A FRAMEWORK FOR AUTOMATIC EXTRACTION AND ANNOTATION OF STRUCTURED DATA FROM DEEP WEB

A thesis submitted to Pondicherry University in partial fulfilment of the requirements for the award of the Degree of

DOCTOR OF PHILOSOPHY IN
COMPUTER SCIENCE AND ENGINEERING

by

B. UMAMAGESWARI

Department of Computer Science and Engineering
Pondicherry Engineering College
Puducherry – 605 014

JULY 2018
CHAPTER 7

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This chapter concludes the thesis with the summary of research contributions made in the domain of Unsupervised Web Data Extraction. It also discusses in detail the merits and demerits of the proposed technique. Finally, it deals with the open research issues in the area of Automatic Web Data Extraction.

7.1 RESEARCH FINDINGS

The presence of high quality information in Deep Web Pages which are dynamically generated by embedding structured data records in server-side templates has made many researchers design techniques for automatic extraction of such data. The limitation of the existing systems includes their dependency on the structure of HTML pages for inducing wrappers. As regards deep web pages, two assumptions are made: templates are Union Free Regular Expressions; and dissimilar attributes are always enclosed between different tags. Such assumptions interfere with the accuracy of the system. Also, they require multiple pages generated using the same server-side template to identify repetitive patterns representing wrappers. Thus the proposed system has addressed the following aspects with respect to web data extraction:

i. **WDE based on Heuristics:** Initially a prototype system for Journal Information Extraction has been built based on the heuristics that the data records which are to be extracted are available as leaf nodes in the DOM tree. Rather than finding repetitive structure of templates, this solution uses a shortcut to locate the nodes corresponding to data records in the DOM tree using the above mentioned heuristics.

The text nodes are matched against the keywords such as SNIP, SJR, etc. Once the attribute label node is identified within the DOM tree, attribute value exists as sibling to
attribute label node. Attribute name-value pairs corresponding to data records are extracted and stored in a structured form.

**Merits:**

This system enables integration of journal information from multiple publishers’ websites and it enables the user to get the complete information about journals in a single interaction with the system.

**Limitations:**

The system can be used to extract data records from websites where the attribute label is present explicitly. It is not always true since some websites like social discussion forums are not having attribute labels mentioned explicitly.

**ii. WDE based on Semantic Labeling:** In order to overcome the limitations of the previous approach, WDE based on Semantic Labeling has been proposed. In this approach, the domain knowledge expressed as IF-Then rules guides the extraction process. It is based on the observation that informative section has repeated semantic content rather than non-informative section.

**Merits:**

a. It has no dependency on the structure of web page and therefore, missing attributes and dissimilar attributes surrounded by the same formatting tags can be handled and it improves the accuracy of extraction compared to template-based techniques such as RoadRunner [4], FiVaTech[5] and Trinity [6].

b. It requires determining the Semantic Rules only once for a business domain of interest. It need not be modified every time when we run the algorithm for a differently structured web page belonging to the same business domain.
c. The technique works even if a single search response page is available whereas the existing techniques require multiple input pages generated using the same server-side template.

d. Automatic labeling post extraction is handled whereas it is done manually in the existing approaches.

e. The versatility of the technique is proved by performing extraction from 31 websites belonging to 7 different business domains.

f. The consistency of the proposed technique is proved by calculating Inter-Quartile Range (IQR) corresponding to F1-Measure. The proposed system has the least IQR value equals to 0.2 which proves its stable performance in carrying out extraction across web pages belonging to heterogeneous domains.

iii. **Automatic Annotation using Multi-Heuristics Annotator:** Annotator is a major component of any Web Data Extractor. Most of the existing works in the literature such as RoadRunner [4], FiVaTech [5], Trinity [6], etc. concentrate on template deduction whereas annotation is done manually. In order to reduce the level of human intervention, annotation has to be handled automatically. The state-of-the-art approaches for annotation are based on heuristics such as table annotation, form-interface-based annotation, ontology-based annotation, in-text prefix and suffix-based annotation viz. Annotation cannot be achieved using a single heuristics. Different attributes require different heuristics in order to determine their label. Multi-Heuristics Annotator uses the domain knowledge as a primary entity to carry out the annotation process. If it cannot determine the label of certain attributes, then it tries the other heuristics such as structural feature-based annotator and form interface-based annotator. Experiments have been carried out on 211 query response pages belonging to TEL (Travel, Entertainment and Living group) dataset present in UUIC repository. It outperforms the existing approaches since it uses a combination of domain knowledge and heuristics to guide the labeling process.
Merits:

a. It has improved accuracy compared to Multi-Source Automatic Annotator and Ontology-Based Annotator.

b. The use of domain knowledge in addition to other heuristics used in the existing systems increases the number of attributes getting labelled.

7.2 OPEN RESEARCH ISSUES

i. The accuracy of the proposed technique is dependent on the domain knowledge expressed in the form If–Then rules. A systematic mechanism needs to be devised for framing and maintenance of If-Then rules.

ii. Building a repository of domain knowledge might help in improving the accuracy of extraction mechanism.

iii. After carrying out extraction and annotation process, in order to facilitate efficient analysis, semantic models could be built. It is used to specify concepts and relationships between concepts. Building such models automatically is a challenging research problem.

iv. Automatic Form Submission to get the desired search response pages is another challenging aspect which could be addressed.