Part II

Acquiring Knowledge and Building Organizational Memories
Chapter V

Semi-Structured Data Extraction from Heterogenous Sources

Xiaoying Gao, University of Melbourne, Australia
Leon Sterling, University of Melbourne, Australia

Introduction

The World Wide Web is known as the “universe of network-accessible information, the embodiment of human knowledge” (W3C, 1999). Internet-based knowledge management aims to use the Internet as the world wide environment for knowledge publishing, searching, sharing, reusing, and integration, and to support collaboration and decision making. However, knowledge on the Internet is buried in documents. Most of the documents are written in languages for human readers. The knowledge contained therein cannot be easily accessed by computer programs such as knowledge management systems. In order to make the Internet “machine readable,” information extraction from Web pages becomes a crucial research problem.

The task addressed in this chapter is information extraction from heterogeneous, semi-structured documents such as Web pages generated from online services. Figure 1 is a fragment of a semi-structured document taken from online real estate advertisements and Figure 2 shows the extracted results in the Object Exchange Model (OEM) (Abiteboul, 1997). We call the smallest concepts “Knowledge Units”, which are the lowest level concepts in the concept hierarchy. For example, in the real estate advertisement domain, “suburb”, “price”, “size”, and “type” are knowledge units. The task is to extract their
Knowledge Creation and Sharing in Japanese Organisations: A Socio-Cultural Perspective on ba
Kiyoshi Murata (2011). *Ethical Issues and Social Dilemmas in Knowledge Management: Organizational Innovation* (pp. 1-16).
www.igi-global.com/chapter/knowledge-creation-sharing-japanese-organisations/48223?camid=4v1a