INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

Scalable and Extensible Web Crawler

ABSTRACT

A Web Crawler is a program, which is commonly used by search engines to find the new brainchild on the internet. The use of crawlers has made the web easier for users. Obtaining Deep Web content is challenging and has been acknowledged as a significant gap in search engines' coverage. Deep web refers to the hidden part of the web that remains unavailable for standard Web crawlers.

OBJECTIVE

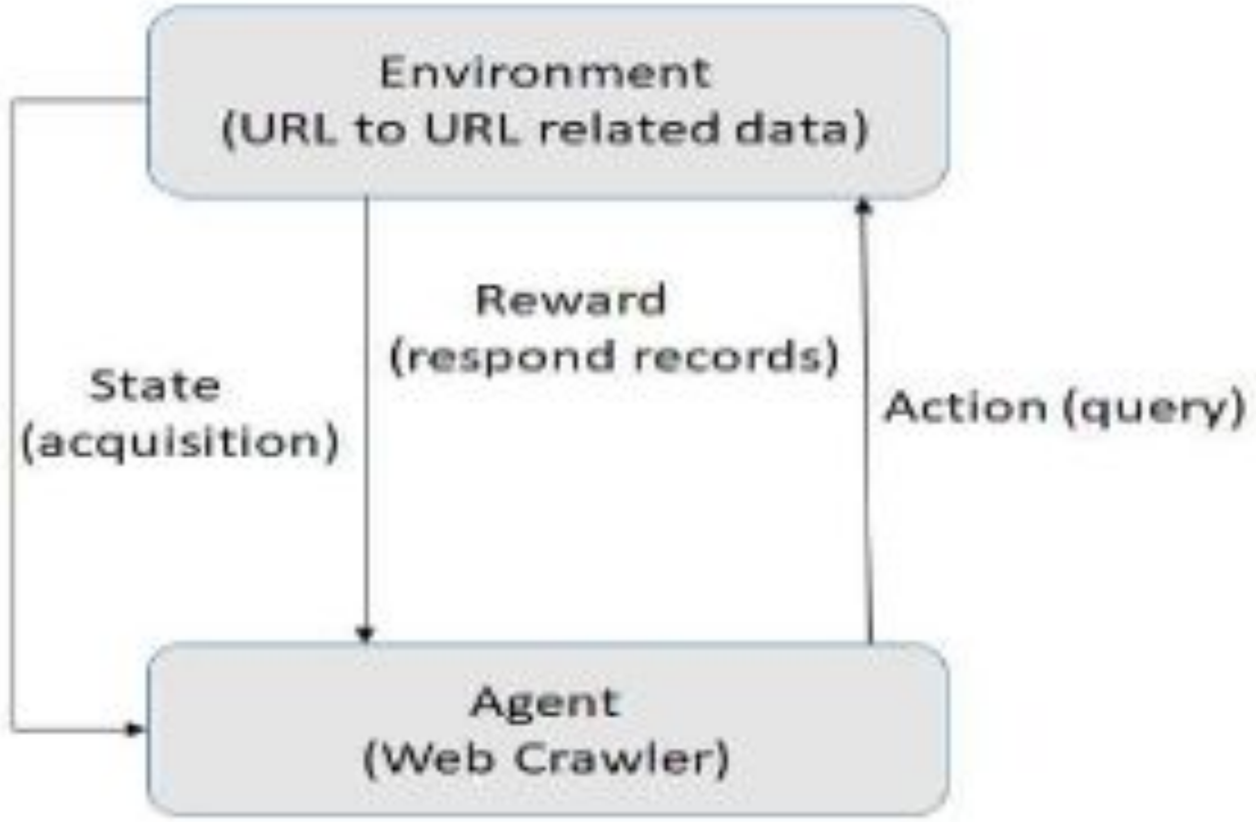
Purpose: Lack of Scalable and Extensible Crawlers makes it difficult to adopt existing Open Source Crawlers for building Search Engines.

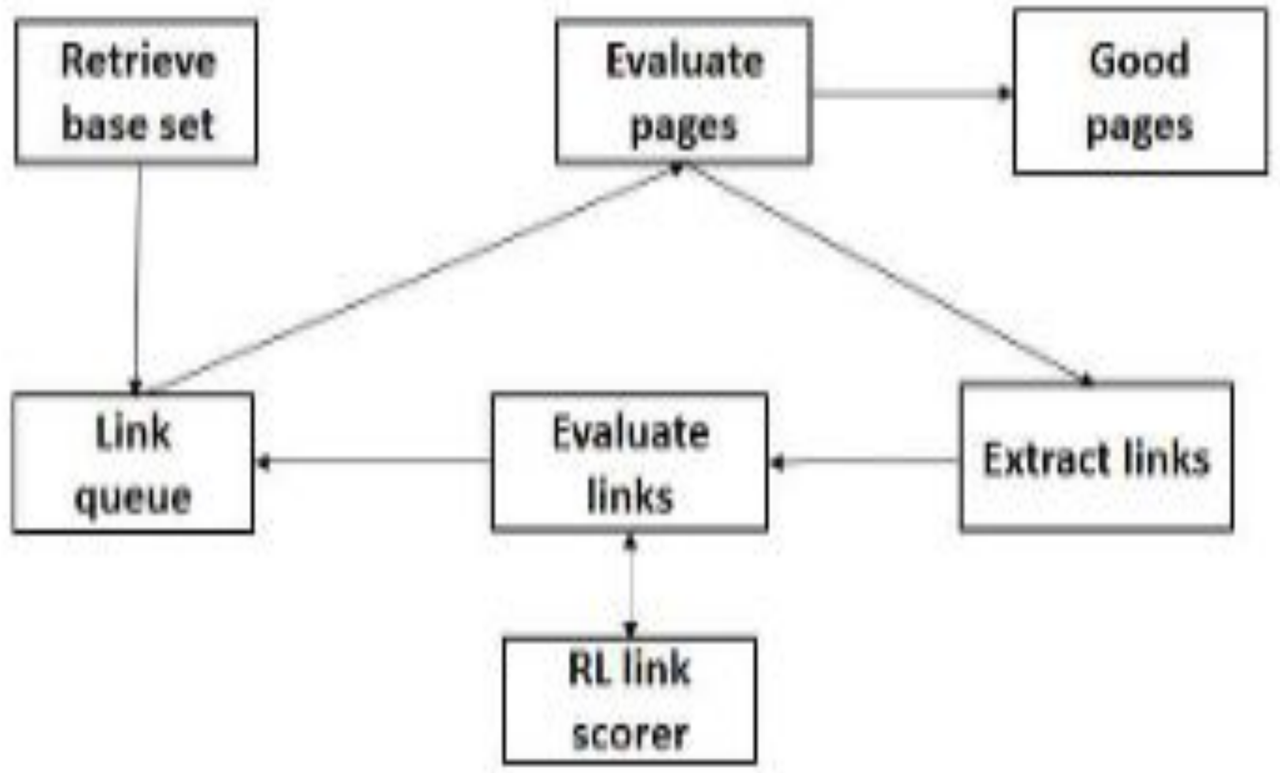
Research:

1. Identifying an approach and building extensible Web Crawler that extracts content from publicly available web content and extends to deep web and dark web.
2. Identifying an approach and Building scalable Web Crawler to extract large volume of content.

METHOD

1. The crawler is regarded as an agent and a deep web database as the environment.
2. According to Q-value, the agent perceives its current state and selects an action (query) to submit to the environment.
3. The framework enables crawlers to learn a promising crawling strategy from their own experience and utilizes various query keywords.





Authors: Dontineni Ganesh Sai, Lalit Mohan S, Y Raghu Reddy

Research Center Name: Software Engineering Research Center (SERC)