Kennesaw State University

College of Computing and Software Engineering

Computer Science Department

CS 4720 – Internet Programming 03

Assignment 2, Paper 2

Grayson Beckman

[gbeckman@students.kennesaw.edu](mailto:gbeckman@students.kennesaw.edu)

September 15, 2019

Shannon Perry

Web Spiders That Crawl and Scrape: 38

Web Spiders That Crawl and Scrape

The modern world is built on information. From GPS, to Banking, and even friends arguing about who is the best quarterback in the NLF today, everything we do is based on information. Nowhere is that truer than in the search engines that consumers visit bullions of times per day. These sites, i.e. Google, Bing, etc. allow users to search through millions of different articles, images, sites, or anything imaginable with just the click of a button. Search engines do this by building lists of words found on Websites (Franklin, 2000) and then displaying them in an organized fashion for the user to see and choose from. This important process in the evolution of the internet is accomplished by deceptively simple pieces of software called spiders.

In the simplest terms’, spiders are scripts, or automated programs, that algorithmically build indexes of data from pages on the world wide web. A spider starts with a list of URLs’ that are determined by the creator of the spider and then scans all those pages and any other pages connected to them recursively and builds a set of all the data found on those sites. These programs are one of the critical factors in transforming search engines like google into companies worth hundreds of billions of dollars. Because of this, the exact algorithms that the spiders run by companies like google use to go through information is a closely guarded secret by those that own them. These programs are so valuable because they gather information into a single location that allow users to easily take advantage of the collective knowledge of what is effectively the entire world.

Although understanding the inner working of spiders is a monumental task that can potentially land a careless individual on the wrong side of the law, their impact can be grasped by even the most technologically illiterate and their value cannot be understated. Before the internet the only way to gather information on a given topic was to either be taught it by someone or to look it up in the local library, both of which have serious problem with efficiency and scalability. The fable of a family member having to embark on a long journey to gather vital information is found throughout history and demonstrates this point perfectly. It is rarely the case that an individual has easy access to experts on any given subject in their local, and somewhat isolated, community or that their local repository of information has a tome that details the secrets of a piece of vital but lesser known knowledge. Thus, the need to travel to a place where that knowledge has a better likelihood of being found, although in the end it is never a guarantee that it will be found. This is in complete contrast to the world we find ourselves in today, where any and all information can be discovered by a few dedicated hours searching the largest warehouse of knowledge ever devised, the internet. Through the use of spiders that diligently catalogue the trillions of pieces of information that can be found, anyone can learn anything that they set their mind to.

However, this technology is not without dangers or downsides. Like any tool, they can be used to help or hurt based solely upon who is wielding that tool. Spiders can be used to gather personal information on unwitting victims in order to sell it or to prey upon any weaknesses that are discovered. They can be used to slow down or lockup different services or sites because they can be designed to use up massive resources while they are traversing these pages. Potentially helping to cripple those that the perpetrator wish to harm. These and many other examples show how even the greatest of tools can be used for ill and that the skill to create such scripts is an enormous responsibility that cannot be taken lightly.

In the end, spiders are one of the most important tools being used on the internet today. They help categorize vast stores of information and put them at anyone’s fingertips, potentially helping to catapult humanity into its next great revolution. But that potential does not preclude the use of spiders for ill, and fighting against this malevolent use of these programs may be one of the next great fights that the technology world will have to face as it continues to grow in importance in the global economy and everyday life.

# References

Franklin, C. (2000, September 27). *How Internet Search Engines Work*. Retrieved from How Stuff Works: https://computer.howstuffworks.com/internet/basics/search-engine.htm