

(12) PATENT APPLICATION PUBLICATION

(21) Application No.202211057644 A

(19) INDIA

(22) Date of filing of Application :08/10/2022

(43) Publication Date : 04/11/2022

(54) Title of the invention : SYSTEM AND METHOD FOR WEB CRAWLERS

(51) International classification :G06F0016951000, G06F0016955000, H04L0067020000, G06F0040143000, H04N0021858000
(86) International Application No :NA
Filing Date :NA
(87) International Publication No : NA
(61) Patent of Addition to Application Number :NA
Filing Date :NA
(62) Divisional to Application Number :NA
Filing Date :NA

(71)Name of Applicant :

1)Chitkara University

Address of Applicant :Chitkara University, Chandigarh-Patiala National Highway, Village Jhansla, Rajpura, Punjab - 140401, India. Patiala -----

2)Bluest Mettle Solutions Private Limited

Name of Applicant : NA

Address of Applicant : NA

(72)Name of Inventor :

1)MANTRI, Archana

Address of Applicant :Chitkara University, Chandigarh-Patiala National Highway, Village Jhansla, Rajpura, Punjab - 140401, India. Patiala -----

2)MISHRA, Rahul

Address of Applicant :ODC-4, Panchshil Tech Park, inside Courtyard by Marriott premises, Hinjewadi Phase - 1, Pune - 411057, Maharashtra, India. Pune -----

3)SINGH, Dhiraj

Address of Applicant :ODC-4, Panchshil Tech Park, inside Courtyard by Marriott premises, Hinjewadi Phase - 1, Pune - 411057, Maharashtra, India. Pune -----

(57) Abstract :

The present disclosure relates to a system (100) for determining web crawlers. The system (100) comprises network (102), page downloader (102), database (106), URL extractor (108), queue (110), and scheduler (112). Page downloader (102) configured to download at least one web page. Database (106) configured to store at least one of text and metadata of at least one web page. URL extractor (108) configured to extract at least one URL of the at least one web page, where initial URL of at least one downloaded web page is retrieved and associated with degree entry. Queue (110) configured to arrange extracted URLs in series which are coupled to initial URL. Scheduler (112) configured to looping the at least one extracted URLs available in queue (110). The at least one extracted URL is provided to page downloader (102) to crawl and extract at least one corresponding web page through the network (102).

No. of Pages : 26 No. of Claims : 10