Chapter II

Web Mining to Create a Domain Specific Web Portal Database

Anthony Scime
State University of New York College at Brockport, USA

ABSTRACT

The dynamic nature of the World Wide Web is causing an evolution of both information access and format. The use of a Web portal to access information about a domain relieves the searcher of the responsibility to know about, access and retrieve domain documents. In a properly constructed portal, a Web mining process has already sifted through pages found on the Web to select domain facts. This Web-generated knowledge is added to domain expert knowledge in an organized database. This chapter details the design and construction of a domain specific Web portal through the combination of domain expertise and Web-based domain facts.

INTRODUCTION

On the World Wide Web, so much information is now available that meaningful information must be sorted out from the seemingly limitless volume of accessible information. In conducting a search for information on the Web, a user must know the desired URL or find the desired information by a search process. The problems that arise in searching for information on the Web are validity and volume. Accessing Web sites in a domain requires the searcher to be able to assess the validity of the site’s content. The sheer volume of sites makes it difficult to find the most comprehensive and most valid site on the Web.

A user initiates a World Wide Web search by using a search engine to find pages that refer to the desired subject. This requires the user to define the domain as a keyword or a collection of keywords that can be processed by the search engine. The searcher may not know how to define the domain as a collection of keywords, thereby limiting the search to the domain name. Even given the ability to break down the domain and conduct a number of searches, the search results still have the two problems of validity and excessive quantity. A solution is a Web portal where all the relevant information in a domain is brought together. The Web portal’s organization relieves the searcher of the necessity to decide the search keywords and to determine the valid and relevant documents from those found in a search.

To construct a Web portal, expertise about the domain provides initial information and a foundation for growth. A Web mining process adds new domain facts by selecting and sifting through the Web pages. This Web-generated knowledge is added to the domain expert knowledge and organized as a database. The value of the database/Web portal information is more than the sum of the various sources.

WEB PORTAL OVERVIEW

A domain Web portal can relieve the searcher from having to decide on validity while having expert assurance as to comprehensiveness. Comprehensiveness reduces the number of sites necessary to visit. In fact, the portal may directly provide domain knowledge. Such a portal requires previously conducted searches and construction of a Web accessible database.

Experts within a domain of knowledge are familiar with the facts and the organization of the domain. In the database design process, the analyst first extracts from the expert the domain organization. This organization is the foundation for the database structure and specifically the attributes that represent the characteristics of the domain. In large domains, it may be necessary to first identify topics of the domain, which may have attributes that are different from each other and occasionally different from the general domain. The topics become the entity sets in the domain data model. Using database design methods, the data model is converted into relational database tables. The expert’s domain facts are used to initially populate the database.

However, it is possible that the experts are either not completely knowledgeable or cannot express their knowledge about the domain. Other sources can be consulted for expert level knowledge. Expert-level knowledge is contained in data, text, and image sources. These sources lead to an expansion of domain knowledge in both domain organization and domain facts.

In the past, the expert was necessary to point the analyst to these other sources. The expert’s knowledge included knowledge such as where to find information about the domain, what books to consult, and the best data sources. Today, the World Wide Web provides the analyst with the capability to find additional information about any
Web Service Integration and Management Strategies for Large-Scale Datasets
www.igi-global.com/chapter/web-service-integration-management-strategies/7893?camid=4v1a