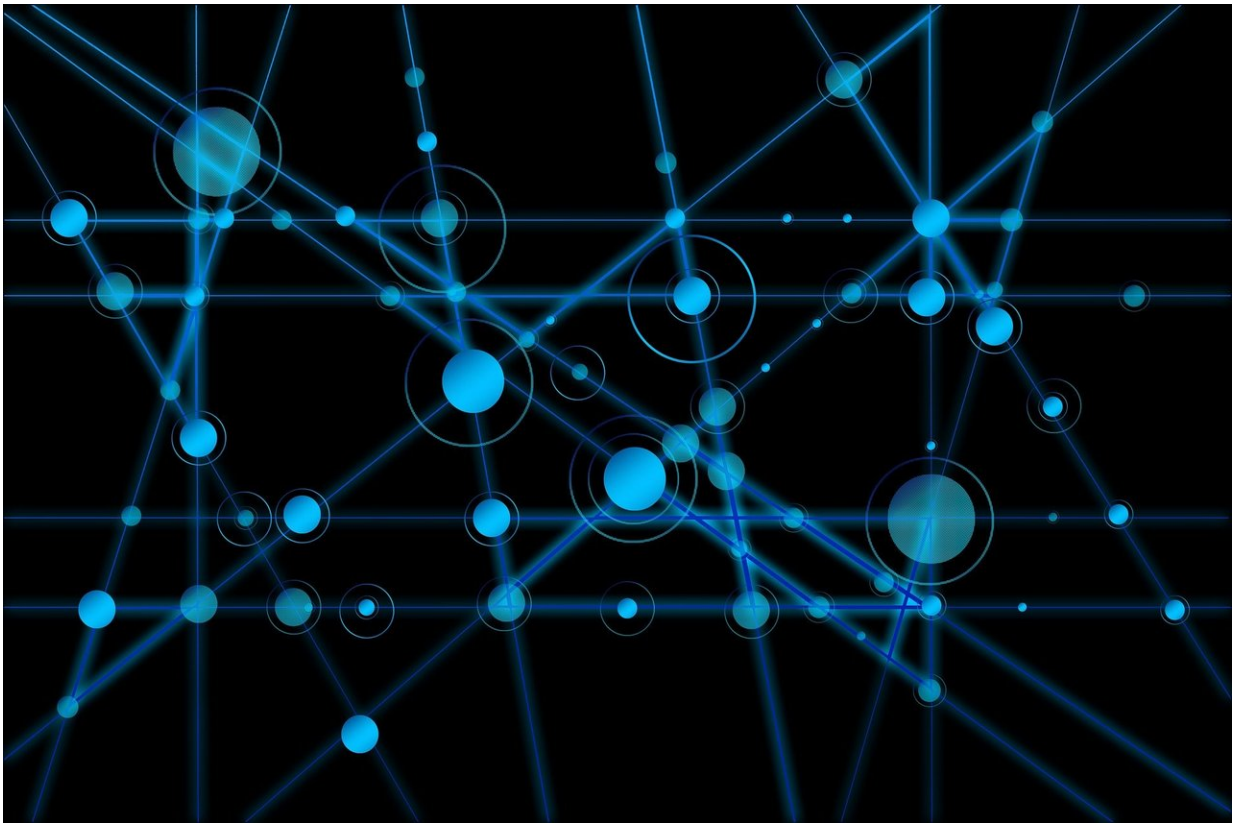
Crawling the invisible web genetically

February 6 2020, by David Bradley



Credit: CC0 Public Domain

The world-wide web has grown immensely since its academic and research inception in 1991, and its subsequent expansion into the public and commercial domains. Initially, it was a network of hyperlinked pages and other digital resources. Very early on, it became obvious that some resources were so vast that it would make more sense to generate

the materials required by individual users dynamically rather than storing every single digital entity as a unique item.

Today, countless websites are dynamic, every unique visit draws information and data dynamically from a back-end database and presents it to the user on-demand. Whereas static pages can easily be spidered by search engines, database content that drives dynamic websites is inaccessible. Even as long ago as 2001 when there were already several terabytes of public, static web data, it was estimated that the "invisible web," or "hidden web," not to be confused with the "dark web," was some 550 times bigger than the visible resources.

Writing in the International Journal of Business Intelligence and Data Mining, a team from India describes how they have developed a genetic algorithm-based intelligent multiagent architecture that can extract information from the invisible web. The tools could allow even materials that are purportedly off-limits to conventional search engines to be spidered, scraped, and cataloged for a wide range of applications.

D. Weslin of Bharathiar University and Joshva Devadas of Vellore Institute of Technology describe the details and benefits of their approach in the latest issue of the journal. "The experimental results show that the proposed architecture provides better precision and recall than the existing web crawlers," the team writes.

More information: D. Weslin et al. Genetic algorithm-based intelligent multiagent architecture for extracting information from hidden web databases, *International Journal of Business Intelligence and Data Mining* (2020). DOI: 10.1504/IJBIDM.2020.104740

Provided by Inderscience