

Schema Extraction and Integration of Tabular Data from Multiple Web Sources

By

Amna Bibi

MCS143006

MASTER OF SCIENCE IN COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD**

2017

Schema Extraction and Integration of Tabular Data from Multiple Web Sources

By

Amna Bibi

MCS143006

A research thesis submitted to the Department of Computer Science,
Capital University of Science and Technology, Islamabad
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE



**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD**

2017



CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD

Islamabad Expressway, Kahuta Road, Zone-V, Islamabad

Phone: +92 51 111 555 666, Fax: 92 51 4486705

Email: info@cust.edu.pk, Website: <http://www.cust.edu.pk>

CERTIFICATE OF APPROVAL

**Schema Extraction and Integration of Tabular Data from Multiple
Web Sources**

by

Amna Bibi

MCS143006

THESIS EXAMINING COMMITTEE

S No	Examiner	Name	Organization
(a)	External Examiner	Dr. Onaiza Maqbool	QAU, Islamabad
(b)	Internal Examiner	Dr. Muhammad Tanvir Afzal	CUST, Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood

Thesis Supervisor

November, 2017

Dr. Nayyer Masood

Head

Department of Computer Science

Dated : November, 2017

Dr. Muhammad Abdul Qadir

Dean

Faculty of Computing

Dated : November, 2017

Copyright ©2017 by CUST Student

All rights reserved. Reproduction in whole or in part in any form requires the prior written permission of Amna Bibi (MCS143006) or designated representative.

Certificate

This is to certify that Ms. Amna Bibi has incorporated all observations, suggestions and comments made by the external evaluators as well as the internal examiners and thesis supervisor. The title of her Thesis is: **Schema Extraction and Integration of Tabular Data from Multiple Web Sources.**

Dr. Nayyer Masood
(Thesis Supervisor)

DECLARATION

It is declared that it is an original piece of my own work, except references mentioned in the text. This work has not been submitted in any form for another degree or diploma at any university or other institution and shall not be submitted to pursue another degree from any other university or institution by me in future.

Amna Bibi
MCS143006
November, 2017

DEDICATION

To,
a Person who is

THE REHMAT for all Universe,

and

To my parents

who taught me to walk and survive in this world,
and who have been a source of inspiration for me

Table of Contents

CHAPTER 1	1
1. INTRODUCTION	1
1.1 HTML and HTML Tables	2
1.2 Schema Extraction	8
1.3 Schema Integration	9
1.4 Schema Matching	10
1.4.1 Element Level	11
1.4.2 Structure Level	11
1.4.3 Instance Level	11
1.4.4 Taxonomy	12
1.5 Problem Statement	13
1.6 Purpose	13
1.7 Scope	14
1.8 Significance of the Solution	14
1.9 Organization of Thesis	14
1.10 Definitions of terms used	14
CHAPTER 2	15
2. LITERATURE REVIEW	15
2.1 Comparison of Schema Extraction Techniques	26
CHAPTER 3	28
3. METHODOLOGY	28
3.1 Data Domain	31
3.2 Preprocessing	31
3.2.1 Data Collection	31
3.2.2 Global Database Creation	32

3.3	Proposed Methodology	32
3.3.1	Research Question 1	32
3.3.2	Research Question 2	35
3.3.3	Research Question 3	39
CHAPTER 4		47
4.	RESULTS & EVALUATIONS	47
4.1	Quantitative Analysis	48
4.1.1	Research Questions	48
4.2	Comparison with Other Techniques.....	60
4.3	Query Validation	61
CHAPTER 5		64
5.	CONCLUSION AND FUTURE WORK	64
5.1	Conclusion.....	64
5.2	Future Work	65
REFERENCES		66
Appendix A.....		69
Appendix B		74
Appendix C		78
Appendix D.....		82
Appendix E		86
Appendix F.....		87

List of Figures

Figure 1-1: A sample html code using tables for layout design ² -----	3
Figure 1-2: A sample HTML code for a simple table ³ -----	4
Figure 2-1: Sample Table Format used by (Nagy et. a., 2014)-----	17
Figure 2-2: Illustration for Algorithm 3 of HTW -----	18
Figure 2-3: Process of integrating Web Tables on web page (Akbar et al., 2015) ----	21
Figure 2-4: Flowchart of proposed technique (Hao et. al., 2011)-----	22
Figure 3-1: Architecture Diagram of Proposed System-----	29
Figure 3-2: Flow chart for Schema Matching Algorithm -----	30
Figure 3-3: Output of Data Extraction Algorithm -----	34
Figure 3-4: Example of web table with Missing Values -----	34
Figure 3-5: Modified Algorithm 3 of HTW+-----	35
Figure 3-6: Algorithm to Predict Headings -----	37
Figure 3-7: Similarity Based Upon Common Sub-String Algorithm -----	41
Figure 3-8: Finding attribute in Taxonomy -----	42
Figure 3-9: Insert Query-----	46
Figure 4-1: Snapshot of Faculty information from NUML university website -----	50
Figure 4-2: Result of schema extraction technique -----	51
Figure 4-3: Comparison of Data Extraction results-----	53
Figure 4-4: Results of Schema and Data Extraction Combined-----	54
Figure 4-5: Schema Extraction of Without Header Row Table-----	56
Figure 4-6: Data Extraction of Without Header Row Table-----	58
Figure 4-7: Graphical Representation of experimental results of SM Algorithms ----	59

ACKNOWLEDGEMENT

I simply bow my head before Almighty Allah for giving me faith in my abilities and enabling me to accomplish this project and granting me with His Special Merci, Blessings and unlimited help throughout the phases of the thesis.

I am especially grateful to **Dr. Nayyer Masood** for his extended cooperation in all regards. I am extremely grateful for his invaluable guidance, mentorship and more importantly his patience with me. Without his consistent support and encouragement, this thesis would not have become a reality.

I owe a great deal to all respected teachers, my family and well-wishing friends who extended towards me whatever help was required.

In the nutshell, it is the blessings of Almighty Allah, mother's orisons, the guidance of my supervisor **Dr. Nayyer Masood, HoD, Computer Science Department**, to complete this research work and thesis successfully.

Amna Bibi

MCS143006

November, 2017

Abstract

World Wide Web is a huge source of data/information. The data is generally in form of text, lists, tables etc. Tables are a common method of representing data in many domains such as entertainment, business platform and educational forums. Applying ad-hoc queries on web tables is a difficult task because for that this data needs to be stored into the database. The problem is addressed in this research by modifying a HTML Wrapper Induction (HTW) technique which is based upon the HTML table components. The new technique is named as HTW+ that has overcome many of the limitations of HTW like, empty cells, missing headings, schema extraction of without heading tables and data extraction. In pre-processing phase, a database is created with necessary tables and the input files containing the HTML code of web tables are prepared. These .txt files are given as input to HTW+. Original HTW+ mainly comprises three algorithms. In first algorithm, HTW+ calculates the number of rows and columns. Second algorithm calculates the boundary of Property (headings) row. These two algorithms have been adopted from HTW as such; however, third algorithm is modified by HTW+ to store the empty cell in array and to extract data along with schema. After schema and data extraction, three schema matching (SM) algorithms, Equal Names, Similar Sub-String Matching and Taxonomy are used. During the process, heterogeneity issues are faced and resolved. From mentioned SM techniques, experimental results show that the Taxonomy is best technique in this research scenario. Once the schema is matched, data from the array is moved into the Master Table. The same process is repeated with all web tables. In the dataset of 134 web tables, 120 web tables have proper and explicit heading row and 14 web tables are without headings. Out of 134 web table technique is able to extract the schema and data of 125 web tables and of these 125 web tables, 113 web tables are successfully inserted into the database table with the precision of 99.12% and 84.32% recall. The data is inserted in a database table resulting in an integrated database.

CHAPTER 1

1. INTRODUCTION

The World Wide Web is home to a large volume of data. This data consists of diverse information in structured, semi-structured and unstructured form. The volume of this data is increasing exponentially (Crescenzi et al., 2001). This data is commonly available to people to use and get benefit (Laender et al., 2002).

Tables (or Web Tables) and lists are commonly used to represent structured data on web (Cafarella et al., 2008). A table is a two dimensional grid of rows and columns. Each column represents an attribute. Billions of tables are available on web containing valuable information (Wang et al., 2012). Tables are a simple, meaningful, effective and popular way of representing data.

Generally, data from web sources can be accessed either using search engines (using keywords), or navigating (browsing) through the web pages. Both are not effective techniques of fetching data as both have considerable limitations. Browsing is not suitable for searching a particular data item as searching through keywords can produce a list of unwanted data that is generally difficult to sort and sift (Laender et al., 2002). Web table data cannot be queried, manipulated and analyzed using either of these methods. Additionally, due to the diverse nature and typology of web data, traditional warehousing techniques and architectures cannot be used. Instead, the issue of accessing this data can be resolved if this data can be extracted and stored in a database.

Extracting data from tables and storing it in a database is very useful for many value added services. It is equally beneficial in business, education and social web domains. In business domain, data extraction techniques help in reducing time and manpower and increase the business efficiency. Using these techniques, companies can collect huge volume of data from web in a relatively short time. This data can also help the analysts and managers to revise or change the business strategies and plans. Context-aware advertising, customer care, comparative shopping, Meta query, opinion mining and database building are the major applications of web data extraction techniques (Ferrara et al., 2014).

With the growth in popularity of social media websites like, Facebook, Tweeter and Instagram, extracting data from these websites has gained special attention. This is relatively a new research area. Scientists of different domains like Sociology, Political Science, and Anthropology have an opportunity to analyze and understand the dynamics of human behavior at a global scale and in a real time-fashion (Ferrara et al., 2014).

If the data about the faculty members of universities can be stored in single database, that can help students and other stake holders in many ways. Student/researchers can easily search and compare the profiles of respective supervisors of different universities of same research area. This can also help students in choosing the university by comparing the faculty profiles of different universities.

To obtain the benefits of storing web table's data into database in business, education and social web domain, the schema of the table needed to be extracted.

Schema is the definition of the table and it needs to be extracted before extracting the data from web tables. In this research, the method proposed by (Purnamasri et. el., 2015) is used to extract schema and the method is further extended to extract the data from web tables. The technique is also being modified to cover additional aspects and existing limitations. The technique was not handling the formatting tags in between the given `<table> </table>` tags. It was also not applied on actual web sources, which this research is aiming to do through implementation on real web tables. This technique was unable to detect column headings for tables without explicit headings. Such tables are also handled in this research. In this research, the empty cells are also taken into account which has not previously been addressed by HTW (HTML Table Wrapper) (Purnamasari et al., 2015). The technique is applied on multiple web sources for schema extraction of web tables and an integrated schema is created for data storage.

In this chapter, schema extraction and integration are defined in detailed. First of all, HTML, HTML tables and different formats of tables used on the web are introduced. In next section, schema extraction, integration and matching is discussed in detail.

1.1 HTML and HTML Tables

HTML is Hypertext Markup Language. It is used to design web pages. It uses tags to

create web pages. The tags usually come in pairs. Most of the tags have opening and closing tag written in angle brackets (Grannell, 2008). As discussed above, HTML uses table to show information in a meaningful manner. It uses <table>..... </table> tags to form tables on websites. This tag has some basic elements which are collectively used to form tables on a website. These elements include <th>...</th>, <tbody>...</tbody>, <tr>...</tr> and <td>...</td>. <tr> tag defines the rows and <td> tag defines a cell of the table. <td> tags are used inside <tr> tag. Number of <td> tags depends upon the number of columns of table. These web tables (commonly called tables) are not only used to represent data on a website but also for other purposes such as designing the layout of web pages. The latter is not an obvious use of tables but it is a commonly used method to divide a web page into sections¹. Fig 1.1 shows an example of the HTML code of tables used as layout². However, the most important purpose of tables on web is to show data in tabular form. Representing data in tables is convenient and easily understandable as people are familiar with this format.

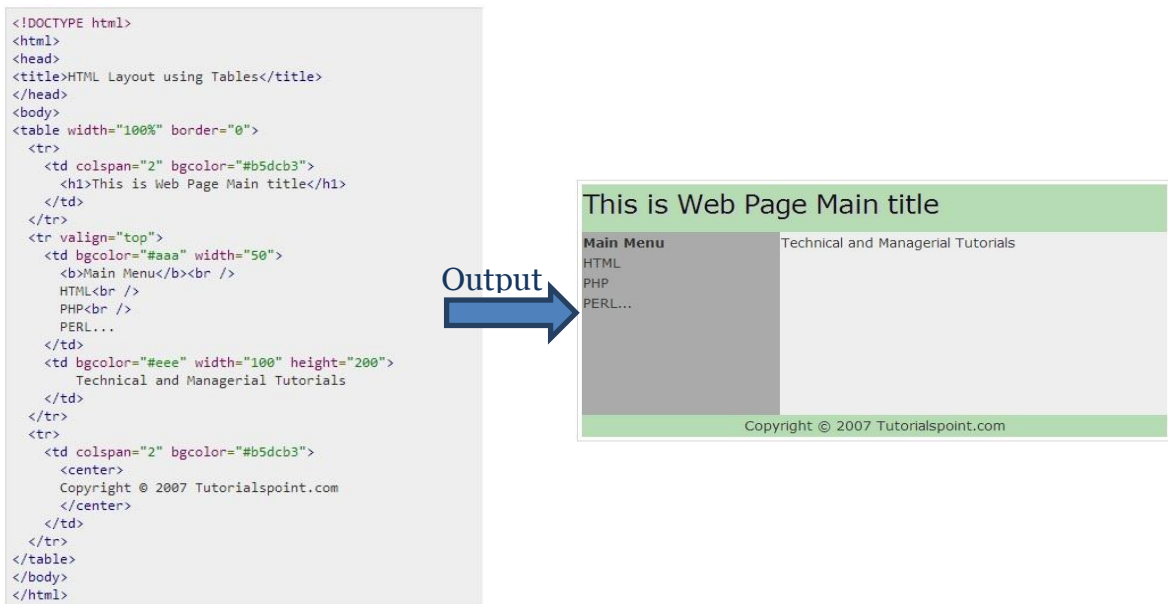


Figure 1-1: A sample html code using tables for layout design²

Web tables (tables) are used to represent different types of information on web. Different

¹ <http://www.echoecho.com/htmltables.htm>

² [https://www.tutorialspoint.com/html/html layouts.htm](https://www.tutorialspoint.com/html/html%20layouts.htm)

websites use these tables differently in different ways. Some examples of web tables taken from different web sources are shown in the Table 1.1, Table 1.2 and Table 1.3. Different formats are due to the different ways of using the <table> ... </table> tags. However, only these tags are not enough to draw the table on a web page. Some other formatting tags with in <table> ... </table> tags are used to create and format the tables. A sample of a simple table containing two rows and three columns is shown in the Figure 1-2;

Different tables format are being followed on websites of single domain data like

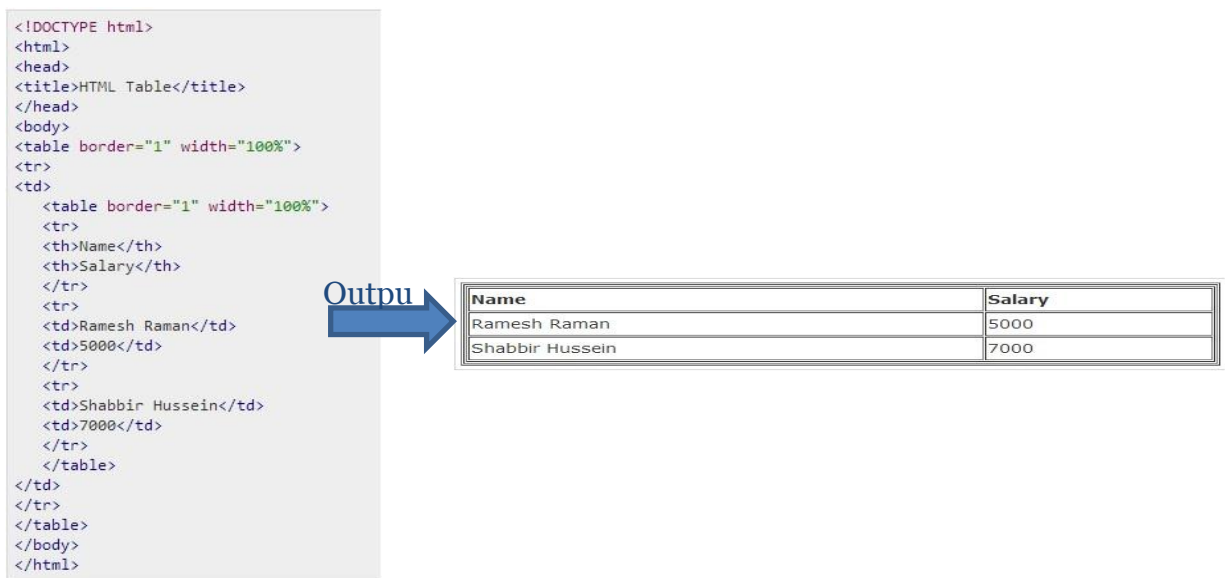


Figure 1-2: A sample HTML code for a simple table³

education, business and media etc. Given below are screen shots of web tables, showing the information of faculty members of some universities taken from the official university websites, along with their HTML codes.

The web table given in Table 1.1 is the required web table format used in this research.

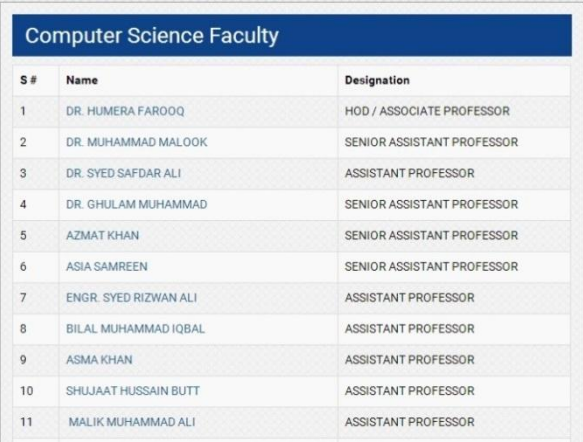
Image of Web Table		Html Code of Web Table	
 <p>Bahria University Karachi website table containing faculty data</p>		<pre> <table class="table table-bordered table-striped"> <tbody> <tr> <td width="48">S #</td> <td width="288">Name</td> <td width="240">Designation</td> </tr> <tr> <td width="48">1</td> <td width="288">DR. HUMERA FAROOQ</td> <td width="240">HOD / ASSOCIATE PROFESSOR</td> </tr> <tr> <td width="48">2</td> <td width="288">DR. MUHAMMAD MALOOK</td> <td width="240">SENIOR ASSISTANT PROFESSOR</td> </tr> </table> </pre>	

Table 1.1: Example of HTML table with multiple records in single <table> tag

Some web tables have merged title or data rows and do not have explicit header row. The web table given in Table 1-2 is an example of such tables.

Image of Web Table	Html Code of Web Table													
<p>Academic Staff</p> <hr/> <p>Emeritus Professors</p> <table border="1"> <tr> <td>CHEN Tien Chi</td> <td>陳天機</td> <td>Emeritus Professor</td> </tr> <tr> <td>WONG Chak Kuen</td> <td>黃澤權</td> <td>Emeritus Professor</td> </tr> </table> <p>Distinguished Professor-at-Large</p> <table border="1"> <tr> <td>YAO Chi Chih Andrew</td> <td>姚期智</td> <td>Distinguished Professor-at-Large</td> </tr> </table> <p>Chairman</p> <table border="1"> <tr> <td>HENG Pheng Ann</td> <td>王平安</td> <td>Chairman and Professor</td> <td>Research Areas visualization, virtual reality, graphics, human computer interaction, medical imaging, surgical simulation</td> </tr> </table> <p>The Chinese University of Hong Kong Karachi website table containing faculty data</p>	CHEN Tien Chi	陳天機	Emeritus Professor	WONG Chak Kuen	黃澤權	Emeritus Professor	YAO Chi Chih Andrew	姚期智	Distinguished Professor-at-Large	HENG Pheng Ann	王平安	Chairman and Professor	Research Areas visualization, virtual reality, graphics, human computer interaction, medical imaging, surgical simulation	<pre> <table class="wide_tb" width="635" border="0" cellspacing="0" cellpadding="4"> <tr> <td colspan="3" class="page_hd_3">Emeritus Professors</td> <td width="40%">&nbsp;</td> </tr> <tr> <td colspan="3">&nbsp;</td> </tr> <tr class="odd_tr"> <td width="25%">CHEN Tien Chi</td> <td width="10%">陳天機</td> <td width="25%">Emeritus Professor</td> <td width="40%">&nbsp;</td> </tr> <tr> <td>WO NG Chak Kuen</td> <td>黃澤權</td> <td>Emeritus Professor</td> <td>&nbsp;</td> </tr> </table> </div> <div class="wide_col"> <table class="wide_tb" width="635" border="0" cellspacing="0" cellpadding="4"> <tr> <td colspan="3" class="page_hd_3">Distinguished Professor-at- Large</td> <td width="40%">&nbsp;</td> </tr> <tr> <td colspan="3">&nbsp;</td> </tr> <tr valign="top" class="odd_tr"> <td width="25%">YAO Chi Chih Andrew</td> <td width="10%">姚期智</td> <td width="65%">Distinguished Professor-at- Large</td> </tr> </table> </div> </pre>
CHEN Tien Chi	陳天機	Emeritus Professor												
WONG Chak Kuen	黃澤權	Emeritus Professor												
YAO Chi Chih Andrew	姚期智	Distinguished Professor-at-Large												
HENG Pheng Ann	王平安	Chairman and Professor	Research Areas visualization, virtual reality, graphics, human computer interaction, medical imaging, surgical simulation											

Table 1.2: Example of without heading web table with merged title and header row

Some of the websites display one record per table. Table 1.3 is showing example of such a table.



Image of Web Table	Html Code of Web Table
<p>Computer Science Faculty</p> <div data-bbox="305 310 847 409">  <p>Prof. Dr. Muhammad Abdul Qadir PhD Parallel & Distributed Computing University of Surrey, Guildford, UK Professor / Dean Faculty of Computing Email: aqadir@cust.edu.pk Research Areas: Semantic Computing, Semantic Digital Libraries, Ontology evaluation, Ontology mapping and merging, Semantic Cache Query Processing, Sem Read More...</p> </div> <hr/> <div data-bbox="305 436 847 535">  <p>Dr. Nayyer Masood PhD Computer Science University of Bradford, UK Professor / HoD Computer Science Email: nayyer@cust.edu.pk Research Areas: Multidatabase Systems, Schema Translation, Schema Evolution, Schema Integration, Data Integration, Data Mining, Read More...</p> </div> <p>CUST, Islamabad website table containing faculty data</p>	<pre> <table class="cvlist"> <tr> <td style="width:100px;height:123px;"> </td> <td > Prof. Dr. Muhammad Abdul Qadir
<p style="padding:0;"> PhD Parallel & Distributed Computing University of Surrey, Guildford, UK </p> Professor / Dean Faculty of Computing
 Email: aqadir@cust.edu.pk
Research Areas: Semantic Computing: Semantic Digital Libraries, Ontology evaluation, Ontology mapping and merging, Semantic Cache Query Processing, SemRead&nbsp;More... </td> </tr> </table> </div><div class="item"> <table class="cvlist"> <tr> <td style="width:100px;height:123px;"> </td> <td > Dr. Nayyer Masood
 <p style="padding:0;"> PhD Computer Science University of Bradford, UK </p>Professor / HoD Computer Science
 Email: nayyer@cust.edu.pk
Research Areas: Multidatabase Systems, Schema Translation, Schema Evolution, Schema Integration, Data Integration, Data Mining, Read&nbsp;More... </td> </tr> </table> </pre>

Table 1.3: Example of HTML table with single record in single <table> tag

The same <table>.....</table> tag is used to form both these tables but in different styles. The layout and design of both HTML codes is same.

Other than the usual <table> tag, some other tags like <div> and lists are also used to

form tables on web. However, such table formats are not within the scope of the present research.

1.2 Schema Extraction

Schema is the description of structure of database (Elamsri, Navathe, 2004). Schema extraction is the process of drawing out schema of a given table (Srivastava, 2010). Schema extraction includes extracting attributes names of the table, their datatypes, relationship with any other table and any constraints applied on table (Srivastava, 2010). Extraction of schema from web tables is quite different as compared to database table because schema related information like primary key, other constraints and Meta data etc. is not explicitly provided in web tables (Adelfio, & Samet, 2013). So, the schema of web tables needs to be extracted.

In this research only web tables with grid like display are considered. In such tables, mostly the first row contains the headings for the columns and the rest of the rows hold data against each column headings. Considering this behavior of tables many schema extraction techniques, (Adelfio et al., 2013; Zhai et al., 2005; Nagy et al., 2014, Purnamasari et al., 2015; Lerman et al., 2001; Gultom et al., 2011) are proposed. Some of these techniques are supervised while others are un-supervised and some of these only use the programming techniques. These techniques differ in their approach, type and format of web tables.

Manual data extraction is one technique used for schema extraction but it involves a lot of human effort. Programmer observes web page and source code and deduces the data patterns from it. This is used to write a program which extracts the data from the web page. This method is not automatic and scalable for very large number of web pages. Other approaches which have automation to some extent are wrapper induction and automatic extraction. In wrapper induction (Cohen et al., 2002; Kushmerick, 2000), firstly the pages/data records are labeled manually and a set of extraction rules is deduced. These rules are then used to extract data from similar pages. This technique still involves manual effort.

Automatic method (Adelfio & Samet, 2013) finds patterns or grammars from similar pages containing similar data records and uses these patterns to extract data from new pages. The pages to extract the patterns are provided manually or by some other system. So far most of the techniques (Adelfio et al., 2013; Zhai et al., 2005; Nagy et al., 2014, Purnamasari et al., 2015; Lerman et al., 2001; Gultom et al., 2011) extract and store data into databases by creating separate tables for each web page. To make comparative queries, data needs to be placed in one database table. This is called database integration or schema integration.

1.3 Schema Integration

The process of integrating multiple database schemas into a single (global) schema is called schema integration (Batini et al., 1986; Elmagarmid et al., 1999). In the literature (Batini et al., 1986; Elmasri et al., 1986; Larson et al., 1989), this process has been mainly applied on typical database schemas or views. However, this research is based on applying the schema integration on the schemas extracted from web tables belonging to the same domain, like, automobiles, medical and universities. The major challenge in schema integration is handling semantic heterogeneities (Elmagarmid et al., 1999). A semantic heterogeneity reflects a situation where same real-world concept is modeled differently in different database schemas (Elmagarmid et al., 1999). Schema integration of web tables introduces even more heterogeneities as schemas in this case are not properly defined rather are extracted from web tables.

Two types of schema integration strategies are being used; binary strategies and n-ary strategies (Batini et al., 1986).

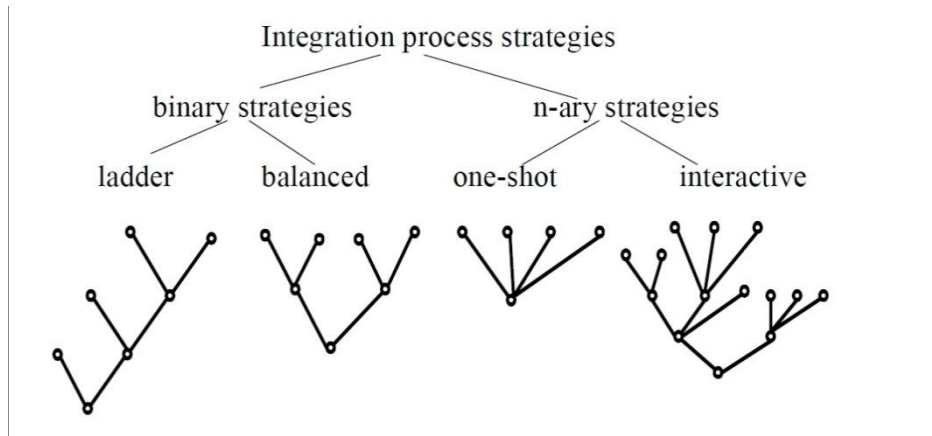


Table 1.4: Integration Strategies

In binary strategy, when two schemas are merged at a time, it is called ladder strategy. In n-ary, n schemas are integrated at a time. In one-shot n-ary, n schemas are merged in one step, otherwise it is interactive (Batini et al., 1986).

Typically, schema integration involves schema translation, schema matching and schema merging (Elmagarmid et al., 1999). However, in the context of web tables, it involves only two phases, i.e. schema matching and schema merging. The most critical task is schema matching, in which schema elements from different databases modeling the same concept are identified. Once the matching elements are identified, merging them into one is relatively straight forward.

In the next section, schema matching is discussed in detail.

1.4 Schema Matching

Schema Matching is a process of comparing and matching the schemas of two or more different database tables. Same type of attributes appears once in the resultant integrated table. The unmatched attributes are also included in the resultant table.

There are many schema matching methods. Three types of schema information are used to perform schema matching, named as Element Level, Structure Level and Instance Level.

1.4.1 Element Level

If element (attribute) level schema information like element's name/domain information such as table name, column name, column data type and column length are used for schema matching, it is called element level matching. It finds correspondence between elements in the first schema with the elements in another schema. Following are some of the element level schema matching techniques;

- Equal Names(EN)
- Edit Distance
- Similarity Based on Common Sub-Strings (SCS)
- Canonical Name checking
- N-Gram
- Hyper name checking
- Hypo name checking
- Using of generic dictionary

1.4.2 Structure Level

Attribute's structure of two tables may differ for example in one table we have attribute "Name" but in another table "Name" can be obtained by combining "First Name" and "Last Name". Same attribute has a different structure in two tables and in order to compare these two one attribute must be transformed. When schema structural information is used to perform schema matching, it is structural level matching.

1.4.3 Instance Level

In Instance-level schema matching information extracted from stored data instance in schema elements is used to find correspondence between the schema elements. This is a more complex method as compared to element level and structure level due to the sheer size of data instances. These operations measure the similarity between two strings and award the score between 0 and 1.

In this research schema is matched based upon the attributes names as no other information about the table schema is provided. For this type of matching, element level

schema matchers give good results.

On the basis of research precedent or previous research two element level techniques, Equal Names and Similarity based on Common Substring and Edit Distance are considered more suitable for the purpose of this research (Zeeshanudin, 2011) based upon the observation of the nature of the attributes for the considered domain .

1.4.3.1 Equal Names

The similarity score of 1 is awarded to the set of strings whose names are equal (Zeeshanudin, 2011).

1.4.3.2 Similarity based on Common Substring

This operation compares two strings and checks for the common sequence of characters of length N. More are the common characters; more chance of being similar strings (Zeeshanudin, 2011).

Along with these two schema matching techniques, taxonomy is built for the common attributes of the domain selected for this research and it is used to match the attributes.

1.4.4 Taxonomy

Taxonomy refers to classification or arrangement of items. It is usually based on predefined criteria that help to arrange items in a manner suited to further discussion and analysis. The information may be used to develop conceptual frameworks furthering research agenda.

The concept of taxonomy is based on being able to categorize items distinctly into sub groups while maintaining each group's particular inherent characteristics and arrangements. In terms of use, a good taxonomy is characterized by simplicity and ease of use³.

Taxonomies are being built and used in many fields like biology, education and computer science. In computer science, taxonomies are being used in many fields like information retrieval, database and software engineering.

³ <http://searchcontentmanagement.techtarget.com/definition/taxonomy>

It is not necessary that the attributes of two different tables of same domain use the same words to represent an attribute. Synonyms or the alternative names may be used instead. Matching such attributes is difficult using Equal Names and Similarity based on Common Substring methods. In such situations, it is better to construct taxonomy. The taxonomy contains the synonyms and alternative names for all the attributes which tables of same domain can share.

1.5 Problem Statement

Most of the schema extraction techniques do not address the missing and without heading tables' issues and none of the methodologies performs the schema integration on schema extracted from multiple web tables. This proposed research aims to overcome these deficiencies that will allow a better querying over the data extracted from multiple web tables belonging to the same domain.

HTW (HTML Table Wrapper) (Purnamasari et al., 2015) is a recent schema extraction technique which has been applied on dummy tables to extract the header row (schema) only, not the data. However, it does not handle the tables with missing heading and with no header row. The questions addressed in this research are based on the gaps found in schema extraction approaches in general and in particular in HTW. Following are the questions that are focus of this research;

RQ1: Can the HTW be enhanced to extract data from real web tables?

RQ2: Can the HTW be enhanced to extract the schema of web tables with missing headings and without header row?

RQ3: Can we use the schema extracted from multiple web tables to build a schema integration tool?

1.6 Purpose

The purpose of this research is to extract the schema and data given in web tables and store it into a database. Furthermore, extracted schema of same domain from different websites is compared and stored in a single database.

1.7 Scope

The proposed solution will have a great impact on the research community as centralized data of single domain will be available for users to applying ad-hoc queries.

1.8 Significance of the Solution

This research proposes a reasonable solution to problem of ad-hoc querying the tabular data on web by storing it into a database and integrating multiple websites data in one database.

1.9 Organization of Thesis

This document is organized in five chapters. Chapter one discuss the introduction of the problem, purpose of the research, and its significance. Chapter two provides a review of the state of the art approaches to relevant to the study. Chapter three presents the proposed methodology of research work. Chapter four describes the results. To evaluate the results, three techniques are used. HTW+ is compared with other schema integration techniques. It is also quantitatively analyzed by calculating Precision, Recall and F-Measure. Inserted data is also validated through query validation. Chapter five of this document concludes the whole research work with conclusion and future work.

1.10 Definitions of terms used

Metadata: Data about data

Schema Schema extraction is the process of drawing out schema of a given

Extraction: table

Schema Matching: It is a process of comparing and matching the schemas of two or more different database tables

Data Integration: Combining data of multiple resources at one place

CHAPTER 2

2. LITERATURE REVIEW

Web is a considerable source of tabular data. Extracting and storing this data into relational schema is a very popular research domain. Schema extraction has applications in business, scientific and social domains. In this section, the researcher will discuss some state of the art schema extraction techniques.

The main objective of such methods according to a conference on World Wide Web (Zhai & Liu, 2005) was to segment the data records, extract data items and save the data in a database. This method provides an automated system to extract data from web and put data in database. HTML uses tags to construct a web page. <table> tag is used to represents table on web. This tag is nested in nature. It inserts data in table row wise. <tr> tag is used to insert the rows and <td> tag inserts the data in a particular cell of that row. Method is divided into two sub tasks. First task is to identify the data records. For this purpose, visual information is used to construct a tag tree. Tag tree is constructed by following the nested structure of HTML code. The root of the tree contains the tag <table> for each row in the table a child is “tr” is created. For each data cell in the row, a child is created for that row. The leaf level contains the actual data in that particular cell of the row of the table. After constructing the tag tree, data regions are identified by comparing tag strings of individual nodes and combination of multiple adjacent nodes. Second task is to align and extract data from the identified data records. Partial tree alignment technique is used. Tree edit distance is used to identify the data records. Trees are matched with each other node by node. Trees are aligned by gradually growing a seed tree T_s . The tree with the maximum records is chosen as the starting seed tree. Then for each node n_i in T_i a matching node n_s is found in T_s . When a matching node is found in T_s , a link from n_i is created to n_s . If no match is found for n_i , then seed tree is expanded by inserting this node into it. Data item nodes are not used during the matching. The data item in the matched nodes' children is inserted into one column in database table.

A conditional random field (CRF) based classification technique with the combination of logarithmic binning is proposed by (Adelfio et. al., 2013). This is a fully automated

method. Each table contains different types of rows. These rows include caption, heading, empty and data rows. Each row is classified based upon the row features. To classify the rows, row features are extracted. The classifier's accuracy is highly dependent upon the input row features. Features for each individual cell is extracted based upon natural language processing principle, formatting, layout, style and values of the cell and then all these attributes are combined using binning feature to construct row features. In next step, logarithmic binning method is applied in which individual cell attributes are used collectively to encode row features. For each possible row feature a bin is formed and each bin is assigned a value which represents its feature. After row features extraction, row labels are assigned to each row based on CRF. CRF is trained with human classified rows. After training the CRF is used to label huge volume of data. The output of the CRF is a sequence of row labels like "TNNHGDDDAGDDDABN". This output helps in extracting schema of the relational table. Column names are decided based upon the header row(s), datatype is determined by the type frequency within the data rows of each column, additional attributes can be determined by the group header rows and data rows are determined by the data records. This method is applied not only on HTML tables, but also on spread sheets. This method has shown improved results over previous methods. Nagy and his co-researcher (Nagy et. a., 2014) introduced a technique to convert the web tables to relational tables. This is the first end to end approach that produces an access compatible canonical table. The HTML table is converted into an excel table and from excel table its CSV file is generated. Table is segmented based upon the indexing property rather than appearance features. To segment the table minimum indexing point (MIP) and four critical cells CC1, CC2, CC3 and CC\$ are calculated using (Embley et. al., 2011). CC1 and CC2 determine the stub headers; CC3 and CC4 indicate the data regions. MIP (CC2) is determined by searching from the cell A1 for unique columns and row header rows. The figure below is an example of segmented table where A2, A3, B4 and G10 are the CCs and MIP is CC2.

	A	B	C	D	E	F	G
1	4 Schools and pupils by school size. School years 2003/04-2009/10. Per cent						
2	School years	Schools	Schools	Schools	Pupils	Pupils	Pupils
3		Less than 100	100-299 pupils	300 pupils	Less than 100	100-299	300 pupils
4	2003/04	36.2	39	24.8	8.7	39.3	52
5	2004/05	35.2	39	25.8	8.7	38.3	53
6	2005/06	35.2	39	25.8	8.8	38.3	52.9
7	2006/07	34.3	40	25.7	8.4	39	52.6
8	2007/08	34	39.6	26.4	8.3	38.2	53.5
9	2008/09	33.3	40	26.7	8.1	38.2	53.7
10	2009/101	32	40.7	27.3	7.7	38.2	54.1
11	1	Preliminary figures.					
12	Explanation of symbols						

Figure 2-1: Sample Table Format used by (Nagy et. a., 2014)

The categories are extracted by comparing the number of unique elements in the cross-product of a pair or header rows with the length of the header. Two rows of the column header of above fig. constitute two categories because the Cartesian product ($\{\text{Schools, Pupils}\} \times \{\text{Less than 100, 100-299 pupils, 300 or more}\}$) of the two rows has 6 elements, which is exactly the length of the column header. From the category extraction output, canonical table is generated. This table is used to query the data. Results show that the proposed approach is more effective for large, complex and heterogeneous tables.

The technique HTW (HTML Table Wrapper), proposed by (Purnamasri et. al., 2015) first finds the area of the table and then extracts data. First of all, table is detected and then property (heading) portion of the table is detected before extracting data. The technique is divided into three steps and algorithm for each step is formulated. In first step number of rows and columns is calculated. To calculate the total rows in table, the algorithm counts the `<tr>` tags in `<table>` tag of the HTML code of the website. Columns are calculated by counting the `<td>...</td>` tags in each `<tr>` tag. The algorithm also checks the `colspan` attribute in the `<td>` tag. It adds the value of `colspan` in the column count. When the number of rows and columns is known, the size of the table is easily calculated by `rows*columns` formula. In second algorithm, the property of the table is detected. Generally, the first row of the table contains the headings of the columns. In many tables, the headings span more than one row. In this case, either the columns, or the rows are merged. The algorithm checks for the “`rowspan`” attribute in each `<td>` tag in `<tr>` tag of table to calculate the length of the property of the table. It returns the value of highest

rowspan attributes amongst every <td> tag. The third algorithm actually extracts the data from the table. It takes the value of the rowspan returned from the second algorithm to extract the heading of the columns. While reading the data in <td> tag of <tr> tag, it checks the value of “colspan”. If its value is greater than 1, it concatenates the content in this cell with the columns below it. After reading the header rows, it reads the cells row by row.

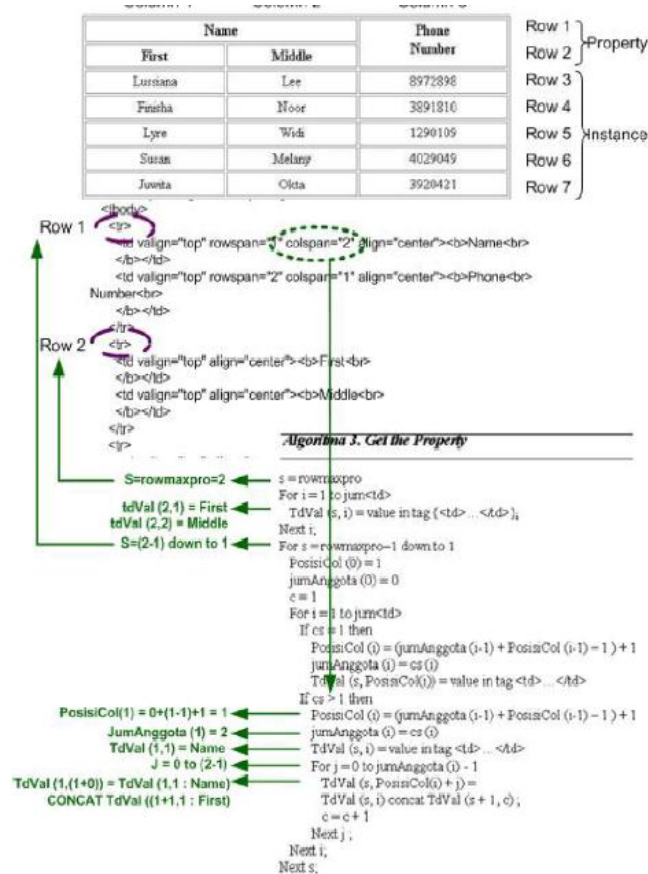


Figure 2-2: Illustration for Algorithm 3 of HTW

Lerman et. al. (Lerman et. al., 2001) proposed the method which groups the rows and columns for extracting data from lists and tables. The technique identifies the lists by computing the page template. To compute the web page template, it tokenizes the whole page into single words. The sequence of tokens appearing once on each page is found by the template finding algorithm. It starts with the smallest page and takes the first token from it. A sequence of tokens is then found, which appears on each web page. The lists

contain more than two rows so they cannot be a part of the template. From this inference the lists and tables are detected from the page. After identifying the list portion of the page, the data is extracted from it. The data cells are generally separated by the separators. Columns are identified by similarity of content, layout and separators. The extracts are clustered on the basis of common separators. The extracts which share at least one separator are grouped in same cluster. It is more evident that the extracts of a cluster are from same column. In next step DataPro algorithm is used to get the pattern of each cluster and each extract is checked similarity of extract with any of the cluster. If the extract is described by any of the pattern of nth cluster, the nth content feature is set to 1. To cluster the extracts, AutoClass is used. Generally the rows have the same type of data. Grammar induction algorithm ALGERIA is used for identifying rows.

Gultom et al. (Gultom, Sari, & Budiardjo, 2011) have proposed a mashup system Xtractorz, which first extracts table data from web pages and then integrates the data from multiple tables. Before integration, it performs data source modeling, data cleaning and data visualization. HTML code of the targeted web page is extracted and a DOM tree is created from it. Xtractorz runs a recursive algorithm which automatically identifies the tags of table from non-table code and create a DOM tree for it. From this tree, Xtractorz will retrieve the Columns of the table by reading the data in Xpath. In source and modeling stage, the DOM tree is converted into table in Excel and during this stage the data of the columns identified in previous stage is extracted. Xtractorz also performs data cleaning. It provides the cleaning rules in its GUI. User can click on any of the options like deletion of space, semicolon and percentage symbol etc. Xtractorz integrate the data in two tables if they share at-least one column. Then, one of the shared attributes is chosen as Primary key based upon characteristics. It compares the attributes in both tables. The common attributes appears once in resultant table and the unique columns are also included in the resultant table.

A lot of experimental data on materials is available on web and many problems are faced while transferring it into database due to the redundancy, diverse structure and isolation. Min et al. (Min, & Zhiyuan, 2017) has proposed a method which is extracting and storing the material data into the database. The table structure is analyzed by creating a parse tree

of table headers to transfer data into database. Rule based approach is used to create the tree. Before creating a parse tree for the headings, the authors has the described the table format they are considering in detail. Only the two dimensional tables with multiple line header rows are used in this research. The location of cells in table is determined by the row and column numbers. They assumed that the cells can have only textual data and cells can be merged. The table cells are divided into two classes, entry and label. The label area is assumed to be independent. Label area has headings that can be turned into a tree structure. It is assumed that no cell in the label area is blank and if found a one, that will be merged with the nearby filled cell. A parse tree is built for the headings describing the relationships between the headings. Duplicate columns are also handled in this process. After that data is inserted at leaves of the tree by locating their respective headings. a number is also saved along with the data to maintain the order of the data. Tables are created before moving the data into the database. The experiments are conducted on 10000 tables. 2620 of them have empty cells so excluded from data set. The results of querying are compared with the original algorithms. It is found that the speed of data insertion query execution is 10 times slower than and speed of query execution is much faster than original algorithm.

In 2015, Akbar et al. have proposed a method/technique to integrate the web tables. It is three a three phase/stage process. In first stage structure/schema of the table is extracted. The table extraction process is further divided into four steps, discovering the table location in a web page, segmenting headings and data values; discovering of table layout; and transforming the table into a tree structure. Table location is determined by identifying <table> tag in HTML document. The tables with the minimum size of 3x3, is considered for integration process. The size of the table is calculated by count the number of <td>/<th> tags in first row and no of <tr> rows in HTML document. <td>s are calculated for number of columns nad <tr> are calculated for no. of rows. Attributes and data values are segmented by considering the syntax of the table. <th> tahs in first row contains the headings and the all the <td> tags in remaining <tr> tags contains the data values. If no separate row for the headings, then it performs (1) pre-processing (2) visual coherency (3) syntactic coherency (4) semantic coherency and (5) post processing. The

output of this phase is a tree structure of the web table.

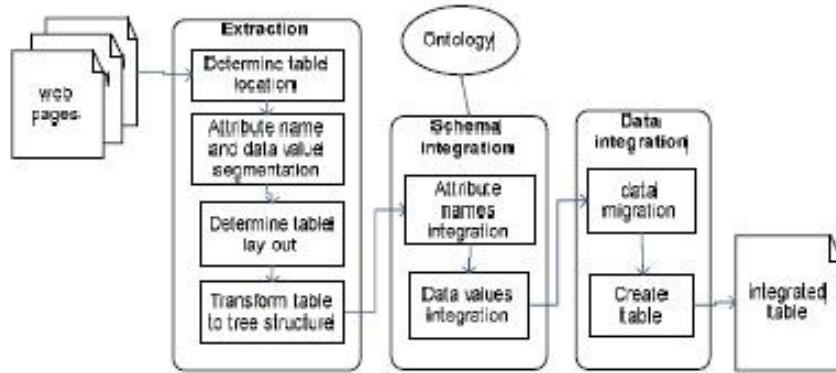


Figure 2-3: Process of integrating Web Tables on web page (Akbar et al., 2015)

In second phase, schema is integrated. This process integrates the trees of both webpage into one. There are two steps in this process, schema integration and data integration. For schema integration, edit distance of two attributes is calculated. If it is less than 50%, then it is checked whether they are synonym to each other or not. Synonyms are found using domain ontology. If they are synonym, integrated into one attribute. A standard format is created for the data values in the data integration process. Third phase is to integrate the data. During the data values integration, duplication is removed by using vector space model. Meanwhile, the integrated tree is transformed into the HTML code. The technique can combine two HTML tables into a single HTML and display the results on web page. The technique is tested on tables of three domains, student thesis, car rental and university rating. Total 8 websites are taken from all these domains and tables of same domain are integrated. The results shows that only three attributes are not identified because they are not found in ontology. Accuracy of results is not yet calculated.

In 2011, Hao with other researchers (Hao et. al., 2011) presents a method which is able to extract the different verticals data without the re-implementation. Vertical here means a specific domain like book, restaurant or university etc. the technique accepts one labeled seed site along with the vertical related to a set of attributes. It is further given some more pages to training the system. The web pages are transformed into DOM tree and the leaf nodes (text nodes) are input to the system. It is a three steps process. First step is feature extraction. The leaf nodes of the DOM tree contain the attributes names. The main focus

is to extract the attributes from these text nodes. From a website related to a specific vertical, a set of pages are parsed into DOM trees to get the text nodes. From these text nodes, three types of features, layout features, content features and context features are extracted.

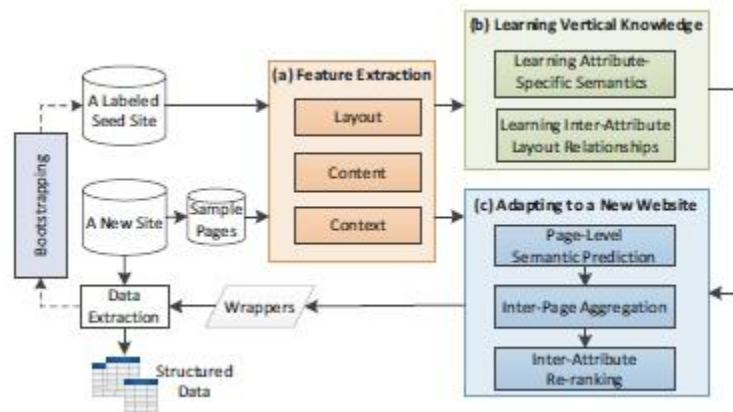


Figure 2-4: Flowchart of proposed technique (Hao et. al., 2011)

Second step is to learn the knowledge about the vertical using all the features extracted in the first step. In this step, attribute specific semantics and inter-attribute layout is determined. In the third step, unseen websites are given as input and the acquired knowledge is adapted to the site by manipulating site-level information. The result is the automatic identification of attributes values for further wrapper induction.

To validate the solution, the technique is run on 124K web pages obtained from 80 websites. The verticals of these websites are: autos, books, camera, jobs, movies, NBA players, restaurants and universities. These verticals are very diverse in nature. 10 websites for each vertical are searched and from these websites, 200 pages each are identified. A set of common attributes and regular expressions for ground truth is also prepared. The results are promising. The authors have proposed to improve the technique by enhancing the solution to combine it with data extraction, by developing a model that can reduce the inter-site gaps in layout and content.

Trinity is a technique of extracting web data proposed by Sleiman and Corchuelo (Sleiman & Corchuelo, 2014). The technique is basically building a trinary tree of the HTML code of the web page and extracting the data from it. Now, most of the

websites/web documents are generated using a same server-side template. This technique is using this fact and works on one or more such web documents and generates a regular expression that is modeling such web documents. This model is later used to extract the data from similar web pages. The basis of this research is a hypothesis; it states that templates have such shared patterns which are not meaningful. So these patterns can easily be ignored without any loss of information.

The first step of this algorithm is to build a trinary tree T of the web document. As the name depicts, in this tree a node can have three child nodes prefix, separator and suffix. Each node is a tuple of the form (T, a, p, e, s) . In the first step a root node is created with the input document and value of s is set to max. The algorithm searches for the shared pattern in the document, if found a new node with three children is created. Along with the tree construction, these nodes are being analyzed to get new shared pattern. This is a recursive process.

After the trinary tree construction, regular expression is formulated. The tree is traversed in pre-order to learn a regular expression. When it reaches at leaf node which is different from the shared pattern achieved in previous step, it means it's the data which is to be extracted and it output it. Otherwise it will output the shared pattern corresponding to the node being analyzed.

On web a lot of information is displayed in the forms of lists. Chu et al. (Chu, He, Chakrabarti, & Ganjam, 2015) has proposed a method in 2015 to convert such lists to relational tables. In lists, generally a record lies in one line separated by any delimiter e.g. space, comma etc. First of all, the records are needed to be separated into atomic values. This is done by tokenization. The line is tokenized based on the delimiters specified by the user. After tokenizing the lines, it needed to separate into values. The token are arranged into columns by looking into their semantics. The whole technique based upon the observation that the values of same column are semantically and syntactically coherent. The semantic coherence is checked by utilizing a corpus of more than 100 million web tables. It is checked that how often two tokens came together in same column. The method is consecutively assigning tokens to a fixed number of segments, with the goal that the whole of coherence score over all sets of qualities in a

similar segment can be augmented. The goodness of the segments is also being analyzed in parallel. Meanwhile, an additional algorithm is build up using thoughts from A* search that can prune away unpromising divisions without affecting quality. Two versions of algorithms are prepared supervised and unsupervised. Unsupervised algorithm has no involvement of user and supervised do involve the users during the segmentation process. The results are compared with two similar purpose algorithm ListExtract() and Judie(). To conduct the experiments, two data sets are prepared. 100 million tables are taken from a document index of a search engine and 500,000 spreadsheet tables from an IT company. The results show that the TEGRA performs significantly better than both the ListExtract and Judie. Moreover, supervised algorithm is showing more potential than unsupervised algorithm. The authors are interested to further investigate this fact.

To query the image web tables, they need to be transformed into a uniform framework. Embley et al. (Embley et. al., 2016) proposed an algorithm HITs for end-to end processing of large number of image tables based upon characterization of header indexed tables. The algorithm proposed in this research is converting the header-indexed table into a category table that can easily be queried by many data stores. The tables regions are segmented by using the distinct indexing of the data areas/regions by header path. They have uses the fact approved by prior work that the image document can be converted into a searchable database by analyzing its physical and logical layout. Physical analysis is assigning literal content to the cells laid out on the grid. Indexing relationship between the headier cells and data cells is determined by the logical analysis. To get the Physical layout of ASCII, scanned and PDF tables, they are first converted into a grid of cells. This grid representation has no explicit information about header row, data cells, footnotes etc. It also does not provide any division between data cells and other parts of table.

After the physical layout, logical layout is analyzed. For all cells in the grid representation, indexing structure for categories and an orderly list of category paths for each data cell is revealed. Heading are indexed using the critical regions, CC1, CC2, CC3 and CC4. CC1 and CC2 are calculated using MIPS (Embley et. al., 2011). CC1 and CC2 are used to identify the stub heading and CC3 and CC4 are used to identify the data cells.

Regional level constraints are handled by using block algebra and cell level constraints are also handled. After the segmentation into header segments and data cell segments, data cells are classified with their corresponding heading. After this, categories are analyzed and a category table is formed. The last step of this algorithm is to produce a group of fact assertion, characterized as relational database tables and also as subject-predicate-object triples in a semantic-web standard.

The experiments are conducted on 200 web tables obtained from a statistical website and 200 spreadsheet tables. The technique was able to correctly produce 197 web tables out of 200. The error is either due to the one of trivial website and wrong calculation of CC2 by MIPS. From 200 spreadsheet tables, 9 tables have errors in critical cells. Out of nine, only 1 table is violating the basic rules of algorithm because it has repeated headings. Query validation is also performed to prove the claim. Authors are keen to update the HITs to fully interpret semantics and syntax of tables, convert egregious tables into HITs, and integrate the interpreted tables into ontologies.

Storing the data of web tables into the database can be really helpful in many ways. To populate a web table (Shaukat et. al., 2016) proposed a method. The HTML tables written with <table> tags are taken in this research. After the data set preparation, Schema is extracted using rule learning or conditional random field (CRF). In first step, from HTML tables relational and non-relational tables are classified. The tables with the relevant data are relation and the tables which have some extra details like audios, videos and simple text and non-relational tables. The classifier, which classifies the tables as relational or non-relational, is trained by CRF. Once the relational tables are identified, header rows are determined. In most of the cases, first row is the header row and the cells in the first row contain the headings. After heading extraction, the domain of the tables is matched by using corpus-based techniques top train the classifier using the table titles. The next step after domain matching is Schema matching. For this purpose, Mapped Knowledge Based (MKB) approach is used. A dictionary is created in the database with all possible keywords in it. The headings of the table are then matched with the dictionary elements. If matched move to the next stage.

After schema matching, data is populated by joining related tables of HTML pages into a

single HTML table. The name is taken as the Primary Key for the table. The reason of this selection is that name appears in almost every website. This is the limitation of this approach as name can have duplicate values. Before insertion, name of new table is compared with the names already inserted into table, if match is found the missing data is appended in that record otherwise a new record is inserted. The results show that, if schema is matched correctly, then 87% tables are populated from 9800 total tables.

2.1 Comparison of Schema Extraction Techniques

The above explained schema extraction techniques differ in many ways from each other. These techniques have adopted different methodologies; operate on different types of file formats and data domains. Comparison of these techniques is presented in the Table 2-1.

Sr. No.	Paper	Schema Extraction	Missing Headings	Without Headings	Data Integration
1	Lerman et. al., 2001	Yes	Yes	Yes	No
2	Zhai et al., 2005	Yes	No	No	No
3	Gultom et. al., 2011	Yes	No	No	Yes
4	Hao et. al., 2011	Yes	Yes	Yes	No
5	Adelfio et. al., 2013	Yes	No	No	No
6	Nagy et. al., 2014	Yes	No	No	No
7	Sleiman et. al., 2014	Yes	-	-	No
8	Purnamasari et. al., 2015	Yes	No	No	No
9	Akbar et. al., 2015	Yes	-	-	Yes
10	Chu et. al., 2015)	Yes	Yes	Yes	No
11	Embley et. al., 2016	Yes	No	No	No
12	Shaukat et. al., 2016	Yes	No	No	Yes
13	Min et. al., 2017	Yes	-	-	No

Table 2.1: Comparison of Schema Extraction Techniques

Most of the techniques can extract the schema of the web data with the proper header row but only some of these can predict the headings for missing headings and without headings web tables. These are the techniques which are basically designed to extract the schema from the lists data available on web and convert it into the table. Only three techniques can integrate the data, but they create an integrated web table. Rest of the techniques mentioned the integration as their future work.

CHAPTER 3

3. METHODOLOGY

In this chapter, the adopted methodology used to accomplish the research objective is discussed. Research objective is to extract web table data from multiple websites and store it into a single database. To achieve the integrated database from multiple web tables, first of all the HTML code portion of tables needs to be extracted from websites. Only the data from a specific type of web tables (Table 1.1) is extracted and stored in database. This will narrow down the selection of websites for data extraction.

In next step, from the extracted code, the schema is extracted using the technique HTW (HTML Table Wrapper) discussed by (Purnamasri et. al., 2015). We have selected this technique as a base for our implementation because it is simple, easy to implement and relatively new. However, this technique has got certain shortcomings that need to be addressed prior to using it for the scenario which is the target of our research work. These shortcomings have been listed below:

- The authors have tested their technique only on the artificial html codes of different table's forms. The technique has not been implemented on the actual websites tables.
- Another shortcoming of the technique is that it only extracts the properties (attributes) of web table; it does not extract data from it.
- The technique is also not creating any database table of extracted properties.
- This technique lacks in extracting the schema of such web tables which do not have explicit heading row.
- The technique fails with the tables having extra caption rows before heading rows.
- Some web tables have missing headings.

HTW is chosen in this research because it is relatively new and has room for a lot of modification some of which are mentioned above. In this research, the technique is not only implemented to extract the attributes headings, but also the data from the web table. It is also moving the data of web table into the database table. The problem of missing

headings is also resolved in this research. Following is the Architecture diagram of the proposed system;

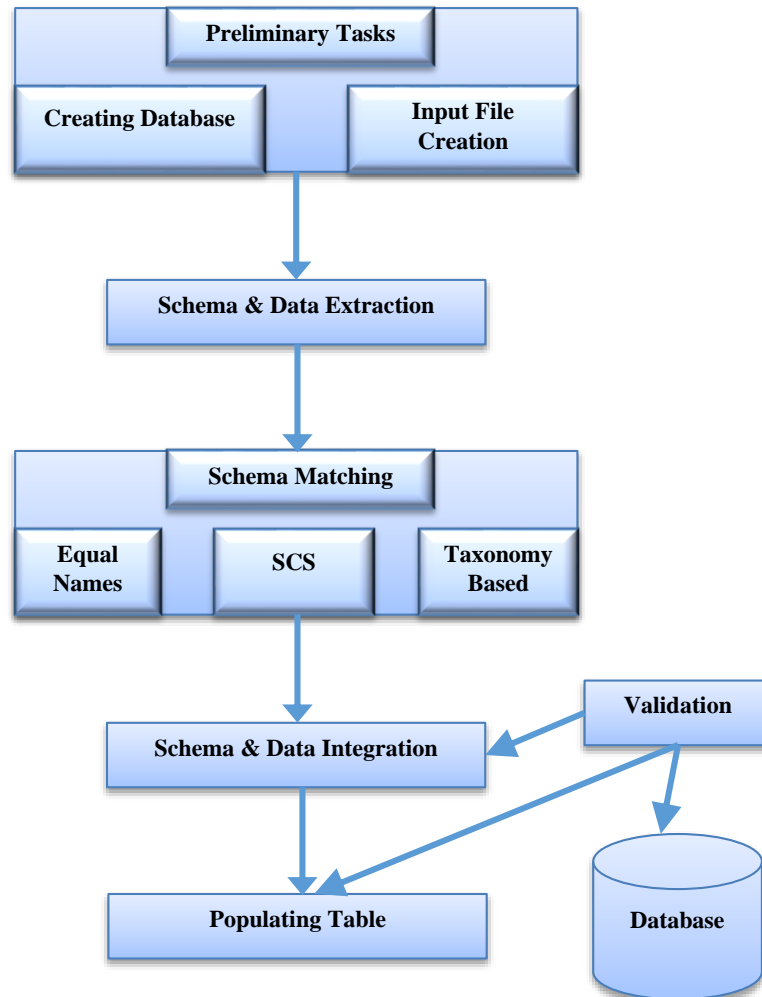


Figure 3-1: Architecture Diagram of Proposed System

While extracting the values, missing values issue is also resolved. Before moving the data into an integrated table, extracted schema is matched with the already created table, “Master Table”. Master Table is created to store the data of all the websites of same domain. All the attributes and data are stored in an array called Data array before moving it into the Master Table. Schema is matched using three different schema matching algorithms, Equal Names (EN), Similarity based on Common Substring (SCS) and using an already built Taxonomy. Following flowchart depicts the working of schema matching algorithms:

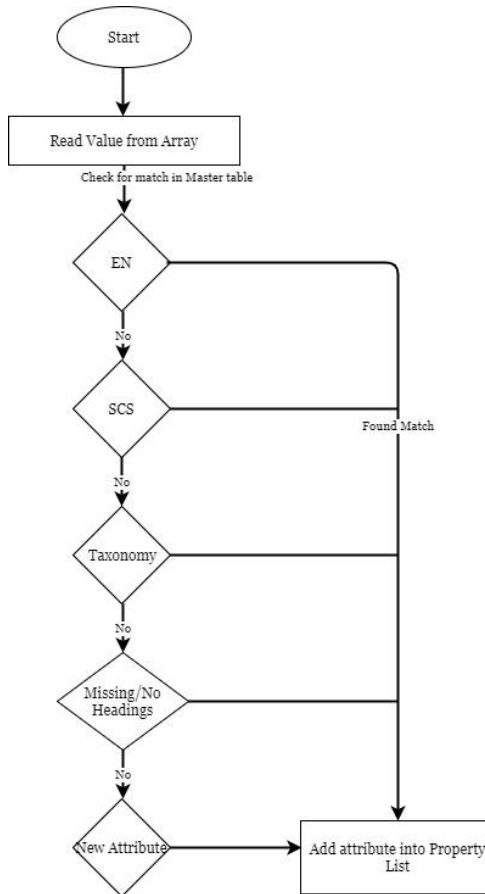


Figure 3-2: Flow chart for Schema Matching Algorithm

In first step the extracted attribute is matched in Master Table using Equal Names, if matched stored in a separate array called Property and started matching next attribute. If the attribute is not matched with any of attributes in Master table using EN, it is passed to Similarity based on Common Substring algorithm. If not matched then passed it to the Taxonomy algorithm. If the attribute is not matched using any of the above three methods, it is tested for missing/no headings. If it is not then it is interpreted as a new attribute or a synonym of any attribute. If it is a new attribute, master table is updated to add it. And the concerning tables are also modified. If it's a synonym, added to taxonomy against the attribute.

Once the schema is matched, the data is transferred into the table. Implicit data integration is performed in this research. Master Table is formed and all the data from different websites is stored in it.

3.1 Data Domain

The domain that we have selected for our research is faculty data from Higher Education Institutes (HEIs). Data on faculty members is required for different purposes, like to find people having a particular research interest, certain degree or searching with name. Similarly, while selecting the institute to get admission, students are more interested in checking the profiles of the faculty members of the HEI. Faculty's profiles play really important role in getting the enrollment in institutes. To check the faculty members, a student has to navigate through all the universities websites one by one. If the information/data available on universities websites can be stored in a single database, that can help students and other stakeholders in different ways.

In this research, the faculty information available on universities websites in a particular table format is being stored in a single database. The websites of universities, with the table format Table. 1.1, are chosen to extract the data. The specific table format is chosen due to the limitation of the HTW, which can only be applied to the table format shown in Table 1.1. The list of the universities with faculty information in this format is given Appendix A.

3.2 Preprocessing

Before the actual implementation, some preprocessing steps were required in order to get started. It was required to prepare proper HTML code files and needed to build a database to store the data and relevant details.

3.2.1 Data Collection

Before the schema extraction technique is actually implemented, the input text file needs to be prepared. The purpose of this step is to prepare an input file which contains only the required HTML code. All the unnecessary code is removed. We are only interested in that portion of HTML code which contains the table with multiple rows containing faculty information. A set of web tables is prepared manually by browsing through different universities website. Google Fusion Tables is a trial information perception web application to accumulate, represent, and share information tables. This web application is also used to search the academic staff details of universities with and without

mentioning the country name. The HTML code of table portion from the website is copied into a text file using SnappySnippet. The file is saved with the name of the university. Data Collection is manual and time consuming. All the text files are placed in a folder and given as input to the Schema Extraction technique.

3.2.2 Global Database Creation

A database is constructed to store the extracted schema and data. Following tables are required for the successful implementation of the proposed algorithm;

- **Master** (Sr No, Name, Designation, Qualification, Contacts_Details, Extention, Res_interest, Experience, Picture, Room_No, Job_Status, Specialization, Profile, Office_Hours, Publications, Building, Mailbox, Symbol)
- **Attributes** (AttCode, AttName)
- **Universities** (UniID, City, UniName, Country, Nick)
- **Departments** (DeptID, DeptName)
- **Taxonomy** (Syn_No, AttCode, Synonyms)

3.3 Proposed Methodology

At the end of this research we will be able to answer three research questions. This will be our research contribution in this particular research activity which falls in the schema extraction domain.

In the following, research questions targeted in this research and the strategy adopted to answer each of them have been described.

3.3.1 Research Question 1

Can the HTW be enhanced to extract data from real web tables?

HTW has been applied on the artificial web tables and the data is also not extracted. This question is answered by extracting schema and data, using table wrapper technique, of 121 web tables and stored in the Master table. While storing the data values, empty data cells are also handled.

3.3.1.1 Schema Extraction

Schema is extracted using a wrapper induction method HTW (HTML Table wrapper) (Purnamasari et. al., 2015). The approach works on the assumption that the first row of the table contains the heading of the table and rest of the rows contains the data against each heading. As described earlier, the technique comprises three algorithms which have been implemented in this work. Output of first algorithm is number of rows and columns of input table, second algorithm calculates the header row boundary and in the end using the values of first and second algorithms third algorithm will return the N headings (attributes names) of the table. The names extracted by third algorithm are stored in first N positions of an array. The third algorithm has been further modified to extract data from the web tables; this aspect was missing in the original algorithm of HTW. Algorithms 1 and 2 are implemented as are proposed in HTW. The algorithm 3 is modified to extract data along with the property.

This function returns the text in the string removing all tags including <td>. The modified algorithm is shown in Figure 3-5.

HTW+ is calculating number of rows and columns of the input web table and after removing all extra tags it is reading the text in the first row of the table and storing it into the array.

3.3.1.2 Data Extraction

In this research, the technique is extended by extracting the data. The same technique of third algorithm of HTW is used to extract the data values from the HTML code and stored in same array in which attributes names are saved. HTML code from text file is read line by line. All the tags other than <td>...</td> tag are ignored. From the cropped line of <td> tag, all the enclosed formatting tags mentioned above are removed and only the text is stored in the array.

It is assumed that if there are N attributes, and then first N positions of array have attribute's names and the next N position have the values of first record against each corresponding attribute and so on. After the Schema Extraction, the array contains the complete table's records along with attributes.

```

-----Results-----
NO. of Columns: 3
NO. of Rows: 10

Schema Extracted

*****Extracted Data and Schema*****
Sr. No
Name
Designation
1
Mr. Farhan S.Sherazi
Head of Department
2
Dr. Nadim Asif
Professor
3
Dr. Muhammad Aasim Qureshi
Senior Assistant Professor
4
Mr. Tahir Iqbal
Senior Assistant Professor
5
Mr. Taimoor Aamer Chughtai
Lecturer
6
Mr. Muhammad Dawood Akram
Lecturer
7
Ms. Munaza Sher
Lecturer
8
Ms. Fakhra Batool
Lab Engineer
9
Ms. Sarah Chaudhry
Lab Engineer

```

Figure 3-3: Output of Data Extraction Algorithm

While reading the data values, some issues can arise. Most important of all is the missing values.

3.3.1.3 Missing Values

In tables, some of records may have missing values. Missing values means some of data cells are empty against some of attributes. As it can be seen in the Figure 3.4, the status of the faculty members is missing in most of the records.

If the missing values are not taken into account, they can cause irregularity while storing the data from array into database. The sequence of the values can change. The disturbed sequence of data can negate our basic assumption about the data stored in array.

The screenshot shows a web application interface for the Department of Computer Science. It features a search bar and a table with the following data:

S.No.	Name	Current Designation	Status
1	Mr.Muhammad Siddiq	Asst. Prof.	Incharge
2	Dr. Muhammad Sarim	Asst. Prof.	Registrar & Director I.T.
3	Dr. Kamran Ahsan	Asst. Prof.	Director QEC
4	Dr. Adnan Nadeem (http://www.researchgate.net/profile/Adnan_Nadeem2)	Asst. Prof.	
5	Dr. Syed Akhter Raza	Asst. Prof.	
6	Dr. Arzoo Ateeq	Asst. Prof.	
7	Ms. Asima Nisar	Asst. Prof.	
8	Mr. Farhan Shafiq	Asst. Prof.	
9	Ms. Uzma Afzal	Asst. Prof.	
10	Mr. Shaikh Kashif Riffat	Asst. Prof.	On Study Leave

Figure 3-4: Example of web table with Missing Values

To handle this discrepancy, while reading data records, if a <td></td> tags combination found which has no data values; “Empty” string is stored in the array. Now the sequence of data values is according to the attributes in first n indexes of array.

The modified algorithm HTW+ is as given below;

```

Algorithm 3: Get the Property and Data
Input: .txt file, cs, rowspan
Output: TdVal[], Property[]

S = rowspan
For i= 1* HTML code line to endlime do
  Read HTML code line
  If line.startswith("<td>") && !line.endswith("</td>")
    Read and concatenate the next lines until "</td>" found
  Remove all tags from line
  If line is empty || NULL
    TdVal[i]="No record in this data cell found"
  Else
    Trim the line for extra spaces/tabs
    TdVal[i]=line
  Next i
If cs==1
  For j=1 to TCol
    Property[j]=TdVal[j]
If cs>1
  .....

```

Figure 3-5: Modified Algorithm 3 of HTW+

In start the algorithm is reading the .txt file line by line, if it is a pair of <td>...</td> tag then it reads the text within the tag ignoring all other tags and storing it into array.

The answer to RQ1 in this research is summarized as follows: the HTW (Purnamasri et. al., 2015) can be utilized for real web tables with some modifications that are necessary to handle certain issues that are faced in real web tables as explained above. The original algorithm presents an approach for data extraction that has been tested on artificial web tables rather than the real ones. This is a major drawback, as the real web tables have more variation in syntax and formats, which has been resolved in the current research.

3.3.2 Research Question 2

Can the HTW be enhanced to extract the schema of web tables with missing headings and without header row?

There is a lot of information given on many websites in form of tables but those tables do not have explicit heading row. The basic assumption of the technique proposed in HTW, i.e., the first row of the table contains the headings/attributes names, is not applicable on

such tables. This research is enhancing the technique by extracting the schema of such tables. Attributes names are extracted by using the values of the tables in each cell of first row. For this purpose, some additional taxonomies like for designations, qualification and research interests, are also built. Stanford library for natural language processing⁴ is used to check whether the string is the name of a person or not.

First two algorithms work fine for such tables too, that is, the first one calculates the number of column correctly which help us in determine how many attributes the table has. And the same value is used to execute the loop to predict the attributes. Figure 3-6 shows the algorithm for predicting headings for the without heading table:

The algorithm receives a string and checks using Stanford library⁵ that whether it is a person's name or not. If it is person's name, Name attribute is stored in Property array.

Email and phone number both are stored in a single attribute Contacts_Details. To check whether it is an email address it is checked that the string contains @ symbol in it. To check the phone number, all the spaces, brackets and all signs like + are removed. The original string is also checked for the alphabets. After removing all unwanted characters from string, it is checked if it is a number and its length is greater than 6 because a phone number can be of minimum 7 digits. It is also labelled as Contacts_Details in Property array.

⁴ <https://nlp.stanford.edu/software/CRF-NER.shtml>

Algorithm: Get the Property of Tables without Headings

Input: DMstr, Valueindex, index, Desig_Taxonomy, Qual_Taxonomy,
ResInt_Taxonomy, TCol

Output: Attribute added in Property array

Begin:

```
If DMstr.contains('@') || DMstr.contains(numbers of 7 or more digits)
    Property[index++]=Contact_Details
else if DMstr.contains(numbers of less than 7 digits)
    Property[index++]=Extension
else if DMstr is identified as person
    Property[index++]=Name
else if DMstr is found in Desig_Taxonomy
    Property[index++]=Designation
else if DMstr is found in Qual_Taxonomy
    Property[index++]=Qualification
else if DMstr is found in ResInt_Taxonomy
    Property[index++]=Research_Interests
else if DMstr is a number and number >0 && <=70
    Property[index++]=Experience
else if DMstr contains Alphanumeric values without brackets, braces
    Property[index++]=Room_No
else if DMstr="Empty"
    Recursive call for TdValue[Valueindex+TCol]
End if
```

Figure 3-6: Algorithm to Predict Headings

For this algorithm, three taxonomies are made for Qualification, Designation and Research Interests respectively. As the headings are being predicted by using the values and the above mentioned attributes can have many values according to the domain selected, so their taxonomies are created. Qualification contains the list of all possible qualification that a person can have to be eligible for working as faculty member in computer science department in all the universities all over the world. Designation contains all the possible designations a person can hold in the academic department. Computer science is a very diverse field. It has so many sub areas like Database, software engineering, computer networks, artificial intelligence etc. all of these sub fields too have so many sub areas/disciplines of work. A lot of research activities in all these disciplines of computer science are being conducted. So list of research interests is really long. Research interests taxonomy is based upon the “**The 2012 ACM Computing Classification System toc**”⁵. Table 3.1 presents the above explained taxonomies.

⁵ <http://www.acm.org/about/class/class/2012>

Attributes	Taxonomies
Qualification	"P.Hd", "Masters", "MCS", "BSCS", "BS", "BS(CS)", "BSSE", "M.Tech", "Bachelors", "MS", "MIT", "BIT", "BS(SE)", "BS(CE)", "BSCE"
Designation	"HOD", "Associate Professor", "AP", "Assistant Professor", "Lecturer", "Instructor", "Junior Lecturer", "Research Assistant", "Research Associate", "Teaching Fellow", "Professor", "Senior Lecturer", "Guest faculty", "Head", "Teaching Assistant for Computer Science"
Research Interests	<p>.....</p> <p>Architecture</p> <p> Serial Architectures</p> <p> Reduced instruction set computing</p> <p> Complex instruction set computing</p> <p> Superscalar architectures</p> <p> Pipeline computing</p> <p> </p> <p>Networks</p> <p> Network architectures</p> <p> Network design principles</p> <p> Layering</p> <p> Naming and addressing</p> <p> Programming interfaces</p> <p> Network protocols</p> <p> </p>

Table 3.1: Taxonomies

The string is searched in Taxonomies build for Qualification, Designation and Research Interests, if found in any of the taxonomies labeled accordingly in Property array.

Many websites display the office extensions of the faculty member in the faculty list. It is observed from the data that the extension is generally a 4 or 5 digit number. So if the string contains only numbers and it is less than 7 digit number, it is interpreted as extension,

The room numbers in universities are mostly alphanumeric. Characters/alphabets represent the building name and number is the room number in that building. This assumption is used to check for the room numbers.

If a column is added in a table, it means it has some values in some of records if its first value is empty. In this case, the legal value is searched by adding the TCol value into the index of the current value which is empty. The process is repeated until we reached the value from which its attribute name can be guessed.

The sequence of the attribute is maintained according to their appearance in the web table to avoid any discrepancies during data insertion.

3.3.2.1 Other Attributes Prediction

In case of an attribute for which the logic for its data values is not include in algorithm, its value is compared with all the values for each attribute already added in database using Equal Name matching algorithm. If the exact match is found in any of the attribute, the attribute name is copied into the Property array. If it is not matched in any of the attribute values, it is asked from the administrator/user to add it whether as a new attribute or as a value against already included attributes.

3.3.3 Research Question 3

Can we use the schema extracted from multiple web tables to build a schema integration tool?

A Master (Global) table is built in advance containing possible attributes to store faculty's data. When the process of schema and data extraction from the web tables is completed for a certain university, the schema is matched with the master table by the below mentioned methods. After schema matching, data is transferred into Master table. The process is repeated for all the web tables obtained from different universities websites. At the end, master table has the combined data for all the university in a single table.

3.3.3.1 Schema Matching

After the attributes are extracted and stored in an array, the next step is to move them into the database. Before moving them into the database, they need to be matched with the attributes in Master Table. As explained above, Master Table has possible attributes related to faculty members of universities. There are 19 attributes in Master Table. These are the common attributes observed on different universities websites. Three schema matching algorithms are applied on the attributes. A main Matcher function is written in which all the three matching algorithms are implemented. Three algorithms are written for all three techniques and called in main matcher function "Match". In Match function,

first of all an attribute is matched using Equal Name, in case it does not match with any of Master table attributes, it is checked whether this is a string with multiple words. If it is such a string it is passed to the substring function otherwise Taxonomy function is called for this attribute.

In schema matching methods strings are compared with each other. To compare two strings, Levenshtein Edit Distance is used. It calculates the number of changes required to convert first string into second string. The methodology of three techniques is explained in next section.

- **Equal Names**

In Equal Names, two strings are compared and if they exactly match with each other returns true. All the extracted attributes are matched one by one with all the attributes of Master Table. The attributes of the Master table are gotten by fetching the schema of the Master table created in SQL Server. The matched attributes are stored in a separate array. The unmatched attributes are passed to the next Schema Matching algorithm. The sequence of the attributes is maintained to avoid the discrepancies while transferring data into Master Table.

The common issue which is faced while integrating the data from multiple web tables is the semantic heterogeneity. A diverse nature of variety is found in the names of the attributes used on different websites for same attribute. The other two schema matching methods are written to handle such heterogeneities.

- **Similarity based on Common Substring (SCS)**

If an attribute is not matched in Master Table using Equal Name, it is passed to Similarity based on Common Substring. In this research this technique is used with a little modification. For example, an attribute's name is "Designation Name", it cannot be matched using Equal Name and if there is no such word added in taxonomy so it cannot be mapped to any of the Master table attribute. Although it can be seen that this attribute can be mapped to "Designation" attribute of Master Table. Now if we can break this string into words, then there is a possibility that we will be able to map it to any of the

Master Table's attribute. So if a string contains more than one word it is break into words.

The method for Substring matching will receive the string as an argument and split it into words on basis of spaces and store into array. Now the elements of the array are compared with attributes of Master Table one by one. As soon as any of the word is matched with any of the attribute of Master Table, the loops breaks and return true. Before exiting, the algorithm replaces the string from the array with the name of the corresponding matched attribute.

Algorithm: Similarity Based upon Common Sub-String Matching Algorithm

```

Input: Var, Property[], Data[], TCol
Output: Property[]

Begin:

Words[]=var.Split(" ")
Connect with database
DataTable SchemaTable= Execute SQL query to select the schema of the Master table
For j=0 to words.Count
    Foreach colrow in SchemaTable.Rows
        Var1= colrow.field("CoulmnName")
        Len= CalcLevenshtein(Var, Syn)
        half=Syn.Length/2
        If Len < half
            Property.Add(Var1)
            found = true
        End if
    End foreach
    If found != true
        Tax_rtn = Taxono(words[j])
        If Tax_rtn==true
            break
        End if
    End if
End for

```

Figure 3-7: Similarity Based Upon Common Sub-String Algorithm

There is a possibility that the attribute is not matched in Master table but can match in Taxonomy. For example if we have string "Position of Employee", Position is not matched with any of the Master table attributes but it is synonym of "Designation" and we have it in Taxonomy against "Designation". If the word is not matched in Master table it is searched in Taxonomy by calling the Taxonomy function. If found in taxonomy its corresponding attribute is placed in Attribute's array and the function will return true to the Sub string function and which return that value to the main matched function.

- **Taxonomy**

If the attribute is not matched using Equal Name or Similarity based on Common Substring, the attribute is passed to the third algorithm which is taxonomy. Taxonomy is a table in database containing the synonyms and alternate names for the attributes of Master Table as described earlier. For example, Qualification of a faculty member can be titled as “Qualification”, “Last Degree”, “Education” and much more.

This algorithm also receives a string (attribute name) from the data array. This string is compared with the values of Synonym attribute from Taxonomy table. If the string is matched with any of the value in Taxonomy table, its corresponding Att_Code is noted.

```
Algorithm: Taxonomy Matching Algorithm
```

```
Input: Var, University Database
Output: Synonym is placed in Property[]

Begin:
Connect with database
Execute SQL query to select the data from the Taxonomy table
While rdr.read()
    Attcode = rdr["AttCode"]
    Syn = rdr[Synonym]
    Len= CalcLevenshtein(Var, Syn)
    Quarter=Syn.Length/4
    If Len < quarter
        Find the corresponding attribute of synonym
        Store the attribute in Property[]
    End if
```

Figure 3-8: Finding attribute in Taxonomy

In next step the noted attribute code is used to search corresponding attribute name in Attribute table. After finding the attribute name, the passed string is replaced by this name.

If the attribute is matched, the loop will continue to match the next attribute until all the attributes get matched. If it returns false, it means we are unable to find the match with all three algorithm.

3.3.3.2 Missing Headings

If the table has some of the missing headings then those cannot be found by any of the above mentioned schema matching algorithms. In place of the attribute name there will be string “Empty” at that index of Data array. To solve the issue the technique used to solve missing headings is also used here. Data_Matcher algorithm is called here by passing the corresponding value against this attribute by adding TCol value into the index

of the Empty string. This process is repeated until the valid value is obtained to compare in the algorithm.

After running all of the above algorithms, if still the attribute is not identified then it means either it's a new attribute or synonym of an attribute.

3.3.3.3 Not Matched Attributes

Now here, three interpretations arise. The value which is not found it can be a new attribute or synonym/ alternate name for any existing attribute or a value. Now the administrator will decide what this value exactly is.

- **New Attribute/Synonym**

If an attribute is not matched by any of three Schema Method methods, it can be either a new attribute or synonym/alternate name of any attribute. At this point the administrator is asked to choose between the two situations. Based upon the admin choice, next action is performed. If it is a new attribute, administrator is asked to enter the suitable data type for attribute. After getting the data type, a new attribute is inserted in both Master table and in Attribute Table with attribute code.

If it is synonym/alternate name, the admin is asked to enter the attribute code from the attribute's list displayed earlier on screen. After getting the attribute code, the value is inserted into the Taxonomy table along with other necessary field's values. The process is not creating any type of redundancy in the database.

3.3.3.4 Schema Integration

Schema is being integrated as new websites are being processed. As explained above, one table with all possible attributes is created. All the attributes extracted from different websites are compared with the attributes of this table. If matched, it means now the data can be inserted into the table. If not matched, it is included into this table as a new attribute.

So schema is not integrated explicitly, it is performed implicitly every time a new website is being processed.

- **Populating Tables**

Once the schema is matched, the next step is to populate the table with the data of websites analogous to all the extracted attributes. The data from all websites is moved into the Master Table, resulting in an integrated table.

After schema matching following heterogeneity issues were faced.

- Issue 1: While building taxonomy, due to the common observation among websites, Title is made synonym for Name attribute. But in some of the websites, Title is used to show the Designation of faculty member. In this case, during the matching process, Title is interpreted as Name although it is storing the Designations. It results with two attributes with the Name caption.
- Issue 2: In some of web tables, name is stored as first name and last name. Now both these attributes are added as synonym of Name attribute so two attributes with same caption exist in the table.
- Issue 3: In the start, two separate attributes were declared for Email and phone number. Contacts Details is made synonym of Phone number on the basis of observation. But in later stages, it is observed that some of the websites have the attribute Contact details having both email and phone number or only email addresses. So, a single attribute Contacts_Details is added into the Master table removing both Email and Phone number. Both these attributes are included into taxonomy as synonym. Now the issue arises when a website has both Email and Phone number separately included in table. So two attributes are mapped to Contacts_Details which causes issues in insertion query.

All of the above issues are related to one another in some way. Now we will discuss the methodology to solve the above stated issues.

Taxonomy for Designation is built in this research for the missing headings. This taxonomy is used to solve the issue 1. After schema matching, an algorithm is executed to check whether the Property array has more than one attributes captioned as Name. If so, their corresponding values of first record form the Data array are compared with the Designation Taxonomy. If the values got matched with any of the value in taxonomy, the attribute name is replaced with Designation.

Issue 2 and 3 are solved by one method. After checking for the Designation, The Property array is checked for the duplicate values. If the array has duplicate values, then the values which are duplicated to each other are merged together. If we merge the duplicates in Property array, we need to merge their corresponding values in each record in Data array. After merging, the repeated values are deleted from both data and Property arrays.

3.3.3.5 Data Insertion into Database Table

After resolving all of the above issues, data is transferred from the Data array into the database table. The data array has all the records fetched from the HTML code of the given website. After schema matching, all the attributes names are stored in Property array and data is present in the Data array. If the website table has n columns than first n positions have the attributes name. Next n positions of array have the values against each of the attribute for the first record and so on. Missing values are also handled by adding “Empty” string in array.

To move the data from array into the Master Table, first N positions of array are ignored as they contain the attributes name which we have already separated in Property array. Starting from the n+1 position of array, next n elements of the array are stored in the database table. The sequence of the values is same as of the attributes. From the name of the University, its city and Uni_ID is extracted from the University table and inserted in the Master table. In this research, only adding the data of computer science faculty, Dept_ID will be 1 for all the universities. SQL Queries used to insert data into Master Table are shown in Figure 3-9.

SQL Queries to insert data from array to Master Table

```
for (int loop = TCol; loop < result-1; loop = loop + TCol)
{
    if (Property[0].ToString() == "Sr_No")
    {
        string query = "INSERT INTO Master(UniID,DeptID,City," + Property[1] + "," + Property[2] + "," +
Property[3] + ") VALUES (" + uni_id + "," + dept_id + "," + city + "," + Data[1 + loop] + "," +
Data[2 + loop] + "," + Data[3 + loop] + ")";
    }
    else
    {
        string query = "INSERT INTO Master(UniID,DeptID,City," + arr2[0] + "," + arr2[1] + "," + arr2[2] +
"," + arr2[3] + ") VALUES (" + uni_id + "," + dept_id + "," + city + "," + arr[0 + loop] + "," + arr[1
+ loop] + "," + arr[2 + loop] + "," + arr[3 + loop] + ")";
    }

    SqlCommand cmd3 = new SqlCommand(query, conn);
    cmd3.Connection = conn;
    rdr3 = cmd3.ExecuteReader();
    rdr3.Close();
}
```

Figure 3-9: Insert Query

Argument of the INSERT query depends upon the no. of attributes. For all possible number of attributes, SQL queries are written.

In this chapter, methodologies used to solve the identified gaps in HTW, is discussed. The issue of missing data is resolved by storing “Empty” string at the respective position of the cell in the array. Missing headings and without header table’s schema is predicted by analyzing the values of the cells. To integrate the data in Master Table, first its schema needs to be matched with Master Table’s schema. Three schema matching algorithms are used; Equal Names, SCS and Taxonomy. The attributes are compared in sequence retrieved from the table, maintaining the order. During the heterogeneity analysis differences like duplicate attributes and naming conflicts etc. are identified and resolved. After resolving conflicts, data is transferred into the Master Table of Database.

CHAPTER 4

4. RESULTS & EVALUATIONS

In this chapter, the results obtained by applying the methodology are discussed. The technique extracts the schema and data of both with and without header row tables and integrates the data into a single database. Primary objective of this methodology was to extract schema of actual web tables and integrating the data from multiple web tables.

Different tasks are performed to achieve the methodology like literature review, pre-processing of HTML tables, schema and data extraction and schema matching to obtain integrated data.

The technique is applied on faculty data of computer science department, of 134 universities websites. Websites differ from each other in many perspectives but they have a common feature that all are displaying the faculty data in desired tabular format. Despite the similar format, tables on different websites have some different features than others. Some of these features are;

- Some of the web tables have empty cells in header row but corresponding data cells contains data values.
- In some web tables, there are empty data cells.
- There are such web tables which have merged rows in between data rows.
- Some of the tables have a caption/title row before the heading row, labeling the title of the table.
- Some web tables do not have explicit header row.

Table 1.1, Table 1.2 and Table 1.3 has example of some websites with above mentioned hitches. In this research HTW is not only applied on real web tables but also modified to HTW+ to solve most of the above stated issues.

To evaluate the technique, three methods of evaluation are used;

- Quantitative
- Comparison with prior techniques
- Query Validation

In following section of this chapter the results of all three methods of evaluation are given.

4.1 Quantitative Analysis

In this chapter I will check how precise and accurate HTW and HTW+ are. Four quantitative measures; Precision (Zhai et. al, 2005; Purnamasri et. al., 2015), Recall (Zhai et. al, 2005; Purnamasri et.al., 2015) and F-measure (Zhai et. al, 2005; Purnamasri et.al., 2015) are used to evaluate the results of proposed approach.

Precision defines how precise our findings are? Precision measures how close the measured values are to each other? The formula used to calculate precision is;

$$Precision = \frac{Correctly\ Extracted}{Correctly\ Extracted + Incorrectly\ Extracted} * 100$$

Recall measures the completeness of the approach. Recall tells us how much the approach has actually found from all that it is supposed to find.

$$Recall = \frac{Correctly\ Extracted}{Ground\ Truth} * 100$$

F-Measure is the average of precision and recall.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

In the following section above formulas are applied on HTW and HTW+ to evaluate their performance.

4.1.1 Research Questions

After performing the experiments, we will be able to answer three research questions. This is our contribution in the area of schema extraction and integration. Following research questions are answered in this research:

4.1.1.1 Research Question 1

Can the HTW be enhanced to extract data from real web tables?

The schema extraction methods (HTW and HTW+) takes a text file containing the HTML code of web table of single university faculty data. The approaches are applied on 120 universities web tables.

The HTW (HTML Table Wrapper) (Purnamasari et. al., 2015) and HTW+ calculates the number of rows and columns in the given web table, maximum value of colspan attribute's value in <td> tags of 1st row and then extract the schema and data of the table.

- **Schema Extraction**

As HTW was originally applied only on artificial web tables, so in this research it is applied on 120 real web tables. The same tables are the input of the HTW+. Amongst these 120 web tables, containing 106 tables with proper header row, 9 are the tables with some missing headings, and 5 web tables are with title/caption row and merged data.

The technique performs better for the web tables with explicit header row. In case of empty headings, it is unable to guess/extract the headings. It completely fails for the without header row tables. There are some tables with merged rows; HTW (Purnamasari et. al., 2015) cannot handle such tables as well. To overcome these gaps, the technique is enhanced and named as HTW+. Table 4.1 given below shows results of the HTW and HTW+ applied on real web tables. In Table 4-1, Uni_Name has the nick names of the universities from which web tables are taken, Actual Attributes column is showing the number columns in the web tables, Corr. Extr. Column contains the number of attributes correctly extracted by the technique, Wr. Extr. Stands for the wrongly extracted attributes and Not Extracted are the number of attributes which both techniques are unable to extract.

Sr No	Uni_Name	Actual Attributes	HTW			HTW+		
			Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn	4	4			4		
2	AIOU	3	3			3		
3	AUP	3	3			3		
.
.
.
120	Awkumshankar	2	1	0	1	1	0	1
121	Awkumtimer	2	1	0	1	1	0	1
	Total	479	426	17	38	441	2	38

Table 4.1: Results of HTW and HTW+ applied on actual web tables

Complete Table 4-1 of results HTW and HTW+ is included in appendix B.

The HTW is working fine for the web tables with proper heading row. However, it fails to identify the headings of such tables which have one or more empty heading cells as shown in Figure 4-1. The technique doesn't even consider it as empty cell and ignores it and does not store any value against this attribute in the array in which we are storing all the data from web tables. As the number of columns N are calculated correctly, so while retrieving the headings from the array, it will count first N indexes values of array as headings. The consequence of this appears at the time of data insertion in the database as it will read some of the data values as column headings.

Name	DesignationName	Email ID	
<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value=""/>	
MUHAMAMD NAVEED ALAM	HOD	nalam@numl.edu.pk	View Profile
Dr M Hanif Zaouq	Assistant Professor	mhzaouq@numl.edu.pk	View Profile
MOHAMMAD RAZA PERWEZ	Assistant Professor	raza@numl.edu.pk	View Profile
DR. FAZLI SUBHAN	Assistant Professor	fazli.subhan@numl.edu.pk	View Profile
SAJJAD HAIDER	Assistant Professor	sajjad@numl.edu.pk	View Profile
KH.MOYEEZULLAH GHORI	Assistant Professor	mghouri@numl.edu.pk	View Profile
HINA ALI	Assistant Professor	hali@numl.edu.pk	View Profile
ATA ULLAH GHAFOR	Assistant Professor	aullah@numl.edu.pk	View Profile
DR NOMAN MALIK	Assistant Professor	mnauman@numl.edu.pk	View Profile
Dr Huma Hayat Khan	Assistant Professor	abc@numl.edu.pk	View Profile
FOUZIA JAMAL GOREJA	Assistant Professor	fjgoreja@numl.edu.pk	View Profile
ABDUL KALEEM	Lecturer	akaleem@numl.edu.pk	View Profile

Figure 4-1: Snapshot of Faculty information from NUML university website

In case of table format shown in Figure 4-1, first three headings will be read correctly and “Mohamad Naveed Alam” will be read as the 4th heading which is incorrect. Due to this error, the sequence of the data is affected while moving data from array into database table. Moreover, one incorrect label will be entered as an attribute heading. HTW+ is handling this situation by storing “Empty” in place of empty cells in array. Now the sequence of the data will not get disturbed.

In case of merged rows containing title/caption of the table before heading row or present within the data records, HTW and HTW+ both fails. They do extract the schema of the table and store it into array along with the text of merged rows but due to the failure of the basic assumption, it consider only the title row as heading and considers the headings as the data records against this title. This is because of number of rows are calculated correctly but it calculates number of columns 0 or 1 which is incorrect. In case of 0, it assumes that there is no data in the table. For number of column 1, title is taken as heading of the attribute so it considers each value at every index of array as data value against the title.

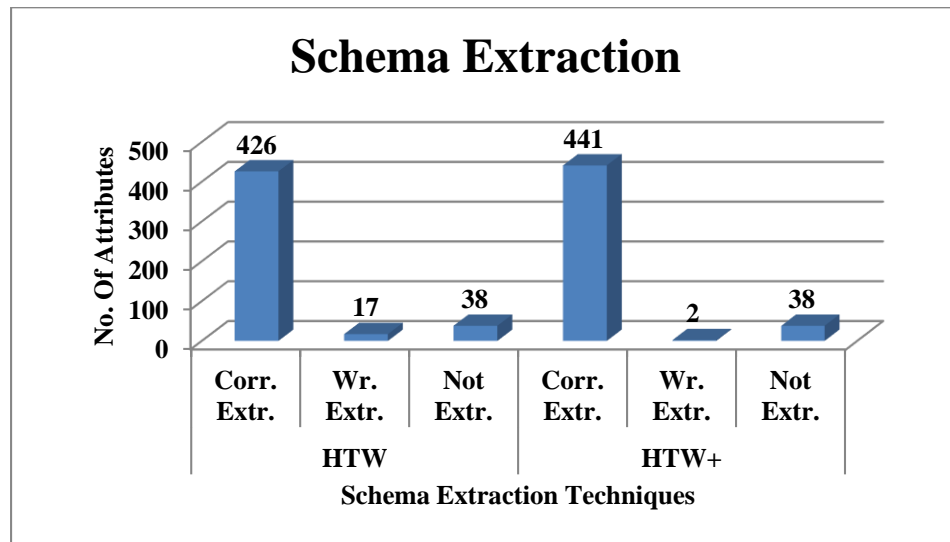


Figure 4-2: Result of schema extraction technique

Amongst the 120 web tables, technique is successfully extracting the schema of the 107 web tables. 13 web tables have either merged table title row or all the attributes are confined in a single <td>...</td> tag. Out of 479 attributes of all web tables, HTW+ is able to extract 441 as compared to 426 extracted by HTW.

	HTW	HTW+
Precision	96.25%	99.56%
Recall	89.16%	92.23%
F-Measure	92.57%	95.75%

Table 4.2: Comparison of HTW and HTW+

The results show that the technique is more precise with the 99.56% precision as compared to HTW. High precision relates to the low incorrectly extracted rate. With the recall of 92.23%, it can be interpreted as the technique is able to correctly identify maximum attributes of a given web page.

- **Data Extraction**

HTW only extracts the attributes names. It does not extract the data from the HTML code of web table. In this research, the technique is modified to extract data as well. The data of the web table is initially stored in the same array in which the attributes headings are stored. HTW+ is also storing the empty data cell. Following Table 4.3 shows the results of data extraction:

Sr No	Uni Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn			✓	✓		
2	AIOU			✓	✓		
3	AUP			✓	✓		
.
.
.
119	Awkumtimer			✓	✓		
120	uttarauniversity			✓	✓		
Total		0	0	120	111	9	0

Table 4.3: Data Extraction results

Complete Table 4-3 is given in Appendix C.

The table shows that the HTW is unable to extract the data of all the input websites. On

the other hand, HTW+ is extracting the data of 111 websites correctly, while it is unable to extract the data in sequence of 9 websites.

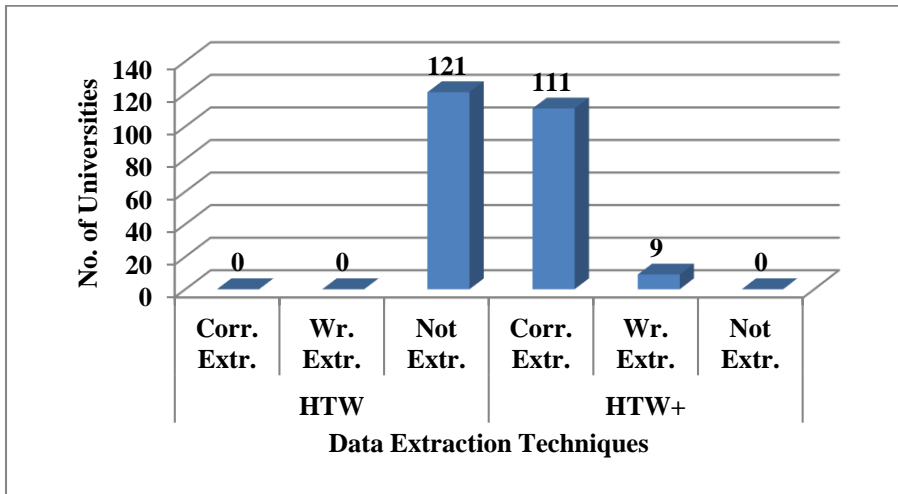


Figure 4-3: Comparison of Data Extraction results

The incorrect extraction of the data from 9 websites is due to either title merged row or merged rows in data portion. The data of 111 website is being read and stored in array in order and the empty data cells are also being handled.

HTW+	
Precision	92.56%
Recall	92.56%
F-Measure	92.56%

Table 4.4: Comparison of Data Extraction Results

Precision, recall and f-measure all are equal because the data of all 120 websites is extracted but among them of 9 websites data is incorrectly extracted. But overall the technique is quite precise and accurate with 92.56% precision and recall.

- **Schema and Data Extraction Combined**

Proposed method is extracting schema and data for most of the web tables included in data set but still there are some website for which schema is extracted but their data is not being extracted. Also, for some of the web tables proposed technique is unable to extract both schema and data. Following table shows the combined results of schema and data extraction of both techniques;

Sr No	Uni_Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn			✓	✓		
2	AIOU			✓	✓		
.
.
.
46	nottingham_UK					✓	
.
.
.
120	Awkumtmer			✓		✓	
		0	1	120	104	16	0

Table 4.5: Results of Schema and Data Extraction

Complete version of Table 4-5 is given in Appendix D. The above table is summarized in the following figure.

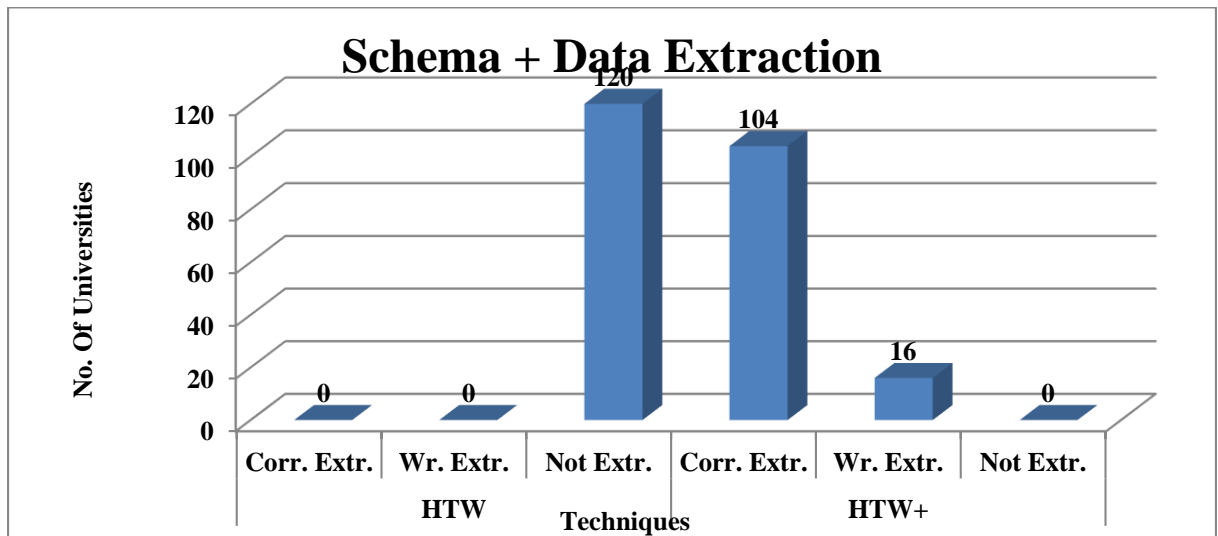


Figure 4-4: Results of Schema and Data Extraction Combined

For some of the websites, schema is being extracted but its data is not extracted and for some schema is not extracted correctly but data is extracted correctly. The web tables obtained from official websites of all campuses of Nottingham University, schema of these websites is extracted correctly but its data portion has some merged rows in it so the

data obtained has some extra rows in it due to which the sequence of data against the attributes is incorrect. Similarly, the web tables obtained from the official websites of all campuses of Abdul Wali Khan University, has only a single row in header portion of table. There are 2 attributes in all these web tables given in one line in one <th> tag. Its schema is not extracted correctly due to this reason, but its data is extracted well in order and correct.

HTW+	
Precision	86.77%
Recall	86.77%
F-Measure	86.77%

Table 4.6: Schema and Data Extraction Combined

The technique is precise with 86.77% precision. The same value for precision and recall is the depiction of the fact that the technique is extracting all attributes and few of them are wrongly extracted.

4.1.1.2 Research Question 2

Can the HTW be enhanced to extract the schema of web tables with missing headings and without header row?

The basic assumption of HTW fails when it comes to the table with no explicit header row. In such tables, first row the data row. HTW+ is enhanced to predict the headings of such tables by considering its values. There are total 14 without headings web tables

- **Schema Extraction**

HTW reads the first row as heading, but in case of without heading tables, it reads first data row as heading. These headings cannot be matched using any of the schema matching technique used in this research. Such values are passed to the Data Matcher algorithm and their appropriate heading is detected. Table 4.7 is showing the results of the algorithm.

Sr No	Uni_Name	Actual Attributes	HTW			HTW+		
			Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	KUST	3	0	3		3		
2	Sbbwu	2	0	2		1	1	
3	Emps	2	0	2		1	1	
.	
.	
.	
12	Uom	3	0	3		3		
13	Nwu	3	0	3		2	1	
14	Vub	3	0	3		2	1	
	Sum	45	0	45	0	41	4	0

Table 4.7: Schema Extraction of Without Header Table

Complete Table 4-7 is given in Appendix E

It can be seen from the table out of 45 web tables without headings, schema of 41 is being extracted correctly. Following diagram shows the above scenario graphically;

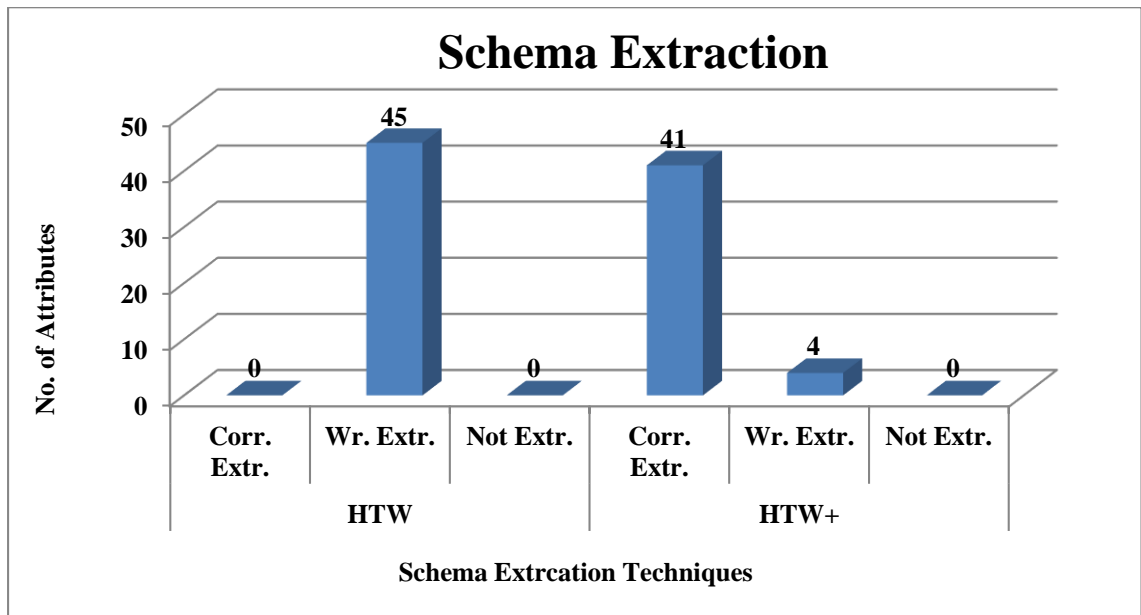


Figure 4-5: Schema Extraction of Without Header Row Table

The four attributes from four different web tables, which are not being detected, are those which have diverse information in one attribute. For example, the web table of Shahid Benizir Bhutto Women University, which is not being detected correctly, is the one

which has designation and qualification together in same cell. Similarly, North Western University, Khulna Bangladesh has name and designation in same cell and Victoria University of Bangladesh has name, designation and contact details in one cell. These values are needed to be separated and stored in their respective attributes.

HTW+	
Precision	91.1%
Recall	91.1%
F-Measure	91.1%

Table 4.8: Schema and Data Extraction of Without Header Row Tables

The technique is able to predict the attributes by their values with the precision, recall, f-measure and accuracy of 91.1%.

- **Data Extraction**

Data extraction is one of the major contributions of this research in the subject area. As HTW is modified to extract the data of the web tables along with the schema so the data of the without headings web table is also extracted. The first row in this case also data row. The technique is successfully extracting data rows and storing them into the data array. Table given below shows the statistics of HTW and HTW+ Data Extraction.

Sr No	Uni Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	KUST			✓	✓		
2	Sbbwu			✓	✓		
3	Emps			✓	✓		
.
.
.
13	Nwu			✓	✓		
14	vub			✓	✓		
Total	14	0	0	14	13	1	0

Table 4.9: Data Extraction of Without Header Row Tables

Complete Table 4-9 is given in Appendix F.

Table 4-9 is summarized in the following Figure;

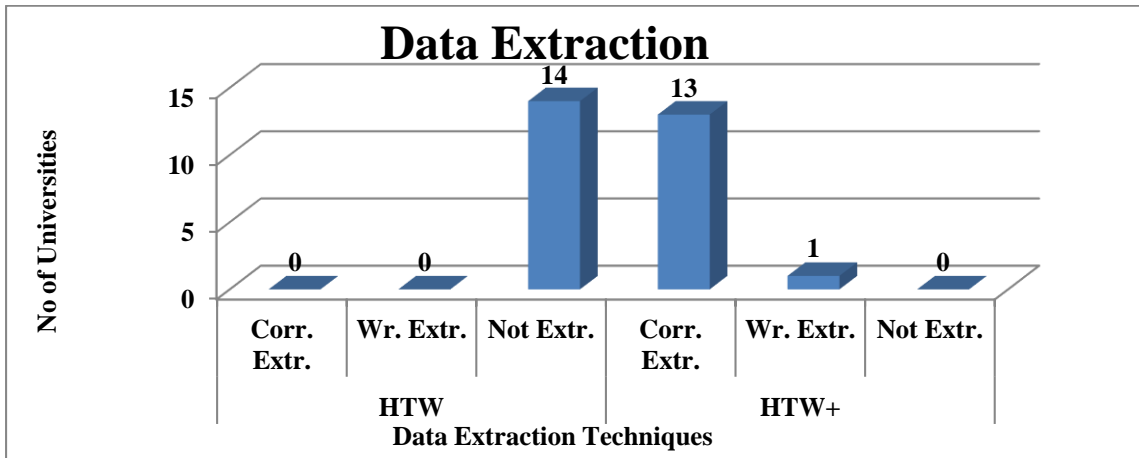


Figure 4-6: Data Extraction of Without Header Row Table

HTW is unable to extract the data of none of the input web tables. On the other hand, HTW+ is extracting data of 13 web tables. Technique is unable to extract the data values of only one website.

HTW+	
Precision	92.58%
Recall	92.58%
F-Measure	92.58%

Table 4.10: Experimental Results of SE of Without Header Row Table

From above table it can be seen that the technique is correctly extracting data of website for which the schema is already extracted.

4.1.1.3 Research Question 3

Can we use the schema extracted from multiple web tables to build a schema integration tool?

To answer this research question, the technique is extended to integrate the above 135 web tables. To integrate the tables, firstly their schema is matched.

- **Schema Matching**

Three methods are used for schema matching, Equal Names, substring Matching and

compared in Taxonomy. The algorithm receives an array containing the names of the attributes used in the website. There are some websites which do not have explicit headings; the given technique is also suggesting the names of the attributes by using its values.

All the attributes firstly matched with Equal Names. If not matched with Equal Names and the string has multiple words, it is matched using Sub-String. If still not matched, it is searched in Taxonomy.

There are total 20 attributes in Master table of the database and 435 attributes are in the web tables collectively. Following table shows the number of attributes matched using each of Schema Matching algorithms.

Equal Name	Common Sub-String Matching	Taxonomy
142	55	166

Table 4.11: Experimental results of Schema Matching Algorithms

Table 4.12: Schema Matching Algorithms

The graphical representation of the above data is as given:

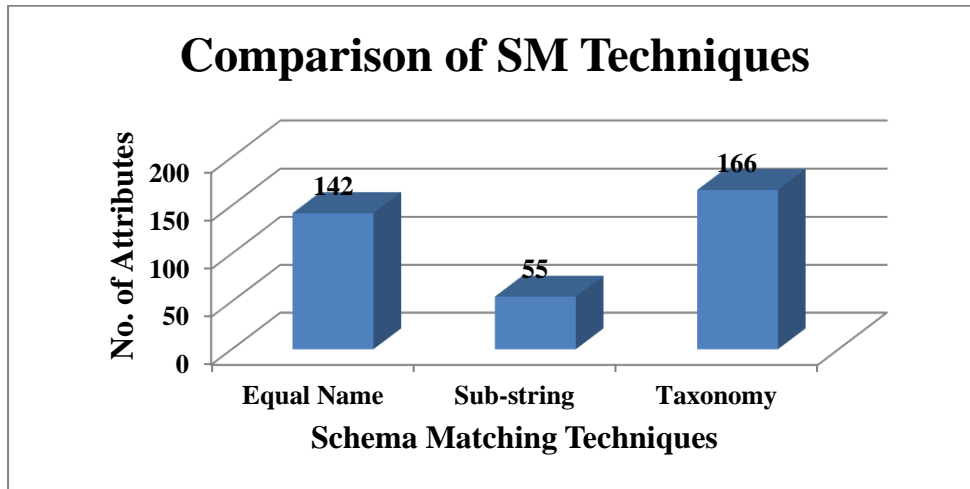


Figure 4-7: Graphical Representation of experimental results of SM Algorithms

As the graph shows that the most of the attributes of maximum websites are matched with in built Taxonomy. Second most used technique is the Equal Names and SCS matching is used only to match 55 attributes.

- **Data Integration**

After schema matching, the Property array has the names of the attributes extracted in the desired format and it is ready to send the data into the Master Table. DML SQL statements are used to insert the data into the database. The data is transferred directly to the master table without inserting it into any intermediate table.

Data set contains HTML code of 134 web tables with and without header row. From these 134 web tables, data of 125 web tables is successfully extracted and stored into the array. Among these 125 web tables, 111 are the web tables in the exact required format and 14 are the without headings web tables. From the 125 data extracted web tables, 113 web tables data is inserted into Master Table. The data of one without header row web table cannot be inserted into the database due to the disturbance in sequence.

HTW+	
Precision	99.12%
Recall	84.32%
F-Measure	91.12%

Table 4.12: Data Insertion

4.2 Comparison with Other Techniques

The proposed technique, HTW+ is compared with the other technique of schema extraction from web data. Following table is summarizing the comparison of the proposed approach with other schema extraction methodologies;

Sr. No.	Paper	Schema Extraction	Missing Headings	Without Headings	Data Integration
1	Lerman et. al., 2001	Yes	Yes	Yes	No
2	Zhai et al., 2005	Yes	No	No	No
3	Gultom et. al., 2011	Yes	No	No	Yes
4	Hao et. al., 2011	Yes	Yes	Yes	No
5	Adelfio et. al., 2013	Yes	No	No	No
6	Nagy et. al., 2014	Yes	No	No	No
7	Sleiman et. al., 2014	Yes	-	-	No
8	Purnamasari et. al., 2015	Yes	No	No	No
9	Akbar et. al., 2015	Yes	-	-	Yes
10	Chu et. al., 2015)	Yes	Yes	Yes	No
11	Embley et. al., 2016	Yes	No	No	No
12	Shaukat et. al., 2016	Yes	No	No	Yes
13	Min et. al., 2017	Yes	-	-	No
14	HTW+	Yes	Yes	Yes	Yes

Table 4.133: Comparison of schema extraction techniques

The above table shows that some of the techniques can predict the headings of missing headings and without headings tables. These are the techniques which are actually designed to convert the lists data available on web into relational tables. So, it can be interpreted that these can predict the headings for the table too. The techniques which are integrating the web tables: integrate the web tables and create another web table, not into database. HTW+ can predict the schema headings for both missing headings and without header row tables and is integrating the web tables from multiple resources into database.

4.3 Query Validation

After successful insertion of 101 web tables into the database, following SQL basic SQL queries are executed to validate the results of proposed methods.

1. Display all the records of Master table.

Following query is executed to produce the required results;

Select * From Master;

The output of the above query is;

Sr_No	Name	Designation	Qualification	Contacts_Details	Extension	Res_Interest
285	Prof. Dr. Abdus Salam	Head of Department	PhD International Islamic University Islamabad	NULL	NULL	NULL
286	Dr. Muhammad Arshad	Assistant Professor	Ph.D John Moores University Liverpool UK	NULL	NULL	NULL
287	Dr. Javed Iqbal Bangsh	Assistant Professor	PhD University of Teknologi Malaysia	NULL	NULL	NULL
288	Abdul Basit	Lecturer	PhD in progress Iqra University Islamabad. MS(Teleco...	NULL	NULL	NULL
289	Mr.Imran Khan	Lecturer	PhD in progress COMSATS Institute of Information Tec...	NULL	NULL	NULL
290	Mr. Tufail Muhammad	Lecturer	MS in System Engineering GIK Topi Swabi	NULL	NULL	NULL
291	Mr. Fahad Masood	Assistant Professor	MS (Telecommunicationand Networking) Gandharar U...	NULL	NULL	NULL
292	Ms. Hina Rabbani	Lecturer	M.Phil in progress Peshawar University. MSc in Maths ...	NULL	NULL	NULL
293	Mr. Saqib Shahid Rahim	Lecturer	MCS CECOS University Peshawar	NULL	NULL	NULL
294	Mr. Faseeh Ullah	Lecturer	PhD in progress from University of Teknologi Malaysia ...	NULL	NULL	NULL
295	Syed Aizaz Ul Haq	Lecturer	MSc. University of Sunderland UK	NULL	NULL	NULL
296	Mr. Janas Khan	Lecturer	MSCS COMSATS Institute of Information Technology ...	NULL	NULL	NULL
297	Mr. Shahid Muhammad Ali	Lecturer	M.Sc. Telecom-Engineering Sunderland University UK	NULL	NULL	NULL
298	Mr. Muhammad Usman	Lecturer	MS in Software Engineering	NULL	NULL	NULL
299	Mr. Hafiz Ullah	Lecturer	M.Phil in Mathematics University of Peshawar	NULL	NULL	NULL

Figure 4.6: Results of Query 1

After executing the query, the records of all the faculty members is displayed on output grid. The query will select all the columns of the “Master” table.

2. Display the Name, Designation and University name of all faculty members.

Following query is executed to know the University name of ‘Abdul Basit’

Select Master.Name, Master.Designation, Uni.UniName

From Master

INNER JOIN Uni ON Master.UniID=Uni.UniID;

The output of the second query is;

Name	Designation	UniName
Chia-Yen Chen	Lecturer	The University Of Auckland New Zealand
Associate Professor Patric...	Associate Professor	The University Of Auckland New Zealand
Paul Denny	Senior Tutor	The University Of Auckland New Zealand
Dr Michael J. Dinneen	Senior Lecturer	The University Of Auckland New Zealand
Professor Gill Dobbie	Professor	The University Of Auckland New Zealand
Professor Bob Doran	Emeritus Professor	The University Of Auckland New Zealand
Professor Alexei Drummond	Professor	The University Of Auckland New Zealand
Dr Matthew Egbert	Lecturer	The University Of Auckland New Zealand
Adriana Ferraro	Senior Tutor	The University Of Auckland New Zealand
Professor Mark Gahegan	Professor	The University Of Auckland New Zealand
Professor Georgy Gimelfarb	Professor	The University Of Auckland New Zealand
Professor Bakht Khoussain...	Professor	The University Of Auckland New Zealand
Dr Yun Sing Koh	Senior Lecturer	The University Of Auckland New Zealand
Associate Professor Sebas...	Associate Professor	The University Of Auckland New Zealand
Dr Simone Linz	Senior Lecturer	The University Of Auckland New Zealand
Dr Jiamou Liu	Lecturer	The University Of Auckland New Zealand
Dr Christof Lutteroth	Senior Lecturer	The University Of Auckland New Zealand
Dr Andrew Luxton-Reilly	Senior Lecturer	The University Of Auckland New Zealand
Dr Aniket Mahanti	Lecturer	The University Of Auckland New Zealand
Dr S Manoharan	Senior Lecturer	The University Of Auckland New Zealand
Dr Radu Nicolescu	Senior Lecturer	The University Of Auckland New Zealand

Figure 4.7: Result of Query 2

The query is displaying the Name, Designation and University Name of all added faculty members. It is an inner join query.

3. Display the Name, DeptID, UniID and Designation of all the faculty members whose research interest is Database.

Given query is executed to produce the required results;

```
Select M.Name, M.DeptID, M.UniID, M.Designation
From Master M
where M.Res_Interest like '%Database%' OR M.Specialization like
'%Database%';
```

The output of the third query is;

Name	DeptID	UniID	Designation
Mitch Chemiack Ph.D. Brown University Associate Professor and Undergraduate...	1	134	NULL
Olga Papaemmanouil Ph.D. Brown University Assistant Professor Computer Scie...	1	134	NULL
Dr. Mohammad Rezwanul HuqAssistant Professor	1	33	NULL
Michela Tauffer Associate Professor David L. and Beverly J.C. Mills Career Devel...	1	115	NULL

Figure 4.8: Output of Query 3

Above query is producing the list of teachers with research interest of Database.

The results of queries show that that data from the integrated table can be searched by applying any type of query.

The results of above all three validations methods shows that the HTW+ performance is quiet promising with the precision of 100% and accuracy of 85.12% and the technique is producing correct query results.

CHAPTER 5

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

Web is a source of huge amount of data in form of tables. These web tables need to be converted into the database table and integrating these web tables in one database in really helpful to execute ad-hoc queries. To store these web tables into DB, their schema is required to be extracted. Schema extraction of such tables is different from the schema extraction performed in conventional databases as no explicit schema definition is provided with such tables. HTW (HTML Table Wrapper) (Purnamasari et. al., 2015) [8] is a one of prior techniques, which only extracts the schema of the web tables of specific format and do not extract its data. The technique is also unable to detect the missing headings and data cells and hence not completely extracting complete information from tables. Another limitation of this technique is; it cannot extract the headings of the table in which headings are not explicitly given in the header section of <table> tag.

In this research, HTW is modified to HTW+ to overcome the above mentioned limitations and data from multiple web tables is integrated into a single database table. The values/text read in schema and data extraction is stored in an array. During the process of schema and data extraction to not lose the blank cells, whether they are in header row or in data row, “Empty” string is stored in array. All the extra tags are removed. After storing the whole table’s data into array, its schema is read from the first N indexes. In case of missing headings and without headings tables, data values at their corresponding indexes are used to predict the Property of the Attributes.

After schema and data extraction, extracted schema is matched with the Master table Schema using three methods, Equal Names, Similar Sub- String Matching and an already built Taxonomy. Experiments show that the taxonomy is best Most of the heterogeneity issues are resolved during the process of schema matching. Once the schema is matched, data is transferred from array into the Master table of database.

HTW+ is evaluated using three methods, (i) Comparison with the techniques in literature, (ii) Quantitative Analysis and (iii) Query validation. In comparison, other techniques are

analyzed to conclude which of features like schema extraction, missing headings, without headings and empty data cells are being considered in solution. For quantitative analysis, experiments are conducted on 135 web tables obtained from the official universities websites containing the information about computer science faculty. Among these 135 tables, 121 are tables with explicit header rows while 14 web tables are without headings. In 121 web tables, some tables have title row in the start of the tables, some have merged data rows and some of these have missing headings. The technique is extracting the schema of 108 web tables with the precision of 99.55% and 92.58% accuracy. Out of these 121 web tables, data of 101 tables is being transferred into the database. Some queries are also executed on integrated data to validate HTW+.

5.2 Future Work

In this thesis we have used web tables of specific format which has restricted the collection of dataset. Only 135 websites of this format are found after extensive search. If the technique can be modified to deal with more table format then dataset can be increased. In this thesis, only the data of faculty of computer science department of different universities is taken. Dataset can grow even more if other academic departments' data is also taken into account. During the process of schema and data extraction, merged title rows and data rows are not handled. Although, in this research heterogeneity issues are resolved but multi valued data cells needs to be resolute. Telephone number in some websites has extensions, which needs to be stored in extension instead of Contacts_Details. In the tabular format, very precise about information about the faculty members is given. Details are provided on linked pages. To get more benefits of integrated data, data from linked pages also required to be stored in database.

REFERENCES

- Adelfio, M. D., & Samet, H. (2013). Schema extraction for tabular data on the web. *Proceedings of the VLDB Endowment*, 6(6), 421-432.
- Akbar, M., Azizah, F. N., & Saptawati, G. P. (2015, November). Integration of HTML tables in web pages. In *Data and Software Engineering (ICoDSE), 2015 International Conference on*. IEEE. (pp. 132-137).
- Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM computing surveys (CSUR)*, 18(4), 323-364.
- Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., & Zhang, Y. (2008). Webttables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1), 538-549.
- Chu, X., He, Y., Chakrabarti, K., & Ganjam, K. (2015, May). Tegra: Table extraction by global record alignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM. (pp. 1713-1728).
- Cohen, W. W., Hurst, M., & Jensen, L. S. (2002, May). A flexible learning system for wrapping tables and lists in HTML documents. In *Proceedings of the 11th international conference on World Wide Web*. ACM. (pp. 232-241).
- Crescenzi, V., Mecca, G., & Merialdo, P. (2001, September). Roadrunner: Towards automatic data extraction from large web sites. In *VLDB (Vol. 1, pp. 109-118)*.
- Elmasri, R. Navathe, B. S. (2004) Fundamentals of Database Systems. *Addison-Wesley Longman*, Boston USA
- Elmagarmid, A. K., Rusinkiewicz, M., & Sheth, A. (Eds.). (1999). *Management of heterogeneous and autonomous database systems*. Morgan Kaufmann.
- Elmasri, R., Larson, J. A., & Navathe, S. B. (1986). Schema integration algorithms for federated databases and logical database design. Honeywell Computer Sciences Center.
- Embley W. D., Krishnamoorthy M., Nagy G., Seth S., (June, 2011). Factoring Web tables. In *Procs. EIA/AIE Conf.* (F. Esposito, S. Ferilli, eds.). ACM. p. 253-263.

- Embley, D. W., Krishnamoorthy, M. S., Nagy, G., & Seth, S. (2016). Converting heterogeneous statistical tables on the web to searchable databases. *International Journal on Document Analysis and Recognition (IJ DAR)*, 19(2), 119-138.
- Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-based systems*, 70, 301-323.
- Grannell, C. (2008). *The essential guide to CSS and HTML web design*. Apress.
- Gultom, R. A., Sari, R. F., & Budiardjo, B. (2011). Proposing the new Algorithm and Technique Development for Integrating Web Table Extraction and Building a Mashup. *Journal of Computer Science*, 7(2), 129.
- Hao, Q., Cai, R., Pang, Y., & Zhang, L. (2011, July). From one tree to a forest: a unified solution for structured web data extraction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. pp. 775-784
- Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118(1), 15-68.
- Laender, A. H., Ribeiro-Neto, B. A., Da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2), 84-93.
- Larson, J. A., Navathe, S. B., & Elmasri, R. (1989). A theory of attributed equivalence in databases with application to schema integration. *IEEE Transactions on software engineering*, 15(4), 449-463.
- Lerman, K., Knoblock, C., & Minton, S. (2001, August). Automatic data extraction from lists and tables in web sources. In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* (Vol. 98).
- Liu, B., Grossman, R., & Zhai, Y. (2003, August). Mining data records in web pages. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. pp. 601-606.
- Min, C., & Zhiyuan, M. (2017). Method of Understanding Structure and Building Database with Material Experiment Data. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1).

- Nagy, G., Embley, D. W., & Seth, S. (2014, April). End-to-end conversion of HTML tables for populating a relational database. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on* . IEEE. pp. 222-226.
- Purnamasari, D., Wicaksana, I. W. S., Harmanto, S., & Banowosari, L. Y. (2015). HTML table wrapper based on table components. *International Journal of Computer Applications in Technology*, 52(4), 237-243.
- Shaukat, K., Masood, N., Mehreen, S., Haider, F., Bakar, A., & Shaukat, U. (2016, December). Population of data in web-tables schema. In *Multi-Topic Conference (INMIC), 2016 19th International*. IEEE. pp. 1-6.
- Sleiman, H. A., & Corchuelo, R. (2014). Trinity: on using trinary trees for unsupervised web data extraction. *IEEE Transactions on Knowledge and Data Engineering*, 26(6), 1544-1556.
- Srivastava, D. (2010, October). Schema extraction. In *CIKM* (pp. 3-4).
- Wang, J., Wang, H., Wang, Z., & Zhu, K. (2012). Understanding tables on the web. *Conceptual Modeling*, 141-155.
- Zeeshanudin, S. (2011). A library of schema matching algorithms for dataspace management systems (Master of Science dissertation). A dissertation submitted to the University of Manchester.
- Zhai, Y., & Liu, B. (2005, May). Web data extraction based on partial tree alignment. In *Proceedings of the 14th international conference on World Wide Web*. ACM. pp. 76-85.

Appendix A

Sr. No	NickName	Uni Name	City	Country
1	Abasyn	Abasyn University Peshawar	Peshawar	Pakistan
2	AIOU	Allama Iqbal Open Univrsity Islamabad	Islamabad	Pakistan
3	AUP	The University Of Agricultural Peshawar, Pakistan	Peshawar	Pakistan
4	Awkumbuner	Abdul Wali Khan University Mardan Buner Campus	Buner	Pakistan
5	Awkumchitral	Abdul Wali Khan University Mardan Chtral Campus	Chitral	Pakistan
6	Awkumshankar	Abdul Wali Khan University Mardan Shankar Campus	Shankar	Pakistan
7	Awkumtimer	Abdul Wali Khan University Mardan Timer Campus	Mardan	Pakistan
8	Bahria_ISB	Bahria University Islamabad	Islamabad	Pakistan
9	Bahria_Kri	Bahria University Karachi	Karachi	Pakistan
10	Bahria_Lro	Bahria University Lahore	Lahore	Pakistan
11	CSIT	University Of Balochistan(Department Of Cs & It)	Quetta	Pakistan
12	Dadabhoy	Dadabhoy Institute Of Higher Education	Karachi,	Pakistan
13	fuuast	Federal Urdu University Of Arts, Science And Technology Karachi	Karachi	Pakistan
14	hiast	Hyderabad Institute Of Arts, Science And Technology	Hyderabad	Pakistan
15	icp	Islamic College Peshawar	Peshawar	Pakistan
16	iiu	International Islamic University Islamabad	Islamabad	Pakistan
17	inu	Iqra National University Peshawar	Peshawar	Pakistan
18	MUL	Minhaj University Lahore	Lahore	Pakistan
19	MUST	Mirpur University Of Science And Technology	Mirpur	Pakistan
20	northern	Northern University Nowshera	Nowshera	Pakistan
21	NUML	National University Of Modern Languages Islamabad	Islamabad	Pakistan
22	Riphah	Riphah International University Islamabad	Islamabad	Pakistan
23	Uos	University Of Sargodha	Sargodha	Pakistan
24	uoswabi	University Of Swabi	Swabi	Pakistan
25	ustb	University Of Science And Technology Bannu	Bannu	Pakistan
26	KUST	Kohat University of Science and Technology	Kohat	Pakistan
27	sbbwu	Shaheed Benazir Bhutto University	Peshawar	Pakistan
28	uoli	University of Loralai	Loralai	Pakistan

29	wuajk	Women University Of Azad Jamu And Kashmir Bagh	Kotli	Pakistan
30	iukb	Islamic University Kushtia Bangladesh	Kushtia	Bangladesh
31	bracu	Brac University	Dhaka	Bangladesh
32	ebaub	Exim Bank Agricultural University Bangladesh	Chapai Nawabganj	Bangladesh
33	ewubd	East West University	Dhaka	Bangladesh
34	hstu	Hajee Mohammad Danesh Science And Technology University	Dinajpur	Bangladesh
35	iiuc	International Islamic University Chittagong	Chittagong	Bangladesh
36	jnu	Jagannath University	Dhaka	Bangladesh
37	just	Jessore University Of Science And Trechnology	Jessore	Bangladesh
38	kuet	Khulna University Of Engioneering And Technology	Khulna	Bangladesh
39	manarat	Manarat International University	Dhaka	Bangladesh
40	nub	Northern University Bangladesh	Dhaka	Bangladesh
41	portcity	Port City International University	Chittagong	Bangladesh
42	pub	The People'S University Of Bangladesh	Dhaka	Bangladesh
43	rmstu	Rangamati Science Nad Technology University	Rangamati	Bangladesh
44	ruet	Rajshahi University Of Engineering And Technology	Rajshahi	Bangladesh
45	uoda	University Of Development Alternative	Dhaka	Bangladesh
46	vu	Varendra University	Rajshahi	Bangladesh
47	cusb	Central University Of South Bihar	Gaya And Patna	India
48	mgcub	Mahatma Gandhi Central University Bihar	Bihar	India
49	dimapur	Nagaland University	Bihar	India
50	tripurauniv	Tripura University	Nagaland	India
51	visvabharti	Visva Bharti	Suryamaninagar	India
52	hnbgu	Hemvati Nandan Bahuguna Garhwal University	Santiniketan	India
53	kakatiya	Kakatiya University	Srinagar	India
54	uma	University Of Mohaghegh Adabili	Telangana	Iran
55	fim	University Of Passau	Passau	Germany
56	le	University Of Leicester	Ardabil	United Kingdom
57	dmu	De Montfort University Leicester	Passau	United Kingdom
58	ed	The University Of Edinburgh	Leicester	United Kingdom
59	emps	University Of Exeter	Leicester	United Kingdom
60	aber	Aberystwyth University	Edingurgh	United Kingdom
61	uom	University Of Mauritius	Exeter	Mauritius
62	dur	Durham Univeraity	Llandinam	United Kingdom
63	kent	University Of Kent	Moka	United Kingdom

64	surrey	University Of Surrey	Canterbury	United Kingdom
65	wlv	University Of Wolverhampton	Selangor	United Kingdom
66	wsu	Western Sydeny University	Guildford	Australia
67	reading	University Of Reading	Wolverhampton	United Kingdom
68	liverpool	University Of Liverpool	Berkshire	United Kingdom
69	york	University Of York	Modify Record	United Kingdom
70	uwindsor	University Of Windsor	Liverpool	Canada
71	swansea	Swansea University	Heslington, York	United Kingdom
72	usd	University Of South Dakota	Windsor	United State of America
73	lincoln	University Of Lincoln	Swansea	United Kingdom
74	hull	University Of Hull	Vermillion	United Kingdom
75	manchester	The University Of Manchester	Manchester	United Kingdom
76	duke	Duke University	Hull	United State of America
77	bamu	Dr. Babasaheb Ambedkar Marathwada University	Manchester	India
78	jacobsschool	Jacob School Of Engineering	Durham	United State of America
79	ubc	The University Of British Columbia	Auranabad	Canada
80	sjsu	San Jose State University	California	United State of America
81	swinburn	Swinburn University Of Technology	Vancouver	Australia
82	canberra	University Of Canberra	Washington	Australia
83	unsw	Unsw Sydney	Hawthorn	Australia
84	itee	The University Of Queensland	Bruce	Australia
85	usq	University Of Southern Queensland	Sydeny	Australia
86	utas	University Of Tasmania Australia	St Lucia	Australia
87	columbia	Columbia University	Darling Heights	United State of America
88	uiowa	University Of Iowa	Hobart	United State of America
89	titech	Tokyo Institute Of Technology	New York	Japan
90	utdallas	The University Of Texas At Dallas	Texas	United State of America
91	auckland	The University Of Auckland New Zeland	Tokyo	New Zeland
92	uchicago	The University Of Chicago	Chicago	United State of America
93	cambridge	University Of Cambridge	Cambridge	United Kingdom
94	stanford	Stanford University	Chicago	United State of

				America
95	uga	University Of Georgia	Athens	United State of America
96	eeecs.wsu	Washington State University	Pullman	United State of America
97	stir	University Of Stirling	Stirling	United Kingdom
98	dawsoncollege	Dawson College	Montreal	Canada
99	cl	University Of Cambridge	Cambridge	United Kingdom
100	nottingham_china	University Of Nottingham China	Ningbo	China
101	nottingham_UK	University Of Nottingham Uk	Nottingham	United Kingdom
102	nottingham_Malaysia	University Of Nottingham Uk	Semenyih	Malaysia
103	princeton	Princeton University	Princeton	United State of America
104	dal	Dalhousie University	Nova Scotia	Canada
105	unimelb	The University Of Melbourne	Parkville,Victoria	Australia
106	dcs.shef	The University Of Sheffield	Sheffield	England
107	ucc	University College Cork Ireland	Cork	Ireland
108	oregonestate	Oregonestate University	Corvallis	United State of America
109	bu	Boston University	Boston	United State of America
110	royalholloway	Royal Holloway University Of London	Egham	United Kingdom
111	purdue	Purdue Science University	West Lafayette	United State of America
112	kingston	Kunston University London	Kingston Upon Thames	United Kingdom
113	toranto	Uiversity Of Toronto	Toronto	Canada
114	ucsb	University Of California, Santa Barbara	Santa Barbara	United State of America
115	udel	University Of Delaware	Newark	United State of America
116	piit	University Of Pittsburgh	Pittsburgh	United State of America
117	nyuad	Nyu Abu Dahbi	Abu Dhabi	United Arab Emirates
118	leeds	University Of Leads	Leeds	United Kingdom
119	helsinki	University Of Helsinki	Helsinki	Finland

120	cmu	Carnegie Mellon University	Pittsburgh	United States
121	hw	Heriot Watt University	Edinburgh	United Kingdom
122	andrews	University Of St Andrews	St Andrews	United Kingdom
123	southampton	University Of Southampton	Southampton	United Kingdom
124	brad	University of Bradford	Bradford	United Kingdom
125	biu	Benson Idahosa University	Benin	Nigeria
126	adelaide	The university of Adelaide	Adelaide	Australia
127	kiu	Krakoram International UniversityUniversity	Gilgit	Pakistan
128	alkhair	Alkhair Uniersity	Bhimber	Pakistan
129	marmara	Marmara University	Istanbul	Turkey
130	cukurov	Cukurova University	Adana	Turkey
131	nwu	North Western University, Khulna	Khulna	Bangladesh
132	uttarauniversity	Uttara University	Dhaka	Bangladesh
133	vub	Victoria University of Bangladesh	Dhaka	Bangladesh
134	brandies	Brandeis University	Waltham, Massachusetts	United State of America

Appendix B

Sr No	Uni_Name	Actual Attributes	Schema Extracted	HTW			HTW+		
				Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn	4	Yes	4			4		
2	AIOU	3	yes	3			3		
3	AUP	3	yes	3			3		
4	Bahria_ISB	3	Yes	3			3		
5	Bahria_Kri	3	yes	3			3		
6	Bahria_Lro	3	Yes	3			3		
7	csit	6	yes	6			6		
8	Dadabhoy	4	Yes	4			4		
9	fuuast	4	Yes	4			4		
10	hiast	4	yes	4			4		
11	icp	6	yes	6			6		
12	iiu	3	Yes	3			3		
13	inu	2	Yes	2			2		
14	MUL	3	Yes	3			3		
15	MUST	5	yes	5			5		
16	northern	8	Yes	8			8		
17	NUML	4	Yes	3	1		4		
18	Riphah	3	yes	3			3		
19	UOS	3	Yes	3			3		
20	uoswabi	5	Yes	5			5		
21	ustb	2	Yes	2			2		
22	uoli	5	Yes	5			5		
23	wuajk	6	Yes	6			6		
24	bracu	3	Yes	3			3		
25	ebaub	3	yes	3			3		
26	ewubd	4	yes	4			4		
27	hstu	4		0		4	0		4
28	iiuc	6	yes	6			6		
29	iukb	3	Yes	3			3		
30	jnu	4	Yes	4			4		
31	just	2	Yes	2			2		

32	kuet	5		0		5	0		5
33	manarat	4	yes	4			4		
34	nub	3	yes	3			3		
35	portcity	2	yes	2			2		
36	pub	5	yes	4	1		5		
37	rmstu	3	yes	3			3		
38	ruet	4	yes	4			4		
39	uoda	2	yes	2			2		
40	vu	3	yes	3			3		
41	cusb	3	yes	3			3		
42	mgcub	4	yes	4			4		
43	dimapur	4	yes	4			4		
44	tripurauniv	6	yes	6			6		
45	visvabharti	4	yes	4			4		
46	hnbgu	5		0		5	0		5
47	kakatiya	6	yes	6			6		
48	uma	4	Yes	3	1		4		
49	fim	2	yes	2			2		
50	le	4	yes	4			4		
51	dmu	3	yes	3			3		
52	ed	3	yes	3			3		
53	aber	7	yes	7			7		
54	kent	3	yes	3			3		
55	nottingham_UK	4	yes	4			4		
56	nottingham_china	4	yes	4			4		
57	nottingham_Malasiya	4	Yes	4			4		
58	surrey	5	yes	5			5		
59	wlv	5	Yes	5			5		
60	wsu	2	yes	2			2		
61	york	2		0		2	0		2
62	uwindor	5	Yes	5			5		
63	usd	3	yes	3			3		
64	lincoln	5	yes	2	3		5		
65	hull	4	yes	4			4		
66	manchester	5	yes	5			5		
67	duke	3		0		5	0		5

68	bamu	4	yes	4			4		
69	jacobsschool	2	yes	2			2		
70	swinburn	4	yes	4			4		
71	canberra	4	yes	4			4		
72	unsw	4	yes	4			4		
73	itee	4	yes	4			4		
74	usq	2	yes	2			2		
75	utas	2	yes	2			2		
76	columbia	6	Yes	6			6		
77	uiowa	5	yes	5			5		
78	titech	5	yes	5			5		
79	auckland	5	yes	5			5		
80	uchicago	4	yes	3	1		4		
81	cambridge	4	yes	4			4		
82	stanford	4		0		4			4
83	eeecs.wsu	4	yes	4			4		
84	stir	5	yes	5			5		
85	dawsoncollege	4	yes	4			4		
86	cl	4	yes	4			4		
87	princeton	6	yes	6			6		
88	dal	4	yes	4			4		
89	unimelb	6	yes	6			6		
90	dcs.shef	5	yes	5			5		
91	ucc	5	yes	5			5		
92	oregonestate	4	yes	3	1		4		
93	bu	6	yes	6			6		
94	royalholloway	6	yes	4	2		6		
95	purdue	6	yes	4	2		6		
96	kingston	2	yes	2			2		
97	toronto	4	yes	4			4		
98	ucsb	6	yes	6			6		
99	udel	5	yes	4	1		5		
100	piit	4	yes	4			4		
101	nyuad	2		0		2			2
102	leeds	6	yes	6			6		
103	helsinki	4	yes	4			4		

104	cmu	6	yes	6			6		
105	hw	2	yes	2			2		
106	andrews	7	yes	7			7		
107	southampton	2	yes	2			2		
108	brad	4	yes	4			4		
109	biu	4	yes	4			4		
110	adelaide	5				5			5
111	alkhair	2				2			2
112	marmara	5	Yes	3	2		5		
113	brandies	3	Yes	2	1		3		
114	cukurov	3	Yes	2	1		2	1	
115	dur	5	Yes	5			4	1	
116	uttarauniversity	5	Yes	5			5		
117	Awkumbuner	2		1	0	1	1	0	1
118	Awkumchitral	2		1	0	1	1	0	1
119	Awkumshankar	2		1	0	1	1	0	1
120	Awkumtimer	2		1	0	1	1	0	1
	Total	479		426	17	38	441	2	38

Appendix C

Sr No	Uni Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn			✓	✓		
2	AIOU			✓	✓		
3	AUP			✓	✓		
4	Bahria_ISB			✓	✓		
5	Bahria_Kri			✓	✓		
6	Bahria_Lro			✓	✓		
7	csit			✓	✓		
8	Dadabhoy			✓	✓		
9	fuuast			✓	✓		
10	hiast			✓	✓		
11	icp			✓	✓		
12	iiu			✓	✓		
13	inu			✓	✓		
14	MUL			✓	✓		
15	MUST			✓	✓		
16	northern			✓	✓		
17	NUML*			✓	✓		
18	Riphah			✓	✓		
19	UOS			✓	✓		
20	uoswabi			✓	✓		
21	ustb			✓	✓		
22	uoli			✓	✓		
23	wuajk			✓	✓		
24	bracu			✓	✓		
25	ebaub			✓	✓		
26	ewubd			✓	✓		
27	hstu			✓		✓	
28	iiuc			✓	✓		
29	iukb			✓	✓		
30	jnu			✓	✓		
31	just			✓	✓		
32	kuet			✓		✓	

33	manarat			✓	✓		
34	nub			✓	✓		
35	portcity			✓	✓		
36	pub			✓	✓		
37	rmstu			✓	✓		
38	ruet			✓	✓		
39	uoda			✓	✓		
40	vu			✓	✓		
41	cusb			✓	✓		
42	mgcub			✓	✓		
43	dimapur			✓	✓		
44	tripurauniv			✓	✓		
45	visvabharti			✓	✓		
46	hnbgu			✓		✓	
47	kakatiya			✓	✓		
48	uma			✓	✓		
49	fim			✓	✓		
50	le			✓	✓		
51	dmu			✓	✓		
52	ed			✓	✓		
53	aber			✓	✓		
54	kent			✓	✓		
55	nottingham_UK			✓	✓		
56	nottingham_china			✓	✓		
57	nottingham_Malasiya			✓	✓		
58	surrey			✓	✓		
59	wlv			✓	✓		
60	wsu			✓	✓		
61	york			✓		✓	
62	uwindsor			✓	✓		
63	usd			✓	✓		
64	lincoln			✓	✓		
65	hull			✓	✓		
66	manchester			✓	✓		
67	duke			✓		✓	
68	bamu			✓	✓		

69	jacobsschool			✓	✓		
70	swinburn			✓	✓		
71	canberra			✓	✓		
72	unsw			✓	✓		
73	itee			✓	✓		
74	usq			✓	✓		
75	utas			✓	✓		
76	columbia			✓	✓		
77	uiowa			✓	✓		
78	titech			✓	✓		
79	auckland			✓	✓		
80	uchicago			✓	✓		
81	cambridge			✓	✓		
82	stanford			✓		✓	
83	eecs.wsu			✓	✓		
84	stir			✓	✓		
85	dawsoncollege			✓	✓		
86	cl			✓	✓		
87	princeton			✓	✓		
88	dal			✓	✓		
89	unimelb			✓	✓		
90	dcs.shef			✓	✓		
91	ucc			✓	✓		
92	oregonestate			✓	✓		
93	bu			✓	✓		
94	royalholloway			✓	✓		
95	purdue			✓	✓		
96	kingston			✓	✓		
97	toronto			✓	✓		
98	ucsb			✓	✓		
99	udel			✓	✓		
100	piit			✓	✓		
101	nyuad			✓		✓	
102	leeds			✓	✓		
103	helsinki			✓	✓		
104	cmu			✓	✓		

105	hw			✓	✓		
106	andrews			✓	✓		
107	southampton			✓	✓		
108	brad			✓	✓		
109	biu			✓	✓		
110	adelaide			✓		✓	
111	alkhair			✓		✓	
112	marmara			✓	✓		
113	cukurov			✓	✓		
114	brandies			✓	✓		
115	dur			✓	✓		
116	Awkumbuner			✓	✓		
117	Awkumchitral			✓	✓		
118	Awkumshankar			✓	✓		
119	Awkumtimer			✓	✓		
120	uttarauniversity			✓	✓		
	Total	0	0	120	111	9	0

Appendix D

Sr No	Uni_Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	Abasyn			✓	✓		
2	AIOU			✓	✓		
3	AUP			✓	✓		
4	Bahria_ISB			✓	✓		
5	Bahria_Kri			✓	✓		
6	Bahria_Lro			✓	✓		
7	csit			✓	✓		
8	Dadabhoy			✓	✓		
9	fuuast			✓	✓		
10	hiast			✓	✓		
11	icp			✓	✓		
12	iiu			✓	✓		
13	inu			✓	✓		
14	MUL			✓	✓		
15	MUST			✓	✓		
16	northern			✓	✓		
17	NUML*			✓	✓		
18	Riphah			✓	✓		
19	UOS			✓	✓		
20	uoswabi			✓	✓		
21	ustb			✓	✓		
22	uoli			✓	✓		
23	wuajk			✓	✓		
24	bracu			✓	✓		
25	ebaub			✓	✓		
26	ewubd			✓	✓		
27	hstu					✓	
28	iiuc			✓	✓		
29	iukb			✓	✓		
30	jnu			✓	✓		
31	just			✓	✓		
32	kuet			✓		✓	

33	manarat			✓	✓		
34	nub			✓	✓		
35	portcity			✓	✓		
36	pub			✓	✓		
37	rmstu			✓	✓		
38	ruet			✓	✓		
39	uoda			✓	✓		
40	vu			✓	✓		
41	cusb			✓	✓		
42	mgcub			✓	✓		
43	dimapur			✓	✓		
44	tripurauniv			✓	✓		
45	visvabharti			✓	✓		
46	hnbgu			✓		✓	
47	kakatiya			✓	✓		
48	uma			✓	✓		
49	fim			✓	✓		
50	le			✓	✓		
51	dmu			✓	✓		
52	ed			✓	✓		
53	aber			✓	✓		
54	kent			✓	✓		
55	nottingham_UK			✓		✓	
56	nottingham_china			✓		✓	
57	nottingham_Malasiya			✓		✓	
58	surrey			✓	✓		
59	wlv			✓	✓		
60	wsu			✓	✓		
61	york			✓		✓	
62	uwindsor			✓	✓		
63	usd			✓	✓		
64	lincoln			✓	✓		
65	hull			✓	✓		
66	manchester			✓	✓		
67	duke			✓		✓	
68	bamu			✓	✓		

69	jacobsschool			✓	✓		
70	swinburn			✓	✓		
71	canberra			✓	✓		
72	unsw			✓	✓		
73	itee			✓	✓		
74	usq			✓	✓		
75	utas			✓	✓		
76	columbia			✓	✓		
77	uiowa			✓	✓		
78	titech			✓	✓		
79	auckland			✓	✓		
80	uchicago			✓	✓		
81	cambridge			✓	✓		
82	stanford			✓		✓	
83	eeecs.wsu			✓	✓		
84	stir			✓	✓		
85	dawsoncollege			✓	✓		
86	cl			✓	✓		
87	princeton			✓	✓		
88	dal			✓	✓		
89	unimelb			✓	✓		
90	dcs.shef			✓	✓		
91	ucc			✓	✓		
92	oregonstate			✓	✓		
93	bu			✓	✓		
94	royalholloway			✓	✓		
95	purdue			✓	✓		
96	kingston			✓	✓		
97	toranto			✓	✓		
98	ucsb			✓	✓		
99	udel			✓	✓		
100	piit			✓	✓		
101	nyuad			✓		✓	
102	leeds			✓	✓		
103	helsinki			✓	✓		
104	cmu			✓	✓		

105	hw			✓	✓		
106	andrews			✓	✓		
107	southampton			✓	✓		
108	brad			✓	✓		
109	biu			✓	✓		
110	adelaide			✓		✓	
111	alkhair			✓		✓	
112	marmara			✓	✓		
113	cukurov			✓	✓		
114	brandies			✓	✓		
115	dur			✓	✓		
116	uttarauniversity			✓	✓		
117	Awkumbuner			✓		✓	
118	Awkumchitral			✓		✓	
119	Awkumshankar			✓		✓	
120	Awkumtimer			✓		✓	
	Total	0	0	120	104	16	0

Appendix E

Sr No	Uni_Name	Actual Attributes	HTW			HTW+		
			Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	KUST	3	0	3		3		
2	Sbbwu	2	0	2		1	1	
3	Emps	2	0	2		1	1	
4	Liverpool	4	0	4		4		
5	swansea	2	0	2		2		
6	ubc	5	0	5		5		
7	sjsu	5	0	5		5		
8	utdallas	3	0	3		3		
9	uga	3	0	3		3		
10	kiu	3	0	3		3		
11	reading	4	0	4		4		
12	uom	3	0	3		3		
13	nwu	3	0	3		2	1	
14	vub	3		3		2	1	
	Total	45	0	45	0	41	4	0

Appendix F

Sr No	Uni Name	HTW			HTW+		
		Corr. Extr.	Wr. Extr.	Not Extr.	Corr. Extr.	Wr. Extr.	Not Extr.
1	KUST			✓	✓		
2	sbbwu			✓	✓		
3	emps			✓	✓		
4	liverpool			✓	✓		
5	swansea			✓	✓		
6	ubc			✓	✓		
7	sjsu			✓	✓		
8	utdallas			✓	✓		
9	uga			✓	✓		
10	kiu			✓	✓		
11	reading			✓	✓		
12	uom			✓	✓		
13	nwu			✓		✓	
14	vub			✓	✓		
Total	14	0	0	14	13	1	0