CHAPTER 1

INTRODUCTION

This chapter is introductory in nature. It presents motivation behind the study undertaken in the present thesis and describes overview of Web data extraction. It provides the introduction about Web based environment, Web data extraction and Web mining. Chapter closes with brief summary of the work presented in the subsequent sections of the thesis.

This chapter organizes into different Sections as follows: Section 1.1 consist of motivation behind the study, Section 1.2 presents the introduction to Web based environment, Section 1.3 includes the research areas in Web, Section 1.4 describes the Web data extraction, Section 1.5 elaborates the concept of Web mining and its types, Section 1.6 presents the issues in Web data extraction, Section 1.7 presents the applications of Web data extraction, Section 1.8 describes the efficiency parameters of Web data extraction, Section 1.9 provides the work carried out in this thesis, chapter wise organization is given in Section 1.10 and finally summary of this chapter is given in Section 1.11.

1.1 MOTIVATION

Human beings require a lot of things in daily life to fulfill their routine needs and for the easiness of life. The requirements changes from one person to another. Information is one of such requirement among them. Hence, information processing is essential to provide the desirable information according to the choice of user.

Different information sources are available in today’s life like books, newspapers, Internet etc. But in the modern world Internet has become a single platform to provide all these information. To access the Internet lot of services are available like e-mail, chatting, WWW (World Wide Web) etc.

As per the tremendous usage of Internet from time to time, the information or data available over the Websites have also grown to enormous level. More and more
information is provided over the Website as per the market demand. Therefore, the need of Web data extraction system over the Website comes into demand. Web data extraction systems are used for extracting the required information from a large amount of information available over the Website. These are basically used for saving time and provide the information according to user choice.

### 1.2 WEB BASED ENVIRONMENT

It means the requirements like tools, languages etc. in which the Web works. Constructs required for the Web based environment are described below:

#### 1.2.1 Internet

Internet refers to interconnection of a number of networks for sharing of information from one user to another. Here network means different kinds of networks like WAN’s (Wide Area Networks), wired and wireless networks etc (1). One such design of Internet is shown in Figure 1.1.

![Figure 1.1: Structure of Internet](image)

**Services available over the Internet:** A lot of services are available over the Internet to access the capability of it. Some of the services are:
i) **Web**: It includes collection of a number of Websites for providing different types of information to the users.

ii) **E-mail**: It is used to transfer the messages on personal basis. Today all types of messages like multimedia and text messages can be transferred at significant speed. E-mail system uses SMTP and POP for the transfer of messages.

### 1.2.2 Web
Web is a service available over the Internet for transfer of information. It is basically a collection of Websites. Web works on the basis of n-tier architecture (2-3). It is shown in Figure 1.2.

![n-tier Architecture of Web](Image)

**Figure 1.2**: n-tier Architecture of Web

### 1.2.3 Webpage
A Webpage is an electronic document which is provided to the user through the use of Internet and browser. A Webpage is an integrated display of text, images, sound and video types of information (4).
Structure of Webpage: Every Webpage includes two sections named header section and body section (5).

i) **Header Section:** It comprises of the title (displayed on top of Webpage), meta-data (keywords, author information, Webpage summary and Webpage index), style-sheet (CSS) and script (java-script).

ii) **Body Section:** It contains the actual content of the Webpage. Content here means the information to be included in the Webpage.

Figure 1.3 shows the basic structure of a Webpage i.e. how a Webpage is developed

<table>
<thead>
<tr>
<th>Header Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Consist of the information about the Webpage)</td>
</tr>
<tr>
<td>Not displayed on the Webpage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Body Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Consist of the information of the Webpage)</td>
</tr>
<tr>
<td>Displayed on the Webpage</td>
</tr>
</tbody>
</table>

**Figure 1.3:** Basic Structure of a Webpage

The header section of Webpage comprises of the following parts:

- **Title:** It is the title of the Webpage which is displayed in the title bar. It explains about which the Webpage is developed.

- **Meta-data:** It consists of Keywords, author information, summary etc. of the Webpage. Keywords are used to enhance the popularity of the Webpage. It helps the search engine in performing search and provides the results accordingly. Keywords help in search engine optimization (SEO) of the Webpage. Author information consists of the name of the author or programmer of the webpage.

- **Style-sheets:** Style-sheets provide different styles to the content used in the Webpage. Commonly used style-sheet is CSS (Cascading Style Sheet).
• **Scripts**: Scripts contains the java-script programs generally which are used to give dynamic effect in the Webpage on the client side.

The body section of the Webpage generally comprises of the following parts:

• **Text**: It is textual content which provides the main information of the Webpage.

• **Related Links**: It consists of links to the information which are related to a particular Webpage. These help to distribute the content of Website in different parts.

• **Sections**: It includes the main information about a specific topic which is to be displayed in the particular Webpage.

• **Sub-sections**: It consists of the sub-information about a specific topic which is to be included in a particular Webpage.

• **Forms**: These are used to get input from the user and further processing. These are also used to collect the feedback from the visitor of the Webpage.

• **Footer**: It is the address section of the Webpage.

One of the Webpage look is as shown in Figure 1.4

![Table](https://via.placeholder.com/150)

<table>
<thead>
<tr>
<th>LOGO</th>
<th>Header of Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Link1  Link2  Link3 ..........  Link n</td>
</tr>
<tr>
<td></td>
<td>Content of the Webpage</td>
</tr>
<tr>
<td></td>
<td>Footer</td>
</tr>
</tbody>
</table>

**Figure 1.4**: An Example Webpage
1.2.4 Website

A Website is collection of interlinked WebPages. It used to interconnect homogenous or heterogeneous types of information (6). Interlinked structure of Website is shown in Figure 1.5.

![Interlinked Structure of a Website](image)

**Figure 1.5**: Interlinked Structure of a Website

**Types of Websites**

Websites are categorized on the basis of content available in it. Two types of websites are available (7):

i) **Static Websites**: The content available in these types of Websites is unchanging. The user can see the content, read the content which is static in nature i.e. content can’t be decided at the run-time. These types of Websites are generally developed through HTML and JavaScript i.e. client side languages.

ii) **Dynamic Websites**: The content available in such type of Websites is not fixed. The content is decided at the run-time. These Websites normally include form type data. Hence, these types of Websites are generally developed through server-side languages like PHP, ASP.net, and CGI etc.
1.2.5 Conceptual Design of a Website

Design and development of a Website depends upon the size and content to be provided in it. Generally used design of a Website is conceptual design. It is shown in Figure 1.6 (8).

Figure 1.6: Conceptual Design of a Website
The conceptual design of a Website is made up of following parts:

- **Home Page:** It is main page of the Website and is always named as index or default.

- **Organization System:** A Website may be large scale or small scale Website according to the amount of information to be provided on it. The information to be available on the small scale Website is organized by partitioning the information in the form of different WebPages and interlinking these WebPages through hyperlinks. The information on large scale Website is organized in form of subsites. A subsite is a complete Website in itself.

- **Searching/Navigation System:** Both of these are used for finding the information available on the Website. Navigation is done by the user and searching is performed through the searching software. Thus searching systems are placed on large scale Websites only and navigation systems are placed on both small scale as well as large scale Websites due to cost factor of searching system.

- **Feedback System:** This is placed on the Website for having the views of the visitors (about graphical interface, content or navigation system of website etc.). It is generally a form in which the visitor of Website may input some text and submit it to website owner.

**1.2.6 HTTP (Hyper Text Transfer Protocol)**

HTTP is used as a communication protocol between client and server. The client sends the request to the server and the server gives back the response. Each time a request is made a fresh connection is generated and when web server gives back the response the connection breaks out. This process is known as request-response cycle. In HTTP communication, client sends the request in the form of client side commands (pseudo codes) and server gives back the response in the form of numeric codes (each having a particular meaning and series) (9). The response is either the content of the webpage or the error message. The working of HTTP is shown in Figure 1.7.
1.2.7 Types of Data in WebPages

There are three categories of data used in WebPages depending upon the structure of the data i.e. unstructured, semi-structured and structured form (10).

i) **Unstructured:** A Webpage consists of different kinds of data like text, images, videos etc. Hence, the mixed form of data which doesn't follow a pre-defined model or pattern is known as unstructured data. Examples of such type of documents are e-mail messages, WebPages with audio or image type data, pdf files etc. One such document is shown in Figure 1.8.
ii) **Semi-structured:** When unstructured form of data is provided some structure so that it can be further processed, then it is treated as semi-structured form of data. Examples of semi-structured type are tables in HTML WebPages, tree like or hierarchical structure, Xml documents etc. One such document is shown in Figure 1.9.
iii) **Structured:** In this form, the data is placed in proper database tables and can be processed in a proper manner. Data mining concepts are basically applied over the structured form of data. The data can be extracted using SQL queries from this type of data. An example of structured type data is shown in Figure 1.10.

![Structured Form of data](image)

**Figure 1.10:** Structured Form of data (13)

1.2.8 **Tools and Languages used for Web**

Development of content to be available in the WebPages or Websites requires following tools and languages:

i) **Browser:** It is used as client in client-server communication. It is used to generate the request which is to be sent to server. It understands client-side scripting language (HTML, JavaScript etc.) directly. Hence every response generated by the server is always in the form of client-side language.

ii) **Web Server:** It is software or combination of software and hardware to accept the requests from the client and generate responses according to those requests. Web servers are used according to server-side scripting language like IIS for ASP, Tomcat for JSP and Apache for PHP etc.

iii) **HTML:** It is client side scripting language and used for development of WebPages. Statements used are in the form of tags. It is interpreted language and directly understandable by the browser. It uses different tags which are fixed in nature and performs a specific task like img (for inserting image on webpage), table (for
formatting of data), form (for getting input from user) and formatting tags (for formatting the text) etc. for the development and display of content in a Webpage.

iv) **XML**: The developer gets the flexibility to generate own tags (user-defined tags) by using XML. It is used to describe the data instead of display of data. It provides database like environment on which different operations can be applied.

v) **Server-side Scripting Languages**: These are used for including dynamic content in WebPages. These require a Web server like Apache for PHP, IIS for ASP etc. for processing the commands available in these languages. These are used in different activities where server side processing is required like form processing, calculation on server etc. Examples of server-side languages are CGI Perl, .net, JSP, PHP and ASP etc.

### 1.2.9 Searching Systems

Searching systems helps in providing required information to the user as per their needs. Search engine is basically used as information retrieval system. Generally searching systems uses keyword based search for performing this work. Searching systems consists of interconnected computer servers in a datacenter (14). A unique system having interconnected datacenters performs the working of a complete search engine. The basic idea is that the required information by different users is located on one or more servers hence interconnection of these servers becomes a necessary requirement.

Another requirement is to rank the information available according to quality or content of information. A searching system may be used for providing information available on the Web or on a large scale Website. Mostly used searching systems are Google, AltaVista, Bing and MSN etc. The working of the searching system is shown in Figure 1.11.

**Working of Searching System:**

Search engine basically consists of two components for its internal working. The components are spider or crawler and database. The internal structure of spider is explained below
I. Spider or Crawler: It is the backbone of the search engine. It is software developed for internal working of search engine. The basic functions of spider include the following:

a) **Graphical Interface**: It is the interface provided to user for input of the query.

b) **Searching**: To search the content from the database according to user input.

c) **Ranking**: It is main function of the crawler which provides rank according to the quality of content. Mainly used algorithms for this purpose are HITS and VIPS.

d) **Display the Results**: Once the search is complete, the results found must be provided to the user in a proper manner. Hence it is also the function of crawler.

![Diagram of Search Engine](image)

**Figure 1.11**: Working of a Search Engine

The internal working of crawler or spider depends upon following four modules

i) **Document Processor**: The document processor processes the document or sites the user search against. The major steps performed by the document processor are

- To transform the input stream to already specified format.
- Tokenize the document stream into required format.
• Separate the subdocument pieces.
• Find main elements of the document.
• Remove extra or stop words.
• Extracts index entries.
• Computes weights.
• Update the main file according to the queries match by the search engine.

ii) Query Processor: Query processor matches query input by user with the files in which the related content lies. These are developed according to two factors named time and quality of results. It gives importance to the time instead of quality because of number of documents retrieved.

iii) Searching and Matching Function: The main function of this module is to search and match the results according to the query.

iv) Ranking Module: It arranges the searched documents according to the quality of content available in these documents.

II. Database: It is second component of searching system. The fields of database basically consist of keywords and URL corresponding to the keyword.

1.3 RESEARCH AREAS IN WEB
There are number of research areas in web. Major areas in which scope of research evolves are described below

• Web Data Extraction and Web Data Mining: It is used to extract meaningful data from a large amount of data according to user choice. It helps in search engine optimization (SEO). The user doesn’t have to search the data required by itself hence, it saves the time of users. It involves issues like user choice data, noise removal, efficient program design for Web data extraction etc.

• Web Caching: Web caching basically deals with locality of reference and traffic control issues. Locality of reference here means to store the content or data locally and provide it whenever demanded. Hence, it is used to decrease the time required to provide the data. The second issue it solves is traffic control which means to store the WebPages in between the client and the server for sometimes if traffic is there
and transfer whenever found free. It uses the systems like proxy server, hence issues like Page Replacement, priority etc. are there.

- **Load Balancing:** It deals with how a particular Web server manages the requests from a large number of clients. As the number of requests from different clients increases on a particular Web server, then all requests must be granted and managed i.e. the load on the server must be balanced. It raises the issues like how the Web server manages the number of requests, how much to be fulfilled, how much to be dropped and to which the priority must be given.

- **Searching and Ranking:** It means searching data available on the Web according to a particular keyword. The main issue is to rank the information searched according to the quality of content available in it. Algorithms like HITS and VIPS are used for this purpose.

- **Search Engine Optimization:** It applies some strategies or techniques to the content or WebPages of a particular Website, so that search engine provide higher rank to it than the other information available on the Web for same topic or fact. Issues like to manage the keywords and content in the Webpage, to know the strategies used by a number of search engine and manage the content according to all of them are involved in it.

- **Web Security:** It requires that whenever data is transfer from client to server and vice-versa the data must be transferred securely. Different issues include authenticity, correctness (integrity) and various other security issues.

### 1.4 WEB DATA EXTRACTION

The expectation of users in today’s world regarding information availability at a specific platform increases day by day. Hence data available on a particular Website is also increasing in very large amount according to demand of users and types of users. To provide data according to user needs in less amount of time from large amount of data, Web Data Extraction systems comes into scenarios.

Web Data Extraction Systems are a broad class of software designed for extracting data from Web sources (14-15). A Web Data Extraction System interacts with a Web data
source through the use of its software and extracts data stored in it. The extracted data is
stored in a particular format like database, excel file or Xml document etc. For example
if the source chosen for extraction process is Webpage, the result may consist of tags
used for development of the Webpage or the textual and multimedia content available in
the Webpage etc. After the data is extracted it must be processed further to translate into
structured form for easiness of users (16-17).

1.4.1 Working of Web Data Extraction Systems
Web data extraction system normally works in three phases (18)

i) Automation and scheduling
ii) Data transformation
iii) Use of the extracted data

i) Automation and Scheduling: Automating means to access the WebPages or scripts
as well as their localization i.e. local storage of their elements or content. It requires
developing macros or programs which execute multiple instances of the same task,
like filling forms, selecting menus and buttons, automatically updating of Webpage
etc. (18).

Scheduling means if a user needs to extract data from a Web source (e.g. share
market, news) which updates very frequently, hence scheduling tools are used to
setup a scheduler, which launch macros or programs and execute scripts
automatically and periodically.

ii) Data Transformation: In this phase information extracted from various sources are
transformed to a structured form so that it can be processed or stored for further use
ey easily. During this phase, functions like data cleaning and resolution are performed
(18).

iii) Use of Extracted Data: After the extraction process is over and required data is
transformed into the specified format, the extracted data becomes ready for use.
This phase delivers or stores the structured form of data to a computerized system
like a database, file or warehouse, so that it can be further processed by applying
analytical or statistical methods and provides more fruitful results (18).

Working procedure of Web data extraction system is shown in Figure 1.12
Figure 1.12: Working of Web Data Extraction System
1.5 WEB MINING

Web mining is a category of Data mining which deals with extraction of data from WebPages or Websites. As the WebPages are unstructured in nature hence the techniques of data mining and database systems cannot be directly applied.

Web mining is conceptually categorized into three types (19):

i) Web Content mining

ii) Web Structure mining

iii) Web Usage mining

Web mining types are shown in Figure 1.13:

![Web Mining and its Types](image)

**Figure 1.13: Web Mining and its Types**

i) **Web Content Mining**

It discovers useful and key information from the deep of the Web source as per the user requirement. Web sources generally consist of unstructured or semi-structured type of information like text, image, audio, and video hence, data mining methods can’t be applied directly to extract the data from these sources. Pattern recognitions approaches like template matching or regular expression based approaches are normally used to extract information from such type of sources. Content mining is further categorized as text mining and multimedia mining. A lot of techniques and tools like statistical, neural
network approaches, rapid miner, Web data extractor etc. maybe use for Web content mining (20).

**ii) Web Structure Mining**

It discovers and extracts the information from the link structure of the Web sources. Therefore, it works on the basis of graph theory and extract patterns from the interlink structure of the Web sources. Web structure mining is mainly used for ranking of a Webpage on the Web (21). Algorithms like Page Rank (22), Weighted Page Rank (23) and HITS (Hyper-link Induced Topic Search) are available for page ranking over the Web.

**iii) Web Usage Mining**

Web usage mining is the process to record the activities of the users while they are browsing and navigating through the Web sources. Weblog files are developed by Web servers which consist of entries about the user interaction with a particular Web source (24). The basic aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals, Webpage ranking or to improve the Web structure and Web server performance. Examples of applications include improving designs of Websites, analyzing system performance as well as network communications, understanding user reaction and motivation, and building adaptive Websites. The process of Web usage mining consists of three phases pre-processing, pattern discovery and pattern analysis phase (25).

### 1.6 ISSUES IN WEB DATA EXTRACTION

A lot of issues are there in Web data extraction for extracting the required data from Web sources (18). Main issues are listed below:

**i) Rate of Growth of Information:** As the information available on the Web sources is growing at a tremendous rate. Hence, to keep track of all the information and extract data from such type of sources is becoming very difficult.

**ii) Nature of Data:** The data available in the WebPages and Websites are generally unstructured or very rarely semi-structured in nature. Therefore the general
approaches like SQL queries of Database or methods of Data-mining like association rule or FP growth algorithm etc. cannot be applied directly.

iii) **Heterogeneity**: Data available on the Web is heterogeneous i.e. same information may be presented in different forms or meaning in different WebPages. Secondly the information from different fields must be managed at a single place and must be linked.

iv) **Automation**: Data extraction process requires a high degree of human expertise for the process of automation. It requires high level of accuracy and fast retrieval of information by Web data extraction system. So the main challenge is to develop an automated system for extracting data from the Web sources which provides high performance.

v) **Processing Capability**: Web Data Extraction systems must have the capability to handle large amount of data in specified period of time. Mainly it is required in business (like share market) and artificial Intelligence type of applications because of the constraints like fast processing and accuracy as per the market requirements.

vi) **Privacy**: Applications where privacy or security is required like in the field of social Web, banking or applications dealing with project related data, the potential (even if unintentional) attempts to violate or break user privacy should be timely and adequately identified and advertised. To provide this feature in Web data extraction is a great challenge.

vii) **Unpredictability**: Web sources are continuously updating hence, structural changes are unpredictable. Due to this unpredictability in real-world scenarios the working of such systems may malfunction.

viii) **Noise**: Noise is the content other than the required content available on a Webpage or script. Removal of noise from Web sources is required for efficient Web data extraction and search engine optimization. Hence, Noise removal is a research challenge in Web content mining and Web structure mining.

ix) **Dynamic Behavior**: The information may get change at run time due to dynamic behavior of data available in the Websites. Hence, to extract data from sources which includes dynamic content is a great challenge.
x) **Semantic Web:** To extract data from semantic Web which is in machine understandable form and in form of ontologies is a research challenge.

xi) **Integration of Web Content and Usage Mining:** Integrating both types of Web mining techniques provides efficient results in terms of parameters of web data extraction. But to use techniques and platforms of both content mining and usage mining at a single place is a very difficult task.

xii) **Webpage Ranking:** It deals with Web structure mining to remove limitations of ranking algorithms like page rank and HITS algorithms to get efficient results.

### 1.7 EFFICIENCY PARAMETERS OF WEB DATA EXTRACTION

Efficiency in case of Web data extraction means how much and which data is made available to the user in how much time. Efficiency is defined on the basis of following parameters (27):

i) **Recall**

ii) **Precision**

iii) **F-measure**

iv) **Data Extraction Time**

Precision and recall are basic parameters to find out efficiency of searching systems or data extraction systems. Whenever a search is made according to a particular topic or keyword, the extracted results are either relevant or non-relevant.

i) **Recall:** It is the ratio of relevant record found in the results to the total number of records available according to that particular topic (measure according to manual records).

ii) **Precision:** It is ratio of relevant records found in the results to the total number of record (total of relevant and non-relevant records) found in the search.

Suppose X is total number of records available according to a particular topic t, if a search is made for t and Y number of records is found as the result and Z is number of records which are relevant to t such that

\[ Z \subseteq X \]
Recall and precision are inversely related to each other i.e. when recall increases precision starts decreasing and vice-versa. In today’s world recall is not a matter of effect as it can be obtained very easily through a number of techniques and methods. But to achieve a significant amount of precision is very difficult hence all the research today keeps stress on Precision. As both recall and precision are parameters used for efficiency of searching systems hence both must be achieved up to a significant level.

iii) F-measure: It is relationship between the Recall and Precision and indicates the overall performance of the system. It is calculated generally as Harmonic Mean (HM) of Recall and Precision.

\[ F - measure(t) = 2 \times (Recall(t) \times Precision(t))/(Recall(t) + Precision(t)) \] (1.3)

iv) Data Extraction Time: It is the total time taken to extract the data according to particular topic. It is also the basic measure to check the efficiency of searching or data extraction system.

1.8 APPLICATIONS OF WEB DATA EXTRACTION

In today’s world data extraction has become a common need in many fields to provide required data and for saving time of user. Some of the applications of Web data extraction are described below (18):

i) Business: Web Data Extraction systems are applicable in broad categories of business applications like analyzing the business documents of a company or organization (e-mails, support forum, technical and legal documents). E-commerce is also an emerging application of data extraction systems.

ii) Social Web: It is usable in different areas like railways, social networking Websites (like face book) and shopping Websites for efficient information retrieval.

iii) Real Time Systems: Web data extraction systems are beneficial where very fast processing is required like in share markets etc.
iv) **Medical:** In medical field data extraction systems can provide updated information about medicines, patients in an efficient manner.

v) **Transportation and Reservation System:** Web data extraction systems are beneficial for transportation systems like in air line reservation systems, railway reservation system for both the customers and the administrative officials.

vi) **Bio-informatics:** Data extraction systems can be used in the field of bio-informatics for efficiency and security purposes.

vii) **Banking:** Data Extraction in banking requires security and efficiency. It is useful for both the customers and administrative officials of the bank.

viii) **Searching Systems:** Web data extraction systems improve the efficiency of search engines by providing limited data.

ix) **Context-aware Advertising:** These are used to present the advertisements with the content of the Webpage in a refined form. The goal is to increase the interest of visitors in advertisement along with the content of the Webpage. Web data extraction system helps to extract the relevant data and then adjust the advertisement at proper place on the Webpage.

x) **Customer Care:** Web data extraction systems are used at customer care centers to know the group of customers according to a particular item, so that interactions can be easily made. Secondly the data available in unstructured form from the customers side can be easily understand and main content can be extracted.

xi) **Database Building:** Web data extraction systems are used to develop databases of the collected content which are used frequently in online marketing. Functional areas where database building helps are financial organizations, real estate markets, job posting websites, online auction websites etc.

xii) **Software Engineering:** Today’s most of the applications available over the Internet uses inbuilt software and run online. All the Web data extraction systems use the concept of software engineering to build the program like Web wrapper.

xiii) **Business Intelligence:** Web data extraction system helps in capturing the market information which is useful for the administrators to set and implement the business strategies for future and provide scalability to the business.
xiv) **Competitive Intelligence:** Web data extraction systems help in decision making and know the consumer choice. Hence the comparative study can be easily applied to improve the results or productivity in the market.

xv) **Web Process Integration and Channel Management:** Web applications perform the task of communication among the business partners for completing various activities like sale, purchase and interaction etc. Web data extraction systems proves very useful in such type of managements like travel industry, hotel industry etc.

xvi) **Comparison Shopping:** It is the main requirement of the consumer having online shopping to compare the products on various parameters like prices, quality etc. Web data extraction systems are very useful in such type of scenarios. It requires extracting the data from the deep of the WebPages from a lot of websites and providing the comparative results. Examples of such type of systems are Lixto, Unitminer etc.

xvii) **Mashup Scenarios:** Mashup scenarios are developed through integrating the views of a number of users either technical or non-technical users. Thus, Web data extraction systems help in integrating the data generating in any form and for promotion of Websites and business.

xviii) **Opinion Mining:** It is used in business strategy for having the opinion of users about the product. These opinions are collected in the form of blogs, feedback forms, charts, tags etc. i.e. having no structure. Hence, Web data extraction systems are used to integrate and provide structure to data collected.

xix) **Citation Databases:** Citation indexes are developed according to a particular researcher by different organizations like Google scholar, Cite Seer etc. Therefore, Web data extraction systems are used to extract the data accordingly (research paper referred by others) and calculate the citation index respectively

### 1.9 WORKS CARRIED OUT

Our contribution to the work is based on the objectives chosen for this thesis. Initial phases for performing the work are:
• **Review:** Some base research papers were chosen according to each objective and on behalf of the research performed in base research papers the objectives were carried out and performed.

• **Data Selection:** As the whole work is based on Web based environment. So websites were developed as per the requirements. Data selection means the number of WebPages in a different Website on which the experiments were performed.

• **Pre-requisite:** Different platform were chosen for the development purpose e.g. client side platforms like HTML and JavaScript, server side platforms like PHP etc. Client side uses browser and server side uses Web servers like Apache.

• **Development:** As the field of the research is Web data extraction, the programs generally known as Web wrapper in Web data extraction were developed according to the objective.

• **Testing:** As the whole work is performed on real environment i.e. on websites hence testing was performed very easily either by online publishing of Website or by making our own system as both client and server.

Following works are carried out in this thesis to fulfill the objectives:

i) Web content mining approach is presented on the basis of conceptual design of the Website. The pattern uses for performing the work is chosen as default WebPages of the Website. The presented system uses content matching instead of keyword matching as in case of traditional systems. The approach is proposed basically for large scale Websites hence named MCMM-LSW (Multilevel Content Mining Model for Large Scale Websites). The results are compared with the approach in the base paper (2-level approach) selected and were found to be better in terms of efficiency parameters like recall, precession, F-measure and extraction time.

ii) A usage mining approach is proposed on the basis of clustering of Weblogs. It is named as Usage mining based on clustering (UMC). Relevancy rank report is prepared on the basis of accessibility of a Webpage by different users. Relevancy rank report is used to know the users choice and Weblogs are developed according to the relevancy rank report in clustered form. It uses partition based clustering approach. The results are compared with the approach in the base paper (complete
weblog using query based approach) selected and were found to be better in terms of efficiency parameters (mainly data extraction time).

**iii)** A noise reduction approach based on template of the Webpage is proposed. It is named as n x 1 approach. It uses n x 1 table template which provides structure to the Webpage. The Webpage is then transformed into XML format and XSLT with filter feature is applied over it to remove the noise from the Webpage. The results are compared with the approach in the base paper selected (image method) and were found to be better in terms of efficiency parameters.

**iv)** A Web wrapper is developed to show the working of Web Wrapper induction system. It is named as Relational Web Wrapper (RWW). It is used to extracts the metadata of the Webpage like title, description, keyword etc. The performance of Relational Web Wrapper is tested on different Websites developed and results are developed in terms of efficiency parameters such as Recall, Precision, F-measure and data extraction time.

**v)** The performance comparison of already existing techniques used for Web data extraction is performed. It uses n x 1 method for Wrapper induction, DEPTA and MDR for Automatic methods. The performance of all the approaches is compared by applying on different Websites in different situations. The results are compared on the basis of efficiency parameters such as Recall, Precision, F-measure and data extraction time. The comparison results found are better in case of n x 1 when using the template accordingly, in other cases DEPTA performs better than the MDR and n x 1.

**vi)** The overall model which integrates both Web usage mining and Web content mining approaches is proposed. It uses MCMM-LSW for content mining purpose, UMC for usage mining purpose, n x 1 for noise removal from WebPages. It prepares an overall model for Web data extraction system. Overall results found will be better in terms of efficiency parameters because MCMM-LSW provides efficient data extraction and UMC extracts data according to user choice and takes less web data extraction time.
The overall framework of Works carried out is shown in Figure 1.14

![Diagram](image)

**Figure 1.14:** Framework of the Works Carried Out

### 1.10 ORGANIZATION OF THE THESIS

The thesis is organized into 10 chapters. Chapter wise overview is as follows:

- **Chapter 1** is introductory in nature. It outlines the motivation behind the study undertaken in the present thesis and describes overview of Web data extraction. It describes the introduction about the Web based environment, Web data extraction and Web mining. The chapter closes with brief summary of the work presented in the subsequent section of the thesis.

- **Chapter 2** provides brief review of literature about Web data extraction and Web mining with its types by different authors. It includes techniques, approaches and
methods proposed by different researchers in the field of web wrappers, web content mining, web usage mining etc.

- **Chapter 3** includes the research objectives chosen for this thesis. It includes the brief introduction about each research objective along with the issues the particular objective solves.

- **Chapter 4** provides solution for objective1 chosen for this thesis. It consists of MCMM-LSW (Multi-level content mining model for large scale websites) as a solution for objective1. It is concluded that MCMM-LSW performs better in term of efficiency parameter than the 2-level approach and can be applied over to small scale as well as large scale websites.

- **Chapter 5** provides solution for objective2 chosen for this thesis. It consists of UMC (Usage mining using clustering approach) as a solution for objective2. It is concluded that UMC model is very efficient in terms of extraction time and also provides solution for Web cleaning and tracking.

- **Chapter 6** provided solution for objective3 chosen for this thesis. It consists of nx1 method as a solution for objective3. n x 1 removes the local noise from the WebPages in efficient manner.

- **Chapter 7** included Relational web Wrapper as an example to show the working of Web wrapper induction.

- **Chapter 8** provides solution for objective4 for this thesis. It consists of performance comparison of already existing web data extraction techniques in terms of Web data extraction parameters. It uses manual, n x 1, DEPTA and MDR for comparison work. It finds out which technique is suitable for use in which situation.

- **Chapter 9** provides the CCMUM which is a combination of nx1, MCMM-LSW and UMC models described in this thesis. It combines content mining and usage mining approaches and provides efficient Web wrapper in terms of efficiency parameters of Web data extraction.

- **Chapter 10** gives the overall conclusion of the thesis which helps to know all the chapters at a glance. It also includes some suggestions for future directions, which
helps to know how the work described in previous chapters can be enhanced to provide more efficient results.

1.11 SUMMARY
This chapter provides overview of Web Data Extraction systems. It consists of introduction to Web based environment, Web Data Extraction and Web Mining. It includes issues, applications, efficiency parameters of Web data extraction and work carried out in this thesis.