International Journal of Information Retrieval Research
Volume 13 • Issue 1

Automated Selection of Web Form Text Field Values Based on Bayesian Inferences

Diksha Malhotra, Punjab Engineering College (Deemed), Chandigarh, India*
Rajesh Bhatia, Punjab Engineering College (Deemed), Chandigarh, India
Manish Kumar, Punjab Engineering College (Deemed), Chandigarh, India

ABSTRACT

The deep web is comprised of a large corpus of information hidden behind the searchable web interfaces. Accessing content through searchable interfaces is somehow a challenging task. One of the challenges in accessing the deep web is automatically filling the searchable web forms for retrieving the maximum number of records by a minimum number of submissions. The paper proposes a methodology to improve the existing method of getting informative data behind searchable forms by automatically submitting web forms. The form text field values are obtained through Bayesian inferences. Using Bayesian networks, the authors aim to infer the values of text fields using the existing values in the label value set (LVS) table. Various experiments have been conducted to measure the accuracy and computation time taken by the proposed value selection method. It proves to be highly accurate and takes less computation time than the existing term frequency-inverse document frequency (TF-IDF) method, hence increasing the performance of the crawler.

KEYWORDS

Automatic Form Filling, Bayesian Inference, Deep Web, Hidden Web, Information Retrieval, Instance Templates

1. INTRODUCTION

The content on Internet is growing at a breakneck pace, as more and more people are connecting to it. Publically Indexable Web (PIW) or surface web consists of a very small part of the Internet, which can be accessed by traversing through hyperlinks. Traditional web crawlers use different approaches to access only PIW. Whereas, its counterpart, hidden web, consists of information generated dynamically where the user needs to fill a searchable form to access data. However, a number of recent studies have shown that a significant amount of data lies outside the PIW. A commercial vendor, BrightPlanet.com, claims that the size of the deep web is 500 times greater than the publically indexable web (Bergman 2001). The hidden web data is very important for various stakeholders. Hence, deep web crawlers use numerous approaches to access hidden web data. However, the deep web can be entered only after filling the search forms and hereby, accessing databases. Whenever a user fills up a search form in

DOI: 10.4018/IJIRR.318399 *Corresponding Author

order to access hidden data, a dynamic webpage is generated. A query is shot to the database and the required results from the database are shown. The results from the database may contain diverse content types such as *Dynamic Data, Unlinked Content, and Non-Text Content.* Dynamic data can only be accessed through the supported query interfaces. These interfaces consist of input elements, and a user query includes providing values for these elements. However, unlinked content cannot be accessed by going through links, and non-Text content consists of various PDF, multimedia files, and non-HTML documents.Following are the main four key phases of the working of deep web crawler:

- Discovery of the entry points to the hidden web *i.e.,* searchable interfaces as these allow searching online databases (Lage et al. 2004; Onihunwa et al. 2017).
- Label extraction (Wang and Lochovsky 2003; Nguyen et al. 2008).
- Updating the LVS table and automatically filling the hidden web forms.
- Response analysis *i.e.* classification into valid and invalid responses.

Figure 1 shows the above-explained key phases of the working of a deep web crawler. Each phase has its approaches and challenges associated with it. While designing a deep web crawler, a designer can face the following challenges:

- *Determining the searchable interface of hidden web* (Wu et al. 2006; Moraes et al. 2013; Liu and Li 2016): As the hidden web crawler needs a searchable query form in order to access a hidden web page, hence, a hidden web crawler must be able to identify the query forms as an entry point to the hidden database.
- Extracting form labels(An et al. 2007; Nguyen et al. 2008): As the labels of a form are not at a specified position in web forms; hence, it is a challenging task to extract form labels. The form labels help to fill form fields automatically.
- *Automatically filling forms*: It requires filling form fields with efficient and most suitable words for the field.

Considering the above challenges, the paper focuses on the challenge associated with process of automatic form filling (Álvarez et al. 2007). It requires the selection of appropriate values for the form fields so that with a minimum number of submissions, maximum records of data can be extracted. In order to assist in values selection, the paper focuses on the automatic filling of searchable web forms (excluding login forms) by generating informative instance templates (explained in the following sections) using fields of the form and selecting values for the fields using Bayesian inferences. The Bayesian inferences provide an automatic and effective way to help filling the searchable forms by creating a network structure, and calculating the joint probability.

The remainder of this paper is organized as follows. In Section 2, we present a brief literature review of the techniques in the existing literature for automatic filling of web form. Section 3 presents the terminologies related to the proposed approach of this paper. Section 4 describes the proposed approach for automatic form filling using Bayesian inferences in detail. Section 5 analyses the mathematical concepts related to the proposed approach, discusses the experimental setup and explains the result followed by conclusion.

2. RELATED WORK

In this section, we present a brief literature review of the existing techniques and solutions for Automatic filling of searchable interfaces. The automatic filling of web forms includes the automatic assignment of corresponding values to the fields and submission of web form to extract records from the hidden web.