An Intermediate-Level Introduction to Web Mining

Kooper Young, Nikitha Vedant Madabhushi, Divya Shree Moka, Bharath Kumar Reddy Mopuru, Susmitha Muppalla, Manohar Sri Vikram Vattikuti

Northwest Missouri State University, Maryville MO 64468, USA
S537651@nwmissouri.edu

Abstract. As data scientists, the process of gathering data is not a task to be overlooked or underestimated. In the digital age, data resides mostly on the internet, but by its nature, the internet is dense and it can be difficult and time consuming to retrieve the correct data efficiently. Web Mining is quickly becoming one of the most efficient ways to gather information from the web. As useful as it is, most everyone in applicable fields should understand the basics. In this paper, we plan to do just that, introduce you to the basics of web mining. We will be beginning with an introduction to the topic as well as some definitions, followed by many real world and abstract examples, ranging from the simplest that anyone could achieve, to the greatest new techniques and methods. Specifically, we will be looking at the Open Information Extraction system, as well as the TextRunner system, both of which are focused on making web mining easier and more accessible, as well as discussing the place of AI and neural networks in web mining, and how they might impact the quickly approaching future of technology.

Keywords: *Web Mining* · *Data Acquisition* · *Web Scraping* · *O.I.E* · *Information Extraction* · *TextRunner*

1 Introduction

It is fair to say that by this point that nearly every person on the planet with an internet connection uses the web in some form each and every day. The internet is great, it provides the average person with more knowledge and information than they are ever likely to understand or even view. However, the internet isn't perfect. It can be slow, or sometimes even frustrating to navigate. For example: have you ever been on a website that had a layout that confused you, or perhaps links and buttons didn't take you where you expected them to? Or, have you ever tried looking for some piece of information that took you much longer than normal to find? Taking a step back, it is natural that such a massive repository of information like the internet would have these sorts of issues, but nevertheless, it can be really off putting to users. To analyze, record, and solve/fix these sorts of issues, we use a technique called web mining. By utilizing web mining, we can analyze the patterns and behaviors of users and use that information to make using sites and finding information easier and more intuitive.