A NOVEL METHOD TO EXTRACT THE STRUCTURED DATA FROM DEEP WEB PAGES

A. Revathi and Ch. Manisha

1II M.Tech. Student, 2Asst. Professor
Dept of CSE, Nimra Institute of Science and Technology, Nimra Nagar, Jupudi, Ibrahimpatnam-521456, Krishna Dt., A.P.

ABSTRACT: Deep Web contents are accessed by queries submitted to Web databases and the returned data records are enwrapped in dynamically generated Web pages (they will be called deep Web pages in this paper). Extracting structured data from deep Web pages is a challenging problem due to the underlying intricate structures of such pages. Until now, a large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. As the popular two-dimensional media, the contents on Web pages are always displayed regularly for users to browse. This motivates us to seek a different way for deep Web data extraction to overcome the limitations of previous works by utilizing some interesting common visual features on the deep Web pages. In this paper, a novel vision-based approach that is Web-page-programming-language-independent is proposed. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction. We also propose a new evaluation measure revision to capture the amount of human effort needed to produce perfect extraction. Our experiments on a large set of Web databases show that the proposed vision-based approach is highly effective for deep Web data extraction.

Keywords: Web mining, Web data extraction, visual features of deep Web pages, wrapper generation.

I. INTRODUCTION

The problem of Web data extraction has received a lot of attention in recent years and most of the proposed solutions are based on analyzing the HTML source code or the tag trees of the Web pages (see Section 2 for a review of these works). These solutions have the following main limitations: First, they are Web-page-programming-language dependent, or more precisely, HTML-dependent. As most Web pages are written in HTML, it is not surprising that all previous solutions are based on analyzing the HTML source code of Web pages. However, HTML itself is still evolving (from version 2.0 to the current version 4.01, and version 5.0 is being drafted [14]) and when new versions or new tags are introduced, the previous works will have to be amended repeatedly to adapt to new versions or new tags. Furthermore, HTML is no longer the exclusive Web page programming language, and other languages have been introduced, such as XHTML and XML (combined with XSLT and CSS). The previous solutions now face the following dilemma: should they be significantly revised or even abandoned? Or should other approaches be proposed to accommodate the new languages? Second, they are incapable of handling the ever-increasing complexity of HTML source code of Web pages. Most previous works have not considered the scripts, such as JavaScript and CSS, in the HTML files. In order to make Web pages vivid and colorful, Web page designers are using more and more complex JavaScript and CSS. Based on our observation from a large number of real Web pages, especially deep Web pages, the underlying structure of current Web pages is more complicated than ever and is far different from their layouts on Web browsers. This makes it more difficult for existing solutions to infer the regularity of the structure of Web pages by only analyzing the tag structures.

II. NEW SYSTEM

In this paper, we explore the visual regularity of the data records and data items on deep Web pages and propose a novel vision-based approach, Vision-based Data Extractor (ViDE), to extract structured results from deep Web pages automatically. ViDE is primarily based on the visual features human users can capture on the deep Web pages while also utilizing some simple nonvisual information such as data types and frequent symbols to make the solution more robust. ViDE consists of two main components, Vision-based Data Record extractor (ViDRE) and Vision-based Data Item
extractor (ViDIE). By using visual features for data extraction, ViDE avoids the limitations of those solutions that need to analyze complex Web page source files. Our approach employs a four-step strategy. First, given a sample deep Web page from a Web database, obtain its visual representation and transform it into a Visual Block tree which will be introduced later; second, extract data records from the Visual Block tree; third, partition extracted data records into data items and align the data items of the same semantic together; and fourth, generate visual wrappers (a set of visual extraction rules) for the Web database based on sample deep Web pages such that both data record extraction and data item extraction for new deep Web pages that are from the same Web database can be carried out more efficiently using the visual wrappers.

III. WEB CRAWLING AND META SEARCHING
Data records and data items in them machine process able, which is needed in many applications such as deep Web crawling and meta searching, the structured data need to be extracted from the deep Web pages. Each data record on the deep Web pages corresponds to an object. For instance, Fig. 1 shows a typical deep Web page from Amazon.com. On this page, the books are presented in the form of data records, and each data record contains some data items such as title, author, etc.

IV. WEB DATA RECORD AND ITEM EXTRACTION
Data record extraction aims to discover the boundary of data records and extract them from the deep Web pages. An ideal record extractor should achieve the following: (1) all data records in the data region are extracted and (2) for each extracted data record, no data item is missed and no incorrect data item is included.

V. VISUAL WRAPPER GENERATION
The complex extraction processes are too slow in supporting real-time applications. Second, the extraction processes would fail if there is only one data record on the page. Since all deep Web pages from the same Web database share the same visual template, once the data records and data items on a deep Web page have been extracted, we can use these extracted data records and data items to generate the extraction wrapper for the Web database so that new deep Web pages from the same Web database can be processed using the wrappers quickly without reapplying the entire extraction process.

VI. PRECISION AND RECALL
The basic idea of our vision-based data item wrapper is described as follows: Given a sequence of attributes Explanation for (f; l; d) fa1; a2; . . . ; ang obtained from the sample page and a sequence of data items item1; item2; . . . ; itemmg obtained from a new data record, the wrapper processes the data items in order to decide which attribute the current data item can be matched to. For item i and aj, if they are the same on f, l, and d, their match is recognized. The wrapper then judges whether itemip1 and ajp1 are matched next, and if not, it judges item i and ajp1. Repeat this process until all data items are matched to their right attributes.

VIII. UML DIAGRAMS

Figure 1

Figure 2
IX. CONCLUSION

With the flourish of the deep Web, users have a great opportunity to benefit from such abundant information in it. In general, the desired information is embedded in the deep Web pages in the form of data records returned by Web databases when they respond to users’ queries. Therefore, it is an important task to extract the structured data from the deep Web pages for later processing. We focused on the structured Web data extraction problem, including data record extraction and data item extraction. First, we surveyed previous works on Web data extraction and investigated their inherent limitations. Meanwhile, we found that the visual information of Web pages can help us implement Web data extraction. “Based on our observations of a large number of deep Web pages, we identified a set of interesting common visual features that are useful for deep Web data extraction. Based on these visual features, we proposed a novel vision-based approach to extract structured data from deep Web pages, which can avoid the limitations of previous works. The main trait of this vision-based approach is that it primarily utilizes the visual features of deep Web pages. Our approach consists of four primary steps: Visual Block tree building, data record extraction, data item extraction, and visual wrapper generation. Visual Block tree building is to build the Visual Block tree for a given sample deep page using the VIPS algorithm. With the Visual Block tree, data record extraction and data item extraction are carried out based on our proposed visual features.”

References


