A Study on Retrieve and Archive of Web pages

Shien-Chiang Yu

Professor, the Department of Information and Communication,
Shih Hsin University, Taipei, Taiwan

Web resources are recording the human societal information at that time, including visual design art, style, and included various kinds of resources in these web pages. The value of very having preserved. But web resources not only grow up quickly, but also disappear fast. Web resources will be probably unable to be utilized again because of factors such as server shutdown, revision, etc. at any time. Even preserved the pages, still face the fragile characteristic of digital information. Therefore, the preserved method must address methodological and practical issues to archive and manage digital preservation. This study adopt two major research methods: content analysis and experiment. Through content analysis to explore the characteristics of web pages mode, the procedure of retrieve web content, structured data processing standards, and the relation of mapping between both. Experimental method implemented web tree down mining techniques.

In addition, to cope with the migration of HTML version, to avoid thereafter browser cannot parse the preserved web content today, this research also study long-term preservation issues, including standardization, in line with long-term application format, data extraction and restructuring and other factors. Based on these requirements, this study covers the solution of long-term preservation of web site archive. Using the way of Topic Maps syntax to reorganize unstructured HTML of original web pages into the semi-structured. Making Web content description can provide automated management, analysis and application.