Using Crawlers for Targeted Data Extraction from a Local Multi-File Database

Alexandru ANGHELESCU[a,*], Ciprian-Viorel STUPINEAN[a,b], and Ariana-Anamaria CORDOŞ[a,c]

[a] Romanian Society of Medical Informatics, Mihai Viteazul Blvd., no. 1, 300222 Timişoara, Romania.
[b] Faculty of Mathematics and Computer Science, Mihail Kogalniceanu Str., no. 1, Babeş-Bolyai University, 400084 Cluj-Napoca, Romania.
[c] Department of Public Health, Faculty of Political, Administrative and Communication Sciences, Pandurilor Str., no. 7, Babeş-Bolyai University, 400084 Cluj-Napoca, Romania.
E-mails: alexandru.anghelescu03@e-uvt.ro; ciprian.stupinean@ubbcluj.ro; ariana.cordos@ubbcluj.ro

* Author to whom correspondence should be addressed;

Abstract

Background: A crawler is a software program used to extract data in an automated manner. This study aimed to demonstrate how a crawler can extract specific data from multiple diverse .pdf files. *Methods and Materials*: To achieve this objective, a C# .NET9 application was developed, capable of processing a folder (local database) containing specific .pdf files. The application sequentially read each file and extracted relevant information. The ability of the crawler to extract the e-mail addresses of corresponding authors from academic papers was evaluated as a .pdf file may have contained multiple articles. In addition to email addresses, names of corresponding authors were extracted where possible. The PdfPig library was used to access the data since the input data were .pdf files. The output consisted of a CSV file containing all extracted email addresses. The input dataset included 19 books of abstracts and 180 articles. *Results*: During testing, the application managed to extract 929 email addresses and 77 names. However, due to pattern inconsistencies, name extraction was possible only for articles, not for books of abstracts. Further, evaluation on precision and accuracy was performed. While there was only 1 line extracted that did not contain emails out of the 880 lines, 34.55% of them needed corrective actions. In 213 instances text was attached to the e-mails (e.g. country names: Spain, Israel etc. or other words like keyword or abstract), country prefixes were attached in 156 cases and in 2 lines there additional full stops at the beginning or end of the e-mail. *Discussion*: Crawlers can be effective in extracting specific data from big databases of files simultaneously. In medical research, this ability can have an impact on productivity when dealing with data collection for research purposes. On the other hand, it poses a risk when personal information, e-mails in this case, become accessible for malicious purposes. Future work should explore compliance with data protection regulations, such as GDPR, and methods to ensure responsible data use. *Conclusion*: Besides the usefulness of crawlers in extracting email addresses, they prove their efficiency while dealing with the data gathering part of the research. By using a crawler, the researcher may be able to save some time, just by not dealing with the data extraction part of the study, while dealing with a large database of studies to cite.

Keywords: Crawler; C#; Automated Data Extraction; Email; Data.