CHAPTER 2

LITERATURE REVIEW

The data available over the Websites is increasing at a rapid rate as per the demand of the market. Hence, to provide data according to the user choice is a great challenge. In addition, the extracted data must also follow the constraints of time and accuracy. Therefore, a lot of techniques and methods are proposed by different researchers based on facts like accuracy, efficiency, security etc.

In the earlier stages, the field of Web data extraction consist two classes named learning techniques and knowledge engineering (27). These categories require human expertise to develop the system for data extraction. These are generally based on the regular expression and artificial intelligence. Methods and approaches proposed by different researchers are described in subsequent sections of this chapter.

This chapter is organized into 5 sections. Section 2.1 provides literature review according to tree based methods, Section 2.2 consist of literature review about Web wrapper, literature review about machine learning approaches is given in Section 2.3, Section 2.4 includes literature survey according to web mining and its types and finally summary of this chapter is given in Section 2.5.

2.1 TREE BASED METHODS

Due to unstructured or semi structured nature of data available in the WebPages, tree based are generally used for providing structure to the WebPages. Hence, DOM (Document Object Model) tree based methods are generally used for Web data extraction. DOM tree is graphical or tree representation of the tags used in Webpage.

Tree based methods consist of following different methods for extracting the data from Web sources (28).

i) XPath Based Methods: It uses XPath language to identify the elements of the tree. XPath performs twofold works i.e. identify individual elements of the document tree and identify duplicate occurrence of the same element The limitation of XPATH based methods is that, it is strictly based on the structure of
the Webpage. A very minor change to the template or structure of the Webpage requires a lot of updation in the entire process. To overcome this limitation tree matching based methods were proposed (28).

**ii) String Edit Distance Matching Algorithms:** This method is performed by matching two trees e.g. tree1 and tree2 so that tree1 can be transformed to tree2 with minimum number of operations. The main requirement is to reduce the number of operations so that overall cost can be minimized. The problems found in this method are that every node must appear only one time and relations among nodes must not be changed (29).

**iii) Simple Tree Matching Algorithm:** This method follows the instruction that during the comparison of two trees the node replacement is not allowed. Overall cost of this algorithm is \( O(\text{nodes}(\text{tree}1) \cdot \text{nodes}(\text{tree}2)) \), where nodes (tree) represents the total number of nodes in the tree. Main limitation of this algorithm is that it is not able to match different combinations of nodes and no hierarchical structure is permissible (30-35).

**iv) Weighted Tree Matching Algorithm:** It is another type of simple tree matching algorithm. It uses re-normalization factor to adjust the similarity values between two trees. The change in structure or template of the Webpage doesn’t affect so much in this case. The weighted value for a match among the two nodes is found by dividing the largest count of siblings with the number of paired comparable nodes (31, 33, 36).

### 2.2 WEB WRAPPER

Web wrapper is a program developed for extracting the data from web sources. It uses unstructured or semi-structured form of data and transform into structured form. Its working is divided into three phases i.e. wrapper generation, wrapper execution and wrapper maintenance. The first two steps are performed either by manual efforts or through a program. Web wrapper uses following approaches to extract data from the WebPages (28):
i) **Regular Expression Based Approach:** It defines a set of rules to find patterns or strings from the unstructured data available and then a program is developed on the basis of these rules.

Kushmerik *et. al* (37) proposed a regular expression based Web wrapper named W4F. W4F works on the basis of split and match expressions. It separates the words of a string by annotating its different elements. It has the limitation that whenever a little change occurs in the template structure of the Webpage, the regular expression must be changed or rewritten accordingly.

ii) **Logic Based Approach:** in this approach tools are developed for extracting the data using Web based languages. It considers the Webpage as semi-structured documents.

Gottlob *et. al* (38-39) proposed a logic based Web wrapper named MSO (Monadic Second Order) logic. It uses wrapping language and designs visual tools which follow the condition, all the usable elements can be implemented through related visual primitives.

Baumgartner *et. al* (40-41) proposed a logic based approach for real world scenario. The language used was named Elog. A web wrapper named Lixto Visual Wrapper system was developed using the Elog language for deep Web data extraction. It provides graphical user interface in hierarchical format which gives relationship among the elements of the Webpage. The extracted data is transformed into XML structure for obtaining the structured format.

iii) **Tree Based Approach:** It is also used for wrapper generation by generating a program for Web data extraction.

Liu *et. al* (42-43) proposed an approach named partial tree alignment. It uses the information in WebPages in the form of contiguous regions known as data record regions. It works on the basis of identifying and extracting these data regions. It is implemented on the basis of tree edit distance matching. This Web wrapper works on the basis of two phases named segmentation and partial tree alignment. The first phase divides the Webpage into different segments without extracting any data. It uses the features like DOM tree, visual clues etc. for performing this work and
provides structure to the Webpage. The second phase identifies the segments developed in first phase. The data is extracted from the DOM tree developed. This method is dependent on the structure of the Webpage hence, a little change will affect very much on the results.

2.3 MACHINE LEARNING APPROACHES

These approaches are used to extract domain specific information from Web sources. These approaches are generally developed on the basis of template or statistical based methods.

Phan et.al (44) proposed a conditional model based on Statistical Machine Learning systems. Turmo et.al (45) proposed a model based on adaptive search and machine learning system for providing a solution to human knowledge and interaction.

Some of the web wrappers based on machine learning systems are described below:

i) WIEN: It was the first wrapper induction system proposed by Kushmerick (46). It was based upon inductive learning approach. WIEN had the capability of labeling training pages automatically hence, it requires very less human efforts. It had the limitation that it cannot deal with the missing values which generally occurs in case of real world scenarios.

ii) SoftMealy: Hsu et.al (47) proposed wrapper induction system to extract data from the WebPages. It works on the basis of learning extraction rules through the use of non-deterministic finite state automata and bottom-up inductive learning approach. It requires training WebPages and from these WebPages it extracts the states for extracted data and transitions states for extraction rules. It uses a method for core representation of WebPages. First step i.e. Pre-processing step tokenize the Webpage by following a combinations of inferential rules, secondly tokens were used to define separators which imagine invisible borderlines among two continuous tokens. Finally, the Finite State machine was fed by sequence of separators, instead of raw HTML strings (as in WIEN), so as to match tokens with particular contextual
rules. SoftMealy has the capability to handle a lot of exceptions, like multiple attribute values, variant attribute permutations, missing values/attributes etc.

iii) STALKER: Muslea et.al (48) proposed a wrapper induction system based on supervised learning system which is similar to Softmealy. STALKER tokenizes the elements of the Webpage by manual positioning. Hence, STALKER has the capability of handling empty values, hierarchical structures and non ordered items which was limitation in Softmealy. STALKER uses hierarchical structures called embedded catalog tree. The root node of the embedded catalog tree is used to record the sequence of all tokens. Each child node of embedded catalog tree is a subsequence of tokens generated from its parent node. Hence, parent node act as a super-sequence of tokens for its children. The super-sequence derived is used at each level of the hierarchy so as to keep track of the content available in the sub-levels of the Embedded Catalog tree. The extraction of content demanded by the user is achieved by using a set of extraction rules on the EC tree. STALKER has the capability to define wildcards, classes of generic tokens that are inclusive of more specific tokens.

iv) RoadRunner: Crescenzi et.al (49-51) proposed a template based Web wrapper system which uses learning based approach. It works on the basis of two criterions. First criterion provides the capability to work in an automatic way and ensure parameters like robustness and reliability as per the user requirement. Second criterion requires labeled WebPages, thus requiring human involvement for choosing the WebPages. Road Runner uses visual look and reasoning methods for recognizing elements of the WebPages. RoadRunner works by taking two WebPages at the same time and discovers patterns by analyzing the resemblances and inequalities among the template structure and data from this pair. RoadRunner has the capability to extract meta-data from any Website which consists of minimum two WebPages having same template. As per the dynamic behavior of the WebPages, RoadRunner uses template of the WebPages to extract relevant data from positioning areas of information. RoadRunner can also handle missing values,
optional values and webpage look differences. Hence, it can be applied over all kinds of Web sources. It is an open source system and available very easily.

v) **Visual Box Model**: Krupl et.al (52) and Gatterbauer et.al (53) proposed a model for Web data extraction based on visual cues. It is generally based on the template using table tag in HTML WebPages. It works on the basis of X-Y cut OCR algorithm. OCR algorithm generates visual grid for the elements of the Webpage according to the co-ordinates. X-Y trees are generated to store the bitmap image which is generated through the Cuts applied in recursive manner. This tree follows the criterion that all non-empty tables are assigned to the ancestor nodes with leaves. Special Operations are applied to check whether extracted content contain beneficial information or not.

### 2.4 WEB MINING

Web mining is used for efficient data extraction in Web based environment. It is used to find patterns and extract content from the Web sources as per the user requirement. Web mining consists of three types named Web Content Mining (for content extraction), Web Structure Mining (for structure or link data extraction) and Web Usage Mining (for user behavior). The software or program developed for all these types of Web mining are generally known as Web wrapper (19-24). Overall structure of Web mining and its types are shown in Figure 2.1. Some of the techniques, method and approaches used for web data extraction according to these types of Web mining are summarized below:

**Fung C.C.** and **Wigrai Thanadechteemapat** (54) proposed a model for extracting content from multiple WebPages of a Website. In this approach two levels of Websites were taken into consideration i.e. main page and link pages from the main page. It uses content matching method to extract content from the Website. It is applied over small scale Websites.

**Fung C.C.** and **Wigrai Thanadechteemapat** (55) propose an approach for extracting the content of the single Webpage from a Website. It uses Xpath language to extract data from a single page of a particular Website. It uses block detection, feature extraction from the WebPages to extract the data.
Figure 2.1: Web Mining Working (25)
**Runkler et.al** (56) proposed an approach named relational clustering. This method was applicable for content mining as well as usage mining. It uses RACE (Relational Alternating Cluster Estimation) to deduce clusters from the content which didn’t consist of direct representation by which the content could be extracted. For content mining it analyzed the actual data of WebPages, and then if the WebPages consist of textual content, then content mining was performed by relational clustering of Levenshtein distances. It can also be used for extraction of keywords and for differentiation among different Web documents. For Web usage mining it find out a number of WebPages visited by different users. From this data it deduce two relations named Graph-based and Sequence-based relations which served as a base for relational clustering for such type of content. On the basis of this data it shows the choice of users and accordingly the content can be updated.

**ChaoWang et.al** (23) proposed a Web content mining approach to extract main (key) information from a huge amount of WebPages. By key information it means menu data, hidden information embedded inside the WebPages and navigation indicator etc. It used a method named KIM which is a two-phase method to automatically extract key information of the WebPages. First phase extracts a list of candidate keys from the WebPages and the second phase uses a method based on entropy evaluation to reduce the noise from the list and stores key information in a storage media.

**Srikantaiah et.al** (57) proposed an approach for Web content mining. It extracts Web content from WebPages using two methods WDES and WDICS. The system provides duplication of search result pages and provides method to remove duplication which helps in efficient browsing. The proposed approach works in two phases, the first phase uses a similarity function and develop a mathematical model known as WDES to find the similarity between the WebPages and in the second phase it uses cosine similarity based model (WDICS) to extract data from the links of the WebPages.

**Wei Liu et.al** (58) proposed how to extract data from depth of WebPages. The authors provide an approach based on vision of the Webpage so that it becomes
language independent. For this research a large number of WebPages are identified and their vision or structure is noted. The approach then works on the basis of four steps i.e. Visual Block tree building, data record extraction, data item extraction, and visual wrapper generation. Visual Block tree for a particular Webpage is built in the first step using VIPS algorithm, in the second step it performs record extraction, the third step performs data item extraction from the data records and then finally a Web wrapper program is generated to extract the content required for improving the efficiency of content extraction.

**Lihui Chen** (59) proposed a system for Web content mining named Semantic Virtual Document (SVD). It uses structure of the WebPages along with content mining approach. It uses summarization techniques to represent content available in actual WebPages. SVD uses HAC single linkage document clustering algorithm for partitioning or categorization of content available in the WebPages. To deduce the efficiency of Web content mining using SVD, it uses a program named iSEARCH (Intelligent Search and Review of Cluster Hierarchy). SVD provides better results in form of parameters like precision, recall and execution time.

**Hang et.al** (60) proposed an approach for Web content mining named FP-Contraction based on incremental FP-Growth algorithm of data mining. It is applied over a real life database. It takes input from the WebPages in semi-structured form and converts the data retrieved into structured form (in form of database). By applying incremental growth algorithm it extract association rules. These association rules are used to provide choice of users. It applies the technique by developing a FP-tree from the data available and divides the tree into many trees according to item sets extracted. The newly developed trees are much smaller than the original one hence extraction process becomes faster and efficient.

**Sayali et.al** (61) proposed a technique using Trinity which uses two or more WebPages originated from same server-side template. The technique searches shared pattern and divide them into three parts prefixes, separator and suffixes for applying different ranking functions. The extracted results are stored into a database. The results obtained are more efficient in terms of data extraction time.
Hao Ma (62) proposed an approach for extracting content from the Web graphs by applying recommendation technique. It uses the concept that various types of recommendation are made by the users in daily routines activities, and these recommendations are represented in the form of graphs. To extract content from the graphs generated the proposed approach works in two phases. The first phase provides similarities between the nodes of the graph by applying a novel diffusion method. The second phase creates graph diffusion framework from different recommendation problems. The proposed approach can be applied for recommendation in many fields like World Wide Web, image annotations, expert finding, image recommendations, query suggestions, tag recommendations, etc.

Velasquez et.al (63) proposed an approach for Web content mining in case of semantic Web. It takes the data in the form of machine understandable form instead of user understandable form. It specifies Website key object (which are of user interest on the Webpage) extraction problem which is solved through semantic Web. It develops relations between the objects by using ontolgoies according to user interests i.e. calculated on the basis of time spent by a specific user on a specific page. A similarity measure is used for comparison of two objects. The approach is applied over a real time Website and results are produced on the basis of objects appearing in the same Website.

Rowena Chau et.al (22) proposed an approach for text mining which is a type of Web content mining. It uses the concept of Cross-Lingual Text Retrieval (CLTR) using multilingual text mining. CLTR works in two phases, the first phase provides the data relevant to a domain based on multilingual concept–term relationships from linguistically diverse textual data. Second phase extracts the content of multilingual text from the multilingual concept–term relationships which is in the form of a document containing related information or query expressing the data required. Both the document and query extracted are compared with chosen common data to show the use of CLTR based on the results provided by the approach.

Etemadi (64) proposed an approach to apply clustering to WebPages content. It proposed an algorithm based on the expressions and keywords in Web pages. It uses
a new similarity criterion obtained from Cosine and Jaccard similarity criterion to extract the content of the WebPages. Dunn index is used to measure the distance between the separated clusters which occurs due to inappropriate allocation of clusters during clustering process. The results are compared with the cosine similarity criterion based on the Dunn index change. The results shows high efficiency in case of clustering approach.

**Kyung-Joong Kim et.al** (65) proposed an approach for Web content mining based user perspective and choice. Hence it provides the content extracted from the combination of Web usage mining and Web content mining. The proposed approach uses fuzzy integral system for this purpose. It uses Structure Adaptive Self-Organizing Map (SASOM) which is a variant of SOM for pattern recognition and visualization. Experimental results obtained by the proposed method are compared with Naive Bayes Classifier method of data mining and are found better in terms of various parameters of Web data extraction.

**Nan Di et.al** (66) proposed a model for Web content mining. It extracts the data from the WebPages in the form of named entities. It extracts the Meta information of the Webpage like name, geographic location, organization and time. It also extracts the relations between these entities which are represented by weight, classification etc. The model is applied on a set of pages for a number of days and as the result it provides a demonstration system which is able to find the page location and can be used for efficient searching.

**Baohua Liao et.al** (67) proposed a method for Web content mining. It uses a concept of mathematics named Gaussian Smoothing. It uses the properties of webpage like tags, text and other elements of Webpage for extracting the content of the Webpage. The framework designed is named as Gaussian Smoothing Content Extractor (GSCE). It extracts the main content of the Webpage along with the metadata like title, date of publishing the Webpage etc. The proposed technique is applied to semi structured WebPages and provides higher accuracy of extraction in the form of recall and precision.
Faustina Johnson et.al (68) proposed a Web content mining approach based on genetic algorithm which extracts good quality WebPages and extracts data from those WebPages. For the data collection it selects the WebPages on the basis of different parameters like time, Website existed, backward link, forwards links etc. from different Websites. It then applies genetic algorithm approaches like mutation, crossover etc. to extract the content from these WebPages.

Majid Javid Moayed et.al (69) proposed an approach for classification of Web-page according to different parameters. It uses ant colony mechanism for this purpose named Ant-miner II. It proposes an approach for Web content mining for text based content only and doesn’t consider the other complicated content. It also faces the challenge of growing information over the WebPages. The results are compared with already existing ant-miner algorithm and are found better in terms of efficiency of mining the text from the WebPages.

Haitao et.al (70) proposed an approach known as image method for local noise reduction from WebPages. In this approach the Web pages are processed as images. To give structure to the Webpage it first develops the DOM tree of the Webpage and then divides the Webpage into different blocks with a particular identification (id). It then extracts the content block according to the id of the block. The results found are stored in xml document. The total time taken to extract the content and remove the noise is DOM tree creation time and time to extract the content.

Sivakumar et.al (71) proposed an approach to enhance performance of content mining by removing the noise from a particular Web page. It uses block splitting method to divide the Webpage. First Simhash is applied to these blocks for removing the identical blocks. For every block, three parameters named Keyword Redundancy, Linkword Percentage and Titleword Relevancy are evaluated to identify the importance of a particular block. A threshold value already specified is used to remove the noise blocks from the identified blocks. The extracted blocks are used for the purpose of Web content mining.

Thanda Hwte et.al (72) proposed an approach for removing noise from the WebPages i.e. to extract the information which is meaningful for the user and
remove the rest from Webpage and provide it to the user. The proposed approach provides three patterns content, noise and mixture using neural network and DOM tree. It uses feed forward feature using back propagation learning algorithm of neural network. The classification results obtained from back propagation neural network are used to remove noise patterns from Webpage. The results are shown on the basis of percentage of content, noise and mixture of content and noise extracted from various WebPages.

**Thanda Hwte** (73) proposed an approach for removing the noise from WebPages of Website. This approach uses Case-Based Reasoning (CBR) method to find noise pattern from content which consist of a mixture of data and noise. The back propagation neural network algorithm is used to identify three patterns named noise patterns, data patterns and mixture patterns from the content of a Webpage. The classification obtained from the neural network results is used for removing noise from the Webpage.

**Dutta Amit et.al** (74) proposed an approach for removing noise from WebPages. The approach is based upon the regular expression followed by Site Style Tree (SST). The proposed method works in two phases. The first phase uses the filtering method based on regular expression to remove noisy HTML tags. The second phase uses the filtered document by using entropy based measured to remove the further noise. The size of the Webpage is reduced by eliminate a number of lines of code. Then the DOM tree is developed for the reduced size document and parallely the Site Style Tree is formed by considering the pages from the same URL path of the Website. The experimental result provides better space and time complexity through elimination of noise.

**Azad et.al** (75) proposed an approach for removing the local noise from the WebPages. The approach works on the basis of keywords parsing in the Webpage and remove the selected keywords. It then uses tag filters for images and links named filters and jsoup. Filter removes the images by considering the tag `<img>` and jsoup removes the links by considering the anchor (`<a>`) tag. It then removes
the unnecessary tables in Webpage. The results found provide better efficiency than the techniques used in the literature.

**Ida Mele** (76) proposed an approach for Web usage mining known as query based approach. In this approach a weblog is generated on the basis of a particular user visit a particular Webpage. The weblog generated is used to extract links and content according to user choice. A query covering approach was used to search the WebPages from cache and weblog generated. The extraction time and efficiency parameters (recall and precision) were calculated on behalf of query based approach. No method for weblog cleaning and duplicate entries in the weblog was proposed in this approach.

**Ford Lumban Gaol** (77) proposed an approach for Web usage mining known as Sequential Pattern Mining. It uses a method based on Apriori-all algorithm of data mining for Web usage mining which provides the data frequently visited by the user and hence provide the track of user behavior. The information captured is used to structure or format the Websites.

**Muslea I et.al** (78) proposed a Web wrapper induction system named STALKER to perform Web data extraction from highly-structured WebPages i.e. in which content is placed in the form of proper tables or hierarchical order. It is applied by using single-slot extraction as well as Multi-slot extraction. STALKER consists of two types of lists which consist of the start rules and end rules respectively. STALKER develops a link tree of the Website from which the content to be extracted. It matches the content or keyword according to which the content to be extracted with the link tree developed and provides the results accordingly.

**Ioannis Paparrizos** (79) proposed an automatic Web mining tool to apply structure, content and usage mining using a single tool. The proposed approach develops a multi-thread Web crawler to automatically extract the data which fits in all faces of Web mining. It provides the results based upon the category of the WebPages like the Webpage is blank (provides 100% usage, 0 % structure and 0-100% content data) or redirected Webpage. The clustering approach is used to cluster together WebPages with same patterns or statistics.
Haibin Liu et al. (80) proposed an approach for automatic Web data extraction which uses integration of usage mining and content mining. It extracts the data according to user movement and find out the future requests based on these patterns. The textual content of WebPages is extracted using character N-grams, which are integrated with weblog files to derive user navigation profiles. This approach can be used for personalization and better organization of a particular Website. The results obtained provide a better accuracy in terms of Web data extraction parameters. The approach is applied over three types of classifications. The first classification is equal weight methods and finds out better efficiency and accuracy than the method available in literature. Second classification is KNN classification which is evaluated through GM based dissimilarity and performs better in form of classification and prediction. Third classification is based on N-gram size i.e. document frequency and results obtained are again better the techniques used in literature.

Sampath et al. (81) proposed a method for Web usage mining. It uses a method named systolic tree for generating the weblog and then generating the pattern to extract the content from this weblog. First association rules of data mining are used to extract patterns from the WebPages. The systolic tree is then used to arrange these candidate sets according to the particular values. The proposed approach also takes the benefit of FP-growth algorithm of data mining due its limited size. It provides better efficiency by dividing the tree into parts and selects a particular tree for content extraction. Automatic weight estimation scheme is used in this system. The results are compared with already existing weighted rule mining process. The results are found to be better in terms of efficiency parameters of data extraction.

Zhang et al. (82) had done the work on data preprocessing in Web usage mining. They presented a new algorithm called USIA (User and Session Identification). It finds the user and session identification details. The same user is identified with the help of IP address and User ID. If the request is from the same IP address, then the algorithm concludes that the request is from same user. The session is identified based on the time in and time out period. This research work
mainly focused on user identification for the particular session and series of WebPages viewed by the user.

Yinghui et.al (83) focused on grouping the customer transactions by using the clustering technique. The set of transactions in a group has some similarities, so it can easily identifies the customer behavior and the Website analyst can able to understand the customer expectation and make the Website customer friendly. In other point of view, it make the Website more personalized and more user friendly. It uses pattern based clustering approach to group the similar types of transactions.

Yi Dong et.al (84) dealt with two types of groups one is Web clustering groups which groups the relative pages from the Web server log files, the second is User clustering groups which groups the user who refers the same type of WebPages. Divisive Hierarchical Clustering Algorithm is used to group the weblog files and user of similar type. Then the association rule mining with support and confidence measure is applied to each group to find the relationship among them.

Hannah et.al (85) focused on the Pre-processing phase of Usage Mining and suggested an approach based on Rough set Theory. The problem in weblog Files is their size and unwanted data. This approach used two algorithms Quick Reduct and Variable Precision Rough Set Algorithm to identify the necessary data from the weblog files. The k-means clustering algorithm was applied first to segment similar patterns before applying the above two algorithms. So the algorithms are applied only to the group of similar items to identify the feature selection. So this technique gives the optimal solution for eliminating the unwanted data in the weblog files.

Suneetha et.al (86) mainly focused on the data pre processing step to remove the unnecessary data such as images, extra click events. Pattern discovery algorithms are used to eliminate the unwanted data from the Web server log files. The data was taken from NASA Website server log files and remove the unwanted data to improve the efficiency of the weblog data analyzing process. No specific data mining techniques are applied to weblog files after pre processing. That work is open for future research workers.
Debnath et al (20) proposed an approach which automatically identify the information section available over the WebPages and then extract the data from those WebPages.

Jinlin et al (24) proposed an approach which is based on clustering method i.e. to divide the information available over Website in different sections and extract data from these sections. The approach was basically applied over news based Websites.

Wang et al (87) has done the comparative study on various sequential association rule mining algorithms with the various sequence and temporal constraints to predict the next request from the user. The result is affected based on the set of constraints. So, choosing the correct constraint gives better predictions result.

Lan Yi et al (88) proposed an approach for capturing the template used for developing the WebPages. It then find the common blocks used in the WebPages. The approach works on the basis of a compacted structure tree. An information based measure is used to evaluate the significance of every node in the compacted structure tree. On behalf of the node significance values it assigns a weight to each word in a particular content block. The approach gives results on the basis of two Web mining tasks, named Webpage clustering and Webpage classification.

Chang et al (89) proposed a string matching method for extracting the content of the WebPages, it extract the data by observing the distance in the tag tree or DOM tree developed.

B. Liu et al (90) proposed an automatic approach called MDR which is based on string matching method. It builds the tag tree of the Webpage and extracts the content based upon tags used for matching. MDR generally works on the basis of table and form types of tags. MDR first develop the DOM tree of the Webpage and then find out generalized nodes from this tree. MDR extract the data from the generalized nodes on the basis of the features of different generalized nodes. Limitation of MDR is that it is HTML tag dependent and doesn’t consider visual feature of the Webpage.
B. Liu et al. (31) proposed a technique called DEPTA which uses tree alignment instead of tag strings, which exploits nested tree structures to perform more accurate data extraction. DEPTA works on the basis of two phases. In first phase it identifies the data regions using MDR described above and in second phase it extracts data from these regions using tree alignment. It provides better results in terms of efficiency parameters (Recall and Precision) but takes some more extraction time than MDR because of tree matching instead of tag matching.

B. Liu et al. (91) proposed an automatic Web data extraction method called NET which was used to extract Web data from both flat and nested types of data records. Given a page as input, NET first builds a tag tree based on visual information of the Webpage then it performs a post-order traversal of the build tree and matches all subtrees of the tree using a tree edit distance method and visual cues. It is an extension of DEPTA and MDR.

Hiremath et al. (92) proposed an automatic Web data extraction approach based on visual clue of the Webpage. It works on the basis two steps i.e. Identification and Extraction of the data regions based on visual clues information and Identification of data records and extraction of data items from a data region. This technique is not dependent upon the tags used in the content and can also extract the data from contiguous as well as non-contiguous content.

Yan Guo et al. (93) proposed an approach ECON for removing the noise from the WebPages. This approach develops Document Object Model (DOM) tree for capturing the structure of the Webpage. The approach was applied over news WebPages. The approach first find out a snippet node from the DOM tree and it backtrack this snippet node recursively until a summary node which consist of entire content of the news Webpage is not found. Noise elimination is done by ECON during the process of backtracking. ECON fulfills the condition of scalable extraction and it also provides high accuracy level in terms of efficiency parameters. ECON was applied over a lot of news pages developed in different languages and it works efficiently in all types of WebPages.
2.5 SUMMARY

This chapter provides different studies carried out by other authors. It includes literature review about Web data extraction systems and Web mining based techniques and methods.