidimensional elements are informally chosen according to a specific format, so the resulting multidimensional

* This work has been partially supported by the following projects: SERENIDAD (PEII-11-0327-7035) from Castilla-La Mancha Ministry, and MESOLAP (TIN2010-14860) from the Spanish Ministry of Education and Science.

O. De Troyer et al. (Eds.): ER 2011 Workshops, LNCS 6999, pp. 336–337, 2011.

model may be incomplete. Regarding current Web engineering approaches, to the best of our knowledge, none of them [2,1] provide mechanisms for defining a multidimensional model at the same time that the rest of the website in order to represent the Web usage. To overcome these drawbacks, a model-driven framework for developing a Web usage warehouse is proposed considering two different scenarios (see Fig. 1).

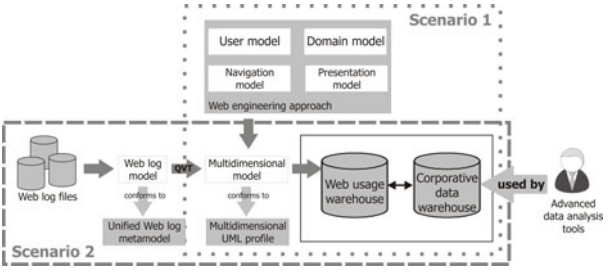

Fig. 1. Model-driven framework for Web usage warehouse development

In our first scenario (*Web usage warehouse within model-driven Web engineering*), several conceptual models have to be defined when designing a website (navigation model, user model, data model, etc.). However, none of these models are intended to represent and understand the Web usage. Therefore, multidimensional concepts (facts, dimensions, hierarchies, etc.) should be identified within the conceptual models of a given application in order to build a Web usage warehouse in an integrated and structured manner. Due to the fact that conceptual models of websites may not be available or out-of-date, within our second scenario (*Web usage warehouse from Web log data*), a Web usage warehouse is developed without requiring these conceptual models, but using Web log files. To this aim, a Web log metamodel is defined which contains the elements and the semantics that allow building a conceptual model from Web log files, which represents, in a static way, the interaction between raw data elements (i.e. the client remote address) and usage concepts (i.e. session, user).

References

1. Ceri, S., Fraternali, P., Bongio, A.: Web modeling language (WebML): a modeling language for designing web sites. Computer Networks 33(1-6), 137–157 (2000)
2. Garrigós, I.: A-OOH: Extending web application design with dynamic personalization (2008)
3. Joshi, K.P., Joshi, A., Yesha, Y.: On using a warehouse to analyze web logs. Distributed and Parallel Databases 13(2), 161–180 (2003)
4. Lopes, C.T., David, G.: Higher education web information system usage analysis with a data webhouse. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3983, pp. 78–87. Springer, Heidelberg (2006)