Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P083-cd

Using Web Crawling for Automated Country Risk Assessment

Beste Ozyurt

Department of Civil Engineering, Middle East Technical University, Turkey. besteozyurt@gmail.com

Joseph H. M. Tah

School of the Built Environment, Oxford Brookes University, UK. jtah@brookes.ac.uk

Irem Dikmen

School of Construction Management and Engineering, University of Reading, UK. i.dikmen@reading.ac.uk

M. Talat. Birgonul

Department of Civil Engineering, Middle East Technical University, Turkey. birgonul@metu.edu.tr

The latest and most accurate host country information is crucial for international market selection and bidding decisions. In today's digital world, this information can be timely gathered from websites on the Internet, the globally connected network system facilitating worldwide access to a wide range of information resources through a massive collection of private, public, academic, and government networks. Within construction companies, the traditional risk assessment process is generally based on searching for country information on the web and necessitates human input, which is subjective by nature and prone to misinterpretation and errors. With the help of automation, the traditional way of identifying and assessing country risks can change for the better. Better, latest, and most accurate data collection with web crawling algorithms, which are used for automated browsing, may be possible for country risk assessment with less human involvement, minimizing human errors and saving effort and time. Studies that explored the automated collection of information from the Internet for country risk assessment in construction are limited in the literature. In this paper, we assert that the web crawling technique can help search for data from the web and can be used to automatically develop country risk registers. Within this context, first of all, the web crawling algorithms will be explained, and how they can be used to automate the risk assessment process will be discussed. Then, a demonstrative example of an international construction company carrying out a country risk assessment process while preparing a bid for a highway project in Serbia will be depicted. The benefits of web crawling for automated country risk assessment will be discussed as well as its challenges.

Keywords: Automation, country risk assessment, international construction, web crawling.

1. Introduction

Acquiring precise, timely, and accurate information regarding the host country is crucial to conducting effective country risk assessments of international construction projects. The Internet, a globally connected network system that enables worldwide access, provides a wide range of information resources through a diverse collection of private, public, business, academic, and government networks. These resources, including news, articles, research papers, interviews, reports, and official documents, are valuable for gathering country information for an effective risk assessment (Moon et al., 2018; Shin, 2015; Pathirage et al., 2007).

The rapid expansion/growth of the Internet has contributed to the accumulation of country data, which is continuing to increase (Shin, 2015; Hjelt and Björk, 2006). However, acquiring timely, precise, and accurate information about host countries can be challenging and time-consuming, requiring qualitative and quantitative assessments that depend on numerous factors and may involve mental shortcuts, heuristics, or biases (Moon et al., 2018).

Despite the promising/encouraging outcomes of web crawling, one of the digital technologies, previous approaches to automating information extraction and retrieval, the use of web crawling specifically for country risk assessment has not received enough attention. An automated process for gathering data from the Internet (i.e., the utilization of web crawling) can aid in reducing errors, enhancing the identification of country risk, and facilitating effective country risk assessment. However, there is a lack of research on the utilization of web crawling for the purpose of country risk assessment.

As a result, this research suggests web crawling as a useful technique for enhancing the country's risk assessment process, and provides insights on how to utilize web crawling for this purpose effectively. To demonstrate how web crawling automatically obtains country-specific data, an example of an international construction company carrying out a country risk assessment process while preparing a bid for a highway project in Serbia was provided.

2. Web crawler

A web crawler, or a robot or spider, is a program that automatically collects and stores content from web pages on the Internet. It searches for relevant information on a specific topic by visiting web pages on demand while keeping indexes (Pinkerton, 1994; Cho, 2001; Olston and Najork, 2010). The crawler receives a prioritized set of URLs (Uniform Resource Locators) to visit and collects hyperlinks from those pages. It iteratively downloads the content of the web pages until the specified or predefined criteria are met, such as a certain number of pages visited, a particular keyword or phrase found, or a specific amount of time elapsed (Moon et al., 2018; Cho, 2001; Olston and Najork, 2010).

Since the 1990s, web crawlers have been used in various ways, particularly as a web search engine. Even though there has been an extensive amount of research on web crawling in general, there have been a few studies conducted specifically on web crawling in construction.

2.1. *Related Work*

To date, there has been limited attention on utilizing web crawling in the construction area. Furthermore, few studies have investigated web crawling as a means of identifying country risks to assist in the assessment of country risks in international construction. Based on the current literature, web crawling algorithms have been used for construction document management systems (Moon et al., 2018), for fire accidents (Kim et al., 2021), for sustainability issues (Hong et al., 2019), for prefabricated construction industry (Dou et al., 2019), and building code (Zhou et al., 2021).

This study explores the potential of web crawling for assessing country risks, considering information about the global market. A demonstrative example is illustrated below to explain how web crawling automatically obtains country-related data.

2.2. *Demonstrative Example*

The example of an international construction company conducting a country risk assessment while preparing a bid for a highway project in Serbia will be used to demonstrate how web crawling can be used for country risk assessment. It is assumed that company professionals are willing to explore country conditions to understand how country-related factors may affect project success or failure. Therefore, they need to assess the identified country risks, taking into account Serbia-specific information.

In this example, company professionals will utilize "web crawling" to obtain country data. Contrary to the traditional method, company professionals will not be browsing through all the relevant websites one by one to find specific information, but will be able to obtain the received information in an orderly spreadsheet. Therefore, they will both save time and make fewer mistakes.

If the risk managers plan to use web crawling for risk assessment, first they would need to identify the necessary country information and determine the related websites as data sources accordingly. Then, they would specify the related risk types for each extracted information, and finally, the rules (e.g., threshold values for each risk level) that would be needed for the assessment process would be assigned. This process can be summarized below in Fig. 1.

This example used Beautiful Soup, a popular Python library for pulling data from HTML (HyperText Markup Language) (Hajba, 2018). Nine data sources were selected for demonstrative purposes. XPath expressions were used in this example to find any element in an HTML DOM page easily.

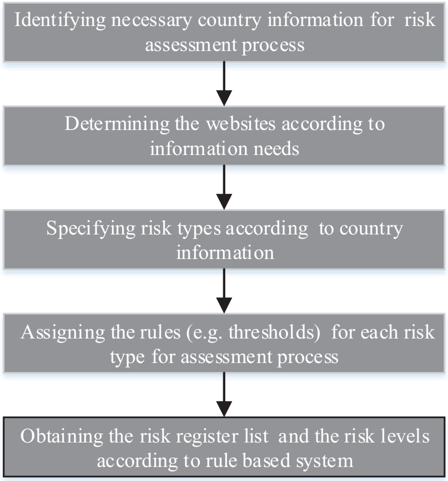
Identifying necessary country information for risk assessment process
Determining the websites according to information needs
Specifying risk types according to country information
Assigning the rules (e.g. thresholds) for each risk type for assessment process
Obtaining the risk register list and the risk levels according to rule based system

Fig. 1. Process chart

In this example, country risk-related information was planned to be extracted from the selected data sources. First, the selected country-related web pages and their URLs were determined by investigating the web page's structure (in HTML). Then, by inputting URLs and XPaths, the programmed web crawler accessed each data page, parsed and processed the webpage to "crawl" the target data needed, and then collected relevant data from the page. Finally, the data was saved in a CSV (comma separated values) format to be more readable in Microsoft Excel, and a tabular output was generated.

Each website's data extraction date ranged from 21 March 2022 to 29 April 2023. The output table of Serbia's example, with the websites (with their URLs) as data sources and the extracted values as corresponding outputs, can be seen below in Table 1. Note that it is also important to indicate the date the data was obtained.

Table 1. Output of Web Crawling Example

WEBSITES AND INDICATORS	**VALUE**
IMF - imf.org/external/datamapper/profile/SRB	
Real GDP growth (Annual percent change)	2
Current account balance, percent of GDP	-6.1
General government gross debt (Percent of GDP)	50.2
WORLDBANK data.worldbank.org/country/serbia	
Poverty headcount ratio at $2.15 a day (% of population)	1.6
Literacy rate, adult total (% of people ages 15 and above)	99
BIS - Bank for International Settlements stats.bis.org	
US dollar exchange rates. End of period	107559,00
ILOSTAT - ilo.org/shinyapps/bulkexplorer	
Labor force participation rate (%)	57.3
Trade union density rate (%)	33.3
Population covered by at least one social protection benefit (%)	48,00
CoFace - coface.com.au	
Budget balance (% GDP)	-2,70
Business climate	A4
National Bank of Serbia nbs.rs/en/indeks/index.html	
Dinar exchange rate	117.2758
NBS interest rates	6.00%
Inflation in March	16.2%
Standard and Poor's Credit Rating	BB+ stable outlook
Statistical Office of the Republic of Serbia stat.gov.rs/en-US	
Employment rate	50.1 %
Unemployment rate	9.2 %
BTI - The Transformation Index -bti-project.org	
Population growth	-0.5 % p.a.
Poverty	8.9 %
Global Edge - globaledge.msu.edu/countries	
Ease of Doing Business Rank	75.7 (44 out of 189)
Corruption Perceptions Index	38
Index of Economic Freedom	63.9
Management Index (Political Leadership Towards Democracy and a Market Economy)	5.39
Status Index (Political and Economic Transformation)	6.94

3. Discussions

The hypothetical demonstrative example shows that country information can be obtained from related web pages efficiently and faster to be used in the country assessment process.

In order to utilize web crawling data in an actual construction project, a country risk taxonomy should be established, followed by identifying corresponding country factors related to these risks comprehensively. Then, web sources containing data about these indicators should be determined. Using inflation and interest rate data to assess "economic risk", and poverty-related data to determine "social risk" can be given as examples of indicators and risk types. After all of these steps, web crawling can be automated to obtain information and assist in risk assessment for a construction company.

As one step further in research, the obtained country data (the output of the web crawling process) can be evaluated in a rule-based system. As a social risk assessment example, if the "poverty rate" is less than 0.1, the system could determine it as "low" risk, whereas if it is higher than 0.3, the system could assess it as "high" risk. Similarly, if the "unemployment rate" is less than 0.15, the system could evaluate it as a "low" risk, whereas if it is higher than 0.35, the system could assess it as a "high" risk. Therefore, according to these hypothetical threshold values (0.1 and 0.3 for poverty; 0.15 and 0.35 for unemployment rate), the output of a basic rule-based system could be below (in Table 2), gathering the Serbia values from the web crawling example. This output could also have been created with color codes (Red for high level; Green for low level) to be more visual.

Table 2. Simple Output of Rule-Based System Example

Social Risk related Indicators	**Country related Values**	**Risk Level**
Poverty	8.9 %	Low
Unemployment rate	9.2 %	Low

This simple rule-based example regarding country risk assessment can show another usage of web crawling outputs in case the threshold values are defined. However, further research is necessary to explore the potential of web crawling for risk management extensively.

4. Conclusions

The collection of recent and accurate information is essential for country risk assessment in international construction projects due to the constantly changing economic, social, and political environment. However, automated data collection through web crawling has not been extensively investigated for this purpose in the construction management literature.

Construction professionals can efficiently and systematically extract, collect, sort, and analyze useful information by implementing automated web crawling processes. This approach can also be updated and utilized repeatedly throughout the construction phases. Instead of manually gathering country data, web crawling can be a more efficient and faster alternative, allowing planners and managers to review and examine web pages as part of the country risk assessment and enhance project management processes.

Despite its advantages, web crawling presents challenges and concerns, such as managing large data structures, deciding which web pages to use as data sources, and revisiting web page content sources. In addition, it may be difficult to design a robust web crawler due to the massive volume of continuously expanding data on the web.

This paper used a demonstrative example to illustrate how simple web crawling can be utilized for country risk assessment. However, more comprehensive implementations are possible, and additional research should be conducted on how web crawling may facilitate risk-informed decision-making.

In this study, copyright issues were considered, and future web crawlers should be aware of selecting websites and target data that adhere to copyright regulations. While human input was necessary in this study for determining the required details of the demonstrative example, more automated utilization of web crawlers could be common in the future.

By leveraging the capabilities of Artificial Intelligence (AI), researchers can significantly enhance web crawling applications, experiencing improvements in speed, accuracy, and effectiveness. Recently, AI-powered large language models have revolutionized the field of web crawling by automating and enhancing the process. These advanced models, with their sophisticated natural language processing capabilities, serve as ideal tools for web crawling tasks. Through automation, they offer substantial time and effort savings, increased efficiency, and the elimination of drawbacks. In the future, renowned and prominent AI chatbots such as ChatGPT, Bing Chat, and Google Bard may find increased usage in country risk assessment and other research domains. Researchers will explore further integration of AI-powered web crawls and investigate the specifics of these intelligent systems in their further studies.

References

Cho, J. (2001). Crawling the web: discovery and maintenance of large-scale web data (Doctoral Dissertation). Department of Computer Science, Stanford University.

Dou, Y., Xue, X., Wang, Y., Luo, X., and Shang, S. (2019). New media data-driven measurement for the development level of prefabricated construction in China. Journal of Cleaner Production, 241, 118353.

Hajba, G.L. (2018). Using Beautiful Soup. In: Website Scraping with Python. Apress, Berkeley, California.

Hong, S. H., Lee, S. K., and Yu, J. H. (2019). Automated management of green building material information using web crawling and ontology. Automation in Construction, 102, 230-244.

Kim, J., Youm, S., Shan, Y., and Kim, J. (2021). Analysis of fire accident factors on construction sites using web crawling and deep learning approach. Sustainability, 13(21), 11694.

Moon, S., Shin, Y., Hwang, B. G., and Chi, S. (2018). Document management system using text mining for information acquisition of international construction. KSCE Journal of Civil Engineering, 22(12), 4791-4798.

Olston, C., and Najork, M. (2010). Web crawling. Foundations and Trends in Information Retrieval, 4(3), 175-246.

Pathirage, C. P., Amaratunga, D. G., and Haigh, R. P. (2007). Tacit knowledge and organisational performance: Construction industry perspective. Journal of Knowledge Management, 11(1), 115-126.

Pinkerton B. (1994). Finding what people want: Experiences with the web crawler, Proceedings of the Second World-Wide Web Conference, Chicago, Illinois.

Shin, Y. (2015). Designing a System Prototype for Construction Document Management Using Automated Tagging and Visualization (Doctoral Dissertation). Department of Civil and Environmental Engineering, The Graduate School, Seoul National University.

Zhou, Y. C., Lin, J. R., and She, Z. T. (2021). Automatic Construction of Building Code Graph for Regulation Intelligence. In International Conference on Construction and Real Estate Management (ICCREM 2021), Beijing, China, 248-254.