CHAPTER 3

PROPOSED METHODOLOGY

3.1 OVERVIEW

The purpose of this research is to design an Automated Web Data Extraction System (AWDES) which can identify the target of data extraction with less amount of human intervention using semantic labeling and also to perform extraction at an acceptable level of accuracy. In AWDES, there always exists a tradeoff between the level of human intervention and accuracy. The goal of this research is to reduce the level of human intervention and at the same time provide accurate extraction results irrespective of the business domain to which the web page belongs. This technique departs from the techniques elaborated in Chapter 2 by using semantic labeling as a basis for extraction rather than considering the structure of web page expressed as templates. This chapter describes the basic model of WDE using set notation and the proposed solution to the problem of WDE and the research goals it meets. This chapter also provides a brief overview of the proposed methodology, data sets used and metrics used to evaluate the accuracy of the proposed system.

3.2 BASIC MODEL OF WDE

The problem of web data extraction can be stated as follows:

Let the website S consists of collection of template generated web pages \( P = \{p_1, p_2, p_3 \ldots p_i \ldots p_m \} \) where each web page \( p_i \) consists of set of data objects \( D = \{d_1, d_2, d_3 \ldots d_r \} \). Each data object \( d_j \) in D is a set of attribute-value pairs \( \{<x_1, y_1>, <x_2, y_2>, \ldots, <x_n, y_n>\} \). The problem of web data extraction involves extraction of \( D \) from every \( p_i \) in \( P \) belonging to \( S \).

**Definition 1: Template-generated web pages**

Template-generated web pages correspond to deep web pages which are generated dynamically by embedding data records obtained from back-end database in server-side
templates generated using server-side scripting languages such as PHP, Perl, JSP, ASP, etc.

3.3 PROPOSED ARCHITECTURE

From the detailed literature review in Chapter 2, it is clear that most of the existing systems have made assumptions such as the availability of multiple input pages; occurrence of missing or optional attributes is rare and so on. Also, these techniques are based on string pattern matching or DOM tree matching and therefore, change in structure of the web page requires occurrence of repeated patterns in the wrappers generated. This problem is addressed in the proposed technique by considering semantic information as the basis of extraction rather than structure or template of the web page. The overall architecture of the proposed AWDES is as shown in the Figure 3.1:

Figure. 3.1 Architecture of Automated Web Data Extraction System (AWDES)
The architecture of AWDES consists of the following three phases:

1. Development of prototype based on heuristic algorithm for extraction of journal metrics from Scientific Publishers’ website.

2. Shortcomings of heuristic algorithm are identified and a refined algorithm based on Semantic Labeling is proposed. Versatility of the proposed approach is proved by applying it to various diverse domains such as Health Discussion Forums, Product Websites, etc. Comparison of the proposed methodology with existing WDES based on qualitative and quantitative metrics has been performed.

3. Automatic Annotation is performed by using Multi-Heuristics Annotator.

**Phase 1: Development of prototype system for extraction of journal metrics from scientific publishers’ website**

*Input:* URL of Scientific Publishers’ website  
*Output:* Metrics such as SNIP, SJR, 5-year Impact Factor, ISSN, Title, etc.  
*Methodology:* A heuristic algorithm is designed based on the analysis of content presented in various websites belonging to Scientific Publishers’ domain such as ScienceDirect [64], Wiley [65], Nature [66], BMJ [67], etc. The technique is based on the fact that the content seen in a rendered page is available as leaf nodes in the DOM tree. The pattern of SNIP, SJR, and Title, etc. can be expressed using regular expression which is used to identify the semantic type of leaf node. Once the leaf nodes are assigned semantic labels, XPath is determined for those nodes. The attributes corresponding to similar data records also have the same XPath. Thus, the XPath acts as a template which can be used to carry out extraction from similarly structured web pages. Chapter 4 contains the elaborate description about the design and development of prototype system.

**Phase 2: Design of Web Data Extraction System using Semantic Labeling**

*Input:* URL of the website
Output: Data records in structured form

Methodology: As a first step, analysis of various websites belonging to the business domain of concern for which extractor has been designed shall be performed. This is done to identify common features expressed in various websites and then, semantic database is built. Semantic database contains rules to identify semantic label of the leaf nodes in the DOM tree representation of the web page. Each non-leaf node is associated with a list of semantic features it represents. Semantic Feature of a parent node is computed based on the semantic feature of the child nodes. This process helps to identify Data Rich Regions (DRR), thereby avoids considering navigation menus and advertisement during the extraction process. Once the data rich region is identified, XPath to data records within the DRR is determined.

Proving the versatility of the Proposed Solution, Comparison and Evaluation of Proposed Methodologies

The AWDES system using Semantic Labeling described in phase 2 has been applied to diverse domains: a) Health Discussion Forums and the results are published in [68] and b) websites belonging to domains such as books, product, movies, health sector, real estate and RISE repository [69].

Implementation Details

The system is implemented using Oracle Java Development Kit 1.7 and Jaunt API [70] is used to convert the web pages into DOM tree.

Data Sets Used

More than 1000 web pages collected from 7 diverse domains are used to prove the versatility of the proposed approach. It is also compared with the state-of-the-art techniques such as FIVATECH [5], RoadRunner [4] and Trinity [6]. In addition to the real world web pages, the technique is applied to data sets available in RISE repository [69]. Evaluation of the proposed system is done based on the following qualitative and quantitative metrics:

a) Qualitative Metrics:
Critical features identified to evaluate web data extractor as mentioned in [71] includes the following:

1. **Support for page flipping**: Ability to crawl through the consecutive web pages if the data records extend over more than a page.

2. **Support for single/multiple input pages**: Ability to perform extraction irrespective of single or multiple input pages.

3. **Data Rich Region Detection**: Identification of informative section rich in semantic content.

4. **Date Records Detection**: Identification of data objects within data rich region.

5. **Labeling post extraction**: Automatic labeling of attributes belonging to data records after the extraction process.

**b) Quantitative Metrics:**

The metrics used for data records extraction as shown in equations 3.1, 3.2 and 3.3 are used to quantitatively evaluate the accuracy of the proposed solution.

**Precision (P)** is defined as the ratio of number of records extracted correctly to the total number of extracted data records.

\[
\text{Precision (P)} = \frac{tp}{tp + fp} \quad (3.1)
\]

**Recall (R)** is defined as the ratio of number of data records retrieved correctly to the total number of data records available.

\[
\text{Recall (R)} = \frac{tp}{tp + fn} \quad (3.2)
\]
\( F_1 \text{-Measure} \) is defined as weighted harmonic mean of precision and recall.

\[
F_1 \text{-measure} = \frac{2PR}{P+R} \tag{3.3}
\]

**Statistical Measure: Inter-Quartile Range:**

It is a measure of variability obtained by dividing the data set into quartiles. The data set is divided into three quartiles, namely, Q1, Q2 and Q3. Q1 is the middle value in the first half of the rank ordered data set, Q2 is the median of the whole data set and Q3 is the middle value in the second half of the rank ordered data set. Inter Quartile Range (IQR) is the difference between Q1 and Q3.

Inter-quartile range of F-measure is determined to prove the consistency of the approach. Least IQR indicates that the variability of F-measure is least across various data sets considered, which proves the consistency of the technique irrespective of the heterogeneity of the structuring of web pages and the business domain to which it belongs.

**Phase 3: Automatic Annotation using Multi-Heuristics Annotator (MHA)**

The task of Web Data Extraction remains incomplete without annotation. Data records containing a set of attribute values are grouped into several Attribute-Value Groups such that all attribute values within a group represent the same concept. The process of grouping attribute values belonging to the same concept into single cluster is known as alignment. Similarity is determined by considering properties such as values of attributes like id, class, etc., XPath [72], prefix/suffix/sub-string similarity and visual similarity. Hierarchical clustering algorithm, namely, agglomerative clustering [73] is used to group similar attribute values into clusters. Annotation refers to assigning meaningful label to each Attribute Value Group. Annotation is performed by taking into account semantic rules used during extraction process, form interface keywords and structure features. Experiment is carried out using TEL dataset from UUIC repository [74]. Comparison of MHA with MSAA [75] and OBA [76] shows that the use of domain knowledge in addition to other features increases the number of attributes getting labeled and also, accuracy of labeling. Chapter 6 discusses in detail about the design and implementation of automatic annotation using MHA.