The Process of Generating an Eco-Material Data Using a Web Crawling

Simhee Hong[1], Jungho Yu[2]
[1]Graduate School, Department of Architecture, Kwangwoon University, Korea
[2]Professor, Department of Architecture, Kwangwoon University, Korea

Abstract

As the problem of harmful factors affecting human body is issued, demands on pleasant and comfortable indoor environment has increased. It is increased an interest to manage pollutants emitted from construction materials and to use a certificated eco-materials, and the requirement of eco-friendly certification has been expanded gradually. Studies on the construction material database, accordingly, to manage effectively information on construction material have been performed constantly. The opened material information on the web, currently, is mainly about certificated eco-materials characteristic and, a website for managing information has been established. This information, however, is expired as the termination of validation period and new information is generated regularly. In order to utilize this information in eco-friendly certification system, a database of automatically collecting, storing, and managing needs to establish. This study, therefore, proposes a process to generate information of eco-friendly material using the Web Crawling technique capable of extracting material information automatically from the Web. It is expected to be used as a certification information in eco-friendly certification system.

Keywords: Web Crawling, Eco- Material Database, Automation of Data Extracting

Contact Author: Jungho Yu, Professor, Architecture Engineering, Kwangwoon University, 521-1 Hwado-gwan, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
e-mail: myazure@kw.ac.kr

1. Introduction

The extension of indoor activity increased demand on comfortable and safe indoor environments, leading to the establishment of eco-friendly policies such as health-friendly construction standard for housing and Green Standard for Energy & Environmental Design (G-SEED). The eco-friendly certification system specifies the management of harmful pollutants emitted from construction material and this prompted interests in using eco-friendly material and pollutants emitted from construction materials. The government, accordingly, implemented construction material certification system and released the information on the certificated materials, and the generation of construction material database for more effective management and use of information on construction materials has been studied constantly.

The currently available certification system on construction materials include Environmental Declaration of Product (EDP), Korea industrial Standard (KS), Good Recycled (GR), and Carbon Footprint of Product (CFP) ones, and the information on these eco-friendly certificated products are managed by Green Construction Materials Information System(http://gmc.greenproduct.go.kr/) supervised by Korea Environmental Industry & Technology Institute (KEITI). The users who are seeking some information from this web site, however, should review the pages one by one, making the procedure to be inconvenient, and the periodical management of information is needed to reflect expired or newly generated ones. This means that a database capable of collecting and managing information is needed for the constant management and use of desired information.

This study, therefore, suggest an algorithm for generating information on eco-friendly material by applying data crawling technique on the structured web page. The crawling technique allow users to extract desired information on construction material to be used in eco-friendly certification system by applying web scraping, a Web Crawling technique that extract information from structured pages, to eco-friendly construction material information system with fixed page structure. This technique classifies the information to be extracted by page and present an extraction algorithm by pages. The extracted information is stored in database and managed by assigning period of validation by each certification agency. The database is expected to be used for items related material information contained in eco-friendly certification system.

2. Preliminary Review

2.1 Web Crawling

The Web Crawling is a process that collects information from web by using Uniform Resource

Locator (URL) as a Seed USL. The Web Crawlers, a program that performs this work, are classified in large into General Web Crawler Server that collects information from seed URL as a starting point and along the hyperlinks, and Distributed Web Crawler that collects information in client environment [4]. The General Web Crawler is used more widely and the process of this Crawler is called Web Scraping, which extracts mainly formalized information, such as text from table or others, among ones written in Hypertext Markup Language (HTML) in Web page. The stored information is processed into the desired form and restored.

2.2 Literature Review of Generating a Web-based Eco-Material Database

The previous studies on Web-based eco-friendly Database (DB) are classified into three categories in large: First, some studies provide classification system to manage construction material information and measures for establishing DB. They proposed items of information needed to utilize materials based on classification system and build DB reflecting the items. Second, some studies developed information system assisting users to search material. They developed information classification system. They developed a system or application that builds information classification system displaying material information according to user's convenience. Third, some studies developed DB that manage material information and that propose method to collect information and information classification system and DB related to specific processes.

Table 1. The Literature Review of Eco-Material Database

Topic	Author(year)	Contents
Method of establishing a database	Kang, D. H. et al. (2006)	Propose a classification and a database for eco-friendly construction materials.
	R. Ozao et al. (2007)	Analyze the information required for materials and build a web page for managing this information.
	Yu, Y. J. et al., (2009)	Establish a database on emission of pollutants according to the classification system used by Ministry of Environment and Korea Air Cleaning Association.
Development of building materials information system to help selecting	M. N. Jadid (2013)	Develop a web page assisting users to search desired materials by web-based system
	Jun, H. J. (2017)	Analyze a material information provided and requested information, and develop an indoor construction material information application used for construction design, by using pattern language
Management of material information related to a specific trade.	Mun, J. Y. et al., (2017)	Develop a web-based database to manage information on construction material when deconstructing buildings by attaching RFID system to construction material.

The most of previous studies focused on proposing information classification system, DB, and system to manage material information, necessitating the additional studies on the method to collect information. This study that proposes processes related to method to collect information, therefore, has uniqueness.

3. Material information creation process using Web crawling technology

The demanded information on material based on the green building certification among the eco-friendly certification system is material name, product name, contact information, and website. The information on material name, product name type and amount of contaminant substance, certification period of validation, and company name can be extracted from the first page of eco-friendly construction material information system and company contact information and website information from second page (Fig. 1). The crawling is performed on each page separately.

Fig. 1 The Web Page to Extracting Information

The crawling is performed on first and second pages of construction material information system separately by using the pages as a seed URLs. The commands performed on the first page to extract information are coded by Python. The eligible information is stored and remaining information is searched by an algorithm for

second page. The unidentified information on the two pages are searched manually.

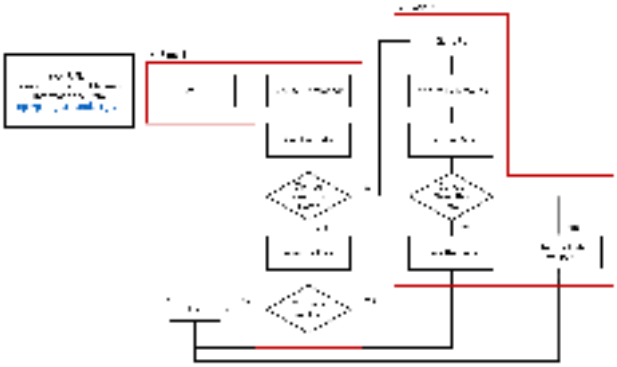

Fig. 2 The Process of Data Crawling

4. Conclusion

This study proposed an algorithm to extract material information from text that are formalized information based on structured pages. The extracted data is compiled as DB and is expected used in eco-friendly certification process and these information can be used to generate certification information of green building certification system related to indoor air qualities. The studies on information extraction from unstructured pages and use them on eco-friendly certification system and apply the methods to real cases are planned

Acknowledgement

This research was supported by a grant(16CTAP-C114926-01) from Infrastructure and transportation technology promotion Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

References

1) Kang, D. H., Yeo, M. S., Kim, K. W. (2006) The Method of Classification of Eco-Friendly Material and Establishing a Database, Magazine of the Society of Air-Conditioning and Refeigerating Engineers of Korea(SAREK), 35(12), 33-40
2) R. Ozao, T. Sawaguchi, H. Ishida, M. Iji, T. Furuyama, Y. Shinohara, K. Yamada, K. Halada (2007) Eco-MCPS : a Newly Developed Web-Based Database for Eco-Matrials in Japan, Materials Transactions, 48(12), 3042-3049
3) Yu, Y. J., Lee, C. W., Kim, M. G. (2009) Development of a Building materials database; Volatile organic compounds, formaldehyde emission rates and chemical compositions, Analytical Science & Technology, 22(1), 57-64
4) Seo, D. M., Jung, H. M. (2013) Intelligent Web Crawler for Supporting Big Data Analysis Services, The Journal of the Korea Contents Society, 13(12), 575-584Kasai, K. and Yashiro, T. (2001) Elicitation of subjective probabilities for risk analysis. Journal of Asian Architecture and Building Engineering, 1 (1), 77-82
5) Mansour N. J. (2013) Development of a Web-Based Decision Support System for Materials Selection in Construction Engineering, International Journal of Civil Engineering and Technology(IJCIET), 4(2), 177-188
6) Mun, J. Y., Hwang S. S. (2017) Construction Materials Management System Based on Web Database, Journal of the KIECS, 12(1), 195-200
7) Jun, H. J. (2017) Development an Eco-Friendly Interior Architecture Material Application Using Pattern Language, MS thesis, KunKook University