# Web Structure Mining of Dynamic Pages

By

**Muhammad Asif Naeem**

**A thesis is submitted in partial fulfillment of the requirements for the degree**

**of**

**MS Computer Science**

**Faculty of Computer & Emerging Sciences**

**Balochistan University of Information Technology and Management Sciences, Quetta**

**April 08, 2006**

# Web Structure Mining of Dynamic Pages

By

**Muhammad Asif Naeem**

**A thesis is submitted in partial fulfillment of the requirements for the degree**

**of**

**MS Computer Science**

**Supervisor: Professor Dr. Muhammad Abbas Choudhary**

**Co-Supervisor: Professor Dr. Abdul Hussain Shah Bukhari**

**Faculty of Computer & Emerging Sciences**

**Balochistan University of Information Technology and Management Sciences, Quetta**

بسم الله الرحمن الرحيم

*YE ARE THE BEST*

**OF PEOPLES, EVOLVED**
**FOR THE MANKIND,**
**ENJOYING WHAT IS RIGHT,**

*FORBIDDING WHAT IS WRONG*

*AND THE BELIEVING IN ALLAH*

**[AL-IMRAN-110]**

**SAY YOU:" THIS IS MY WAY:**

*I DO INVITE INTO ALLAH ON*

*EVIDENCE CLEAR AS THE*

*SEEING WITH ONE'S EYES,*

*I AND WHOEVER*

*FOLLOWS ME*

**[Yousaf-108]**

# Dedicated

To

My loving Parents, Brothers and Sister

# Acknowledgement

I cannot express my feelings of thanks to almighty ALLAH, the most beneficent, the most merciful & gracious, One whose faith encouraged me in every aspect of life and never disappointed me. He paved the path for me leading to success and bright future.

I am whole heartedly grateful to my **supervisor** *Professor Dr. Muhammad Abbas Choudhary*, who himself has become an institution in himself. It is nothing but due to his sincere guidance that I have been able to accomplish my MS thesis. I respect him because of his ambitious nature, keen, quick sense of thinking, creative and imaginative mind. He has brought many radical and revolutionary changes in the Balochistan University of Information Technology and Management Sciences, Quetta.

I pay regards from the core of my heart to my **Co-supervisor** *Professor Dr. Abdul Hussain Shah Bukhari* for his coordination, supervision and valuable comments. Dr. Bukhari's guidance encouraged me to accomplish my thesis work. He is distinguish and discriminate personality in the field if Information Technology, Electronics and Telecommunications. I respect him because of his dynamic and farsightedness nature.

I cannot forget my best friends, *Mr.Imran Sarwar Bajwa* and *Mr. Riaz-Ul-Amin* for their cooperation. I will always remember all my colleagues specially *Professor Dr. Muhammad Nawaz* and other friends for their care, kindness and compassion.

Especially, I am grateful to my loving parents for providing me all sorts of moral and social support in my career endeavors. I have no words to express my deep heart feelings for my loving family.

May God give me courage to serve my country!
(Ameen)

**Muhammad Asif Naeem**

# Abstract

Web structure mining in static web contents decreases the accuracy of mined outcomes and affects the quality of decision making activity. By structure mining in web hidden data, the accuracy ratio of mined outcomes can be improved, thus enhancing the reliability and quality of decision making activity.

Data Mining is an automated or semi automated exploration and analysis of large volume of data in order to reveal meaningful patterns. The term web mining is the discovery and analysis of useful information from World Wide Web that helps web search engines to find high quality web pages and enhances web click stream analysis. One branch of web mining is web structure mining. The goal of which is to generate structural summary about the Web site and Web pages. Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level.

In recent years, Web link structure mining has been widely used to infer important information about Web pages. But a major part of the web is in hidden form, also called Deep Web or Hidden Web that refers to documents on the Web that are dynamic and not accessible by general search engines; most search engine spiders can access only publicly index able Web (or the visible Web). Most documents in the hidden Web, including pages hidden behind search forms, specialized databases, and dynamically generated Web pages, are not accessible by general Web mining applications.

Dynamic content generation is used in modern web pages and user forms are used to get information from a particular user and stored in a database. The link structure lying in

these forms can not be accessed during conventional mining procedures. To access these links, user forms are filled automatically by using a rule based framework which has robust ability to read a web page containing dynamic contents as activeX controls like input boxes, command buttons, combo boxes, etc. After reading these controls dummy values are filled in the available fields and the doGet or doPost methods are automatically executed to acquire the link of next subsequent web page.

The accuracy ratio of web page hierarchical structures can phenomenally be improved by including these hidden web pages in the process of Web structure mining. The designed system framework is adequately strong to process the dynamic Web pages along with static ones.

# Table of Contents

Chapter One

# 1 Introduction

The collection of related records is typically termed as database, and the software used to manipulate this database is referred as the Database Management System or DBMS. Moreover, a data warehouse is a logical collection of information gathered from many different operational databases used to create business intelligence that supports business analysis activities and decision-making tasks.

## 1.1 Knowledge Discovery in Databases

The Knowledge Discovery in Databases is a process of retrieving required set of data from the given repository. This process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the various steps as data cleaning, data integration, data selection, data transformation and data mining, pattern evaluation and eventually knowledge representation (Osmar R. Zaïane, 1999). A brief description of these steps is given on next page.

### 1.1.1 Data Cleaning

Data cleaning also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

### 1.1.2 Data Integration

At data integration stage, multiple data sources, often heterogeneous, may be combined in a common source.

### 1.1.3 Data Selection

In data selection phase the data relevant to the analysis is decided on and retrieved from the data collection.

### 1.1.4 Data Transformation

Data transformation, also known as data consolidation, is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

### 1.1.5 Data Mining

Data mining is the crucial step in which clever techniques are applied to reveal the patterns potentially useful.

### 1.1.6 Pattern Evaluation

In this phase strictly interesting patterns, representing knowledge are identified based on given measures.

### 1.1.7 Knowledge Representation

Knowledge representation is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

## 1.2 Data Mining

Data Mining, also known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases (W. Frawley, G. Piatetsky-Shapiro & C. Matheus, 1992). While Data Mining and KDD are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of analytical tools for analyzing data. It allows users to analyze data from many different dimensions, categorizes it, and summarizes the relationships which are identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases ( [Online], http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining .htm).

## 1.3 Text Mining

Text mining, also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Text mining is often considered a sub-field of data mining and refers to the extraction of knowledge from text documents (Chen H., 2001). Because the majority of documents on the Web are text documents, text mining for Web documents can be considered a sub-field of Web mining, or, more specifically, Web content mining. Information extraction, text classification, and text clustering are examples of text-mining applications that have been applied to Web documents.

Text Mining is an emerging research and development field that address the information overload problem borrowing techniques from data mining, machine learning, information retrieval, natural-language understanding, case-based reasoning, statistics, and knowledge management to help people gain rapid insight into large quantities of semi-structured or unstructured text. Text Mining includes several text processing and classification techniques, as text categorization, clustering and retrieval, information extraction, and others, but it also involves the development of new methods for information analysis, digesting and presentation.

A prototypical application of text mining techniques is internet information filtering. The easiness of Internet-based information publishing and communication makes it prone to misuse. For instance, Websites devoted to pornography, racism, terrorism, etc. are daily accessed easily and influenced by under age persons. Also, Internet email users have to bear intrusive unsolicited bulk email that makes it less valuable and more expensive as a communication means. Internet filtering through Text Mining techniques is a promising work field that will provide the Internet community with more accurate and cheap systems for limiting youngster's access to illegal and offensive internet contents and for alleviating the unsolicited bulk email problem.

## 1.4 Web Mining

One of the important technologies developed at the intersection of data analysis and Web technologies is Web mining. Web mining is a collection of methods and tools offering insights into behavior of the Web server users. The area has started its intensive development in 1999 taking its main ideas from conventional approaches of data mining. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a basic substance, such as mining gold from the earth.) Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help to qualify the success of a marketing campaign. Web mining consists of following three techniques:

- Web content mining
- Web usage mining
- Web structure mining

### 1.4.1 Web Content Mining

Web content mining describes the automatic search of information resources available online and involves mining Web data contents. In the Web mining domain, Web content mining is analogous to data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. The Web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of Web data forces the Web content mining towards a more complicated approach. The Web content mining is differentiated from two different points of view: Information Retrieval View and Database View. For the semi-structured data, all work utilizes the HTML structures inside the documents and some utilize the hyperlink structure between the documents for their representation. As for the database point of view data is fully organized in structured form.

Multimedia data mining is part of the content mining, which is engaged to mine the high- level information and knowledge from large online multimedia sources. Multimedia data mining on the Web has gained many researchers' attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done.

Web content mining aims to extract useful information or knowledge from Web page contents. Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the Web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and text mining techniques and also its own unique approaches.

### 1.4.2 Web Usage Mining

Web usage mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs, which contain information about the referring pages for each page reference and user registration or survey data gathered via CGI scripts. Web usage mining finds patterns in Web server logs. The logs are preprocessed to group requests from the same user into sessions. A session contains the requests from a single visit of a user to the Website. During the preprocessing, irrelevant information for Web usage mining such as background images and unsuccessful requests is ignored the users are identified by the IP addresses in the log and all requests from the same IP address with in a certain time window are put into a session.

Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web. The potential strategic aims in each domain into mining goal are: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources, which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. The Web usage mining is parsed into three distinct phases: preprocessing, pattern discovery, and pattern analysis. It is a better approach to define the usage mining procedure. It also clarified the research sub direction of the Web usage mining, which facilitates the researchers to focus on each individual process with different applications and techniques.

### 1.4.3 Web Structure Mining

Web structure mining focused on the analysis of the link structure of the Web, and one of its purposes is to identify documents, which are pointed to or pointed by many relevant Web pages. The idea is to generate Web communities among pages linked with each other.

Most of the Web information retrieval tools only use the textual information, while ignore the link information that could be very valuable. The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites. Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema. The structural information generated from the Web structure mining includes the followings:

1. The information about measuring the frequency of local links for the Web tuples in a Web table.

2. The information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document.

3. The information measuring the frequency of Web tuples in a Web table that contains links, that are global and those that span different Web sites.

4. The information measuring the frequency of identical Web tuples, appear in a Web table or among the Web tables.

In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, and both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlinks in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient.

Web structure mining has a relation with the Web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the Web. It's quite often to combine these two mining tasks in an application.

## 1.5 Background of Study

Web mining research can be divided into three categories: Web content mining, Web structure mining, and Web usage mining (Kosala R. & Blokeel H., 2000).

Web content mining refers to the discovery of useful information from Web content, including text, images, audio, and video. Web content mining is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization.

Web usage mining focuses on using data mining techniques to analyze search or other activity logs to find interesting patterns. One of the main applications of Web usage mining is to develop user's profile (Armstrong R., Freitag D., Joachims T. & Mitchell T. , 1995).

Web structure mining studies potential models underlying the link structures of the Web. It usually involves the analysis of in-links and out-links, and has been used for search engine result ranking and other Web applications (Brin S. & Page L., 1998).

In recent years, Web link structure has been widely used to infer important information about Web pages. Web structure mining has been largely influenced by research in social network analysis and citation analysis. Citations (linkages) among Web pages are usually indicators of high relevance or good quality. We use the term in-links to indicate the hyperlinks pointing to a page and the term out-links to indicate the hyperlinks found in a page. Usually, the larger the number of in-links, the more useful a page is considered to be. The rationale is that a page referenced by many people is likely to be more important than a page that is seldom referenced. As in citation analysis, an often cited article is presumed to be better than one that is never cited. In addition, it is reasonable to give a link from an authoritative source (such as Yahoo!) a higher weight than a link from an unimportant personal home page.

## 1.6 Problem Statement

The goal of Web structure mining is to generate structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites. In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, one would like to discover the relationships among those Web pages. The relations may fall in one of the types, such as they are related by synonyms or ontology, they may have similar contents, and both of them may sit in the same Web server created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient.

At present the Web structure mining is applied only upon static Web links. As the majority of documents on the Web are dynamic and not accessible by general Web mining applications because links of target pages are hidden behind user forms, specialized databases, and dynamically generated Web pages. Therefore, Web structure mining technique can not be applied upon these dynamic links that decreases the accuracy of mined outcomes and also affects the quality of decision making activity.

## 1.7 Proposed Solution

Dynamic content generation is used in modern Web pages and user forms are used to get information from a particular user and stored in a database. The link structure lying in these forms can not be accessed during conventional mining procedures. To access these links, user forms are filled automatically by the designed system which has robust ability to read a Web page containing dynamic contents as *activeX* controls like input boxes, command buttons, combo boxes, etc. After reading these controls dummy values are filled in the available fields and the *doGet* or *doPost* methods are automatically executed to acquire the link of next subsequent Web page.

By applying the Web structure mining in Web hidden data, the accuracy ratio of mined outcomes can be improved that enhance the reliability and quality of decision making activity.

## 1.8 Scope of Research

In the scope of thesis, from three defined types of Web mining only Web structure mining has been considered. Web structure mining focuses on using the analysis of the link structure of the Web, and one of its purposes is to identify documents, which are pointed to or pointed by many relevant Web pages. The idea is to generate Web communities among pages linked with each other, categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

The approach used in the thesis is to solve the problem by automatically filling the dynamic pages with dummy data using a designed algorithm which has robust ability to read a Web page containing dynamic contents.

Chapter Two

# 2 Literature Review

Web mining techniques are used to extract the required information efficiently and more accurately. Following areas, shown in Figure 2.1, represent the current research advancements related to Web mining.

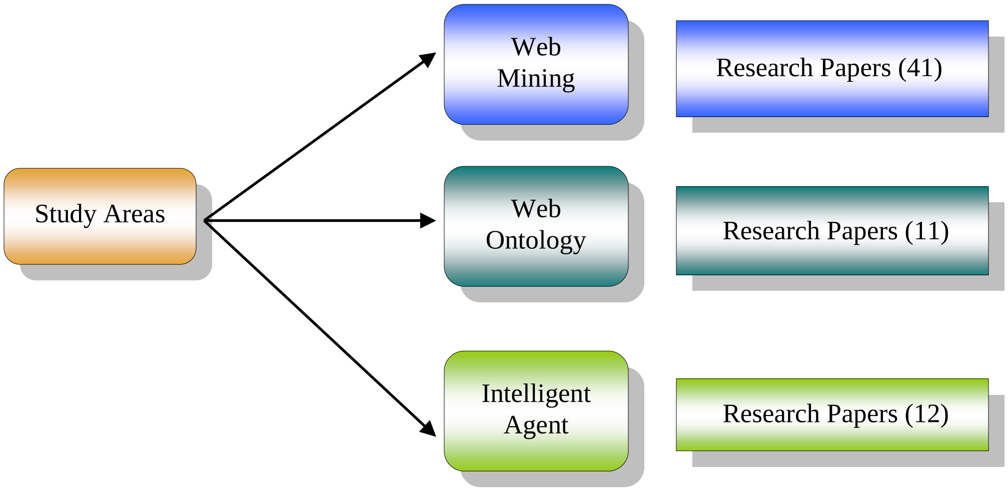


**Figure 2.1: Research survey in related areas**

- Web Mining
- Web Ontology
- Intelligent Agents

## 2.1 Web Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and Web spiders. Structure mining is used to examine data related to the structure of a particular website and usage mining is used to examine data related to a particular user's log files as well as data gathered by forms the user may have submitted during Web transactions. Figure 2.2 represents the Hierarchical representation of Web mining structure.

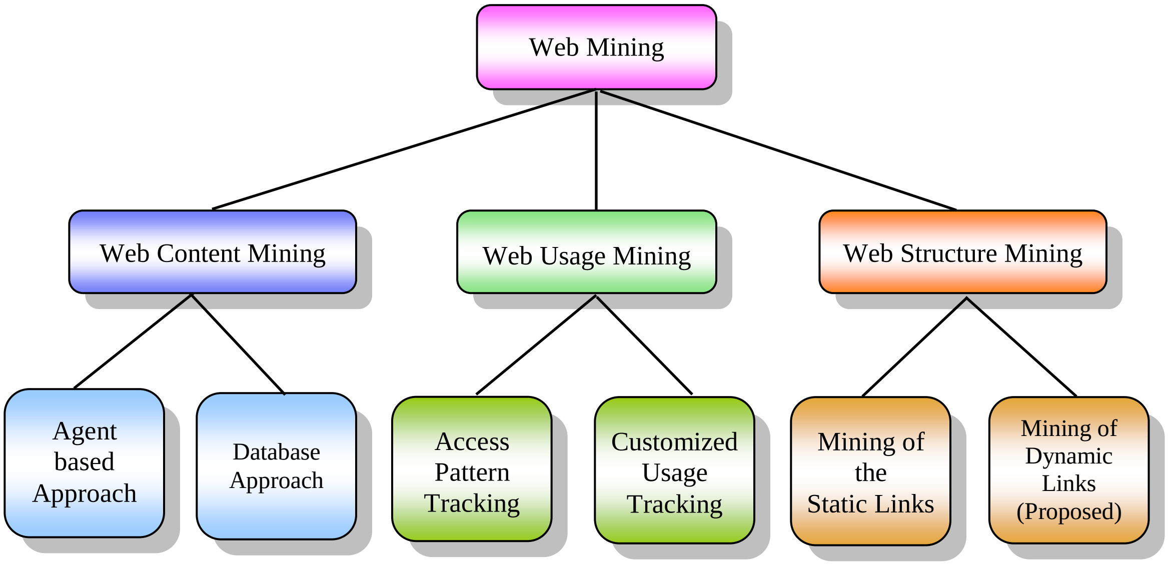


**Figure 2.2: Hierarchical representation of Web mining structure**

Web mining is a very hot research topic which combines two of the activated research areas: Data mining and World Wide Web. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Although there exist some confusions about the Web mining, but the most recognized approach is to categorize Web mining into three areas: Web content mining, Web structure mining, and Web usage mining. Web content mining focuses on the discovery/retrieval of the useful information from the Web contents/data/documents, while the Web structure mining emphasizes to the discovery of how to model the underlying link structures of the Web. The distinction between these two categories isn't very clear sometimes. Web usage mining is relatively independent, but not isolated category, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behavior (Yan Wang, 2000). Generally, Web usage mining consists of three phases: Pre-processing, Pattern discovery and Pattern analysis. A detailed description will be given for each part of them, however, special attention will be paid to the user navigation patterns discovery and analysis while, the user privacy is another important issue. An example of a prototypical Web usage mining system, WebSIFT, will be introduced to make it easier to understand the methodology of how to apply data mining techniques to large Web data repositories in order to extract usage patterns.

Currently, search tools are plagued by the given four problems. Firstly, the phenomenon of hundreds of irrelevant documents being returned in response to a search query is known as abundance problem; secondly, a limited coverage of the Web; thirdly, a limited query interface based on syntactic and fourthly a limited customization to individual users. Data mining techniques, like association rules classification, clustering and outlier detection are reviewed in (Minos N., Garofalakis, Rajeev Rastogi & Seshadri S., 2001). His study provides a brief description of each technique as well as efficient algorithms for implementing Web hyper text and hyper link structure.

A problem that there is no established vocabulary, leads to confusion when comparing research efforts was pointed out. To achieve this goal a definition of Web mining was proposed, and developed taxonomy of the various ongoing efforts related to it (Cooley R., Mobasher B. & Srivastava J., 1997). This study also provided a general architecture of a system for Web usage mining after performing a detailed survey of the efforts in this area.

Currently, most Web usage mining research has been focusing on the Web server side. Since the main purpose of the research is to improve a Web site’s service and the server’s performance (Yongjian Fu & Ming-Yi Shih, 2002). But it was argued that an equally important and potentially fruitful aspect of Web usage mining is the mining of client side usage data. It was also proposed the mining of client side Web usage data in term of personal Web usage mining and designed a framework for it. The proposed approach uses multi agents and works automatically. In addition the applications of suggested framework are not limited to agents, but may include many others such as Web personalization, learning and security.

In real time approaches that involve explicit semantics and human quality control are needed in high commitment domains that require correct and exhaustive knowledge such as science or business. In addition, an explicit conceptualization enables people and programs to explain, reason, and argue bout meaning and thus rationalize their trust, or lack of trust, in a system. It was suggested that Semantic Web mining may be the better support for the development of principled feedback loops that consolidate the knowledge extracted by mining into information available for the Web at large (Battina Berendt, Andreas Hotho & Gerd Stumme, 2001). In addition, it will enable to integrate results from machine learning and mass collaboration, and to reason about the newly emerging objects in today’s highly dynamic Web spaces.

Web personalization is the process of customizing the content and a structure of a Website to the specific and individual needs of each user without requiring them to ask for explicitly (Magdalini Eirinaki & Michalis Vazirgiannis, 2003a). This can be achieved by taking advantage of the user’s navigational behavior as revealed through the processing of a Web usage logs, as well as the user’s characteristic’s and interests. Such information can be further analyzed in association with the content of Website, resulting in improvement of the system performance, user’s retention and site modification. The overall process of Web personalization consists of five modules, namely: user profiling, log analysis and Web usage mining, information acquisition, content management and Website publishing. User profiling is the process of gathering information specific to each visitor to a Website either implicitly, using the information hidden in the Web logs or technologies such as cookies, or explicitly, using registration forms, questionnaires, and the like. Such information can be demographic, personal or even information concerning the user’s navigational behavior. However, many of the methods used in user in profiling raise some privacy issues concerning the disclosure of the user’s personal data, therefore they are not recommended. Because user profiling seems essential in the process of Web personalization, a legal and more accurate way of acquiring such information is needed.

The main component of the Web personalization system is the usage miner. Logs analysis and Web usage mining is the procedure where the information stored in the Web server logs is processed by applying statistical and data mining techniques such as clustering, association rules discovery, classification and sequential pattern discovery in order to reveal useful patterns that can be further analyzed. Such patterns differ according to the method and input data used and can be user and page clusters, usage patterns and correlations between user groups and Web pages. Those patterns can then be stored in database or a data cube and query mechanism or OLAP operations can be performed in combination with visualization techniques. The most important phase of Web usage mining is data filtering and processing. In that phase Web log data should be cleaned and enhanced, and user, and session and page view identification should be performed. Web personalization is a domain that has been recently gaining the great momentum not only in the research area, where many research teams have addressed this problem from different perspective, but also in the industrial area, where there exists a variety of tools and applications addressing one or more models of the personalization process.

With the explorative growth of information on the Web, it has become more difficult to access relevant information from the Web. Web Personalization is possible approach to solve this problem. In semantic Web, user access behavior model can share as ontology. Agent software can then utilize it to provide personalized services such as recommendation and search. A Web usage mining approach for semantic Web personalization was proposed (Baoyao Zhou, Siu Cheung Hui & Alvis C. M. Fong, 2005). The proposed approach first incorporates fuzzy logic into formal concept analysis to mine user access data for automatic ontology generation and then applies approximate reasoning to generate personalized usage knowledge from the ontology for providing personalized services.

Today huge amount of expertise knowledge are readily available online. However content creation tools concentrate on the publishing of information without considering the need for structural organization. As a result the flat distribution of knowledge makes consuming these materials difficult. It was suggested that Web mining is believed to be one of the solutions to manage these materials and to uncover other hidden insights (Kok-Leong Ong, Wee-Keong NG & EE-Peng Lim, 2003). Using data mining on Web data, the efficiency and accuracy of search engines can be improved. Knowledge contained in Web documents can be organized hierarchically. Correlated documents can be identified and hence, improve the Web as a better repository of knowledge. The architecture is built using component concepts with XML as the fabrics for inter component communication. Interfaces are part of the platform in order to interpolate with variety of technologies such as software agent sand services provider models.

As knowledge mining is a statistical approach and it leverages massive amounts of Web data but there are so many natural language processing challenges that are major bottleneck in better accuracy. It was proposed a strategy for the combination of knowledge annotation and knowledge mining, used into a question answering system that is capable of providing users with concise answers (Jimmy Lin & Boris Katz, 2003). The design system arena accepts user questions and they are sent to sub components as knowledge annotation and knowledge mining. Both components the World Wide Web to generate candidate answers, which are then piped through a knowledge boosting modules, that checks the candidate answer against a number of heuristic to ensure their validity. Knowledge annotation techniques are used to handle the head of the curve and use knowledge-mining techniques to handle its tail.

Text research engines scan individual assets and return ranked results. A theory was presented that enables query processing and joining of text resources by structuring them (Karsten Winkler & Myra Spiliopoulou, 2001). Derivation of an XML Document Type Definition DTD is overall domain specific text. The semantic characterization of text unit is core of the presented approach as well as the derivation of XML tags from these characterizations. These tags are combined and they reflect the semantics of the archives contents and have derived a set of statistical properties that reflect the quality of this approximation for whole DTD and for relationships among these tags. The statistical properties of tags and of the relationship from the basis for combining them into complete DTD in the XML schema.

It is difficult to detect terror related activities on the Web and study the typical behavior of terrorist.  However an innovative knowledge based methodology was presented for terrorist detection using Web traffic content as the audit information (Y. Elovici, A.Kandel, M.Last, B.Shapira & O.Zaafrany, 2001). Typical behavior of terrorist is study by applying a data mining algorithm to the textual content of error related Websites methodology. A knowledge based methodology for terrorist activity on the Web consists of multi steps.

First is document representation that captures relationships between terms in a textual document. Second step is similarity measure, which classifies the pages on the basis of similarity of the contents. Third step is detection methodology in which an anomaly detection system is developed for detecting abnormal contents. It may be an indication of terrorist for other criminal activity. Detection may also be pictures or binaries that are downloadable from terror related sites.

With the explosive growth of data available on the internet, personalization of this information space becomes a necessity. An important component of Web personalization is the automatic knowledge extraction from Web logs. However, analysis of large Web logs is a complex task not fully addressed by existing Web access analyzers. Using commercial software, some well-known data mining techniques (association rules and clustering) were applied to analyze access log records collected on a Web newspaper (Paulo Batista & M´ario Silva J., 2002). This paper identifies several reading patterns and discusses approaches for mining this data.

These patterns will define user profiles that integrate a news recommendation system based on Web user preferences. Frequent sets and clustering produced different patterns. Frequent sets show groups of sections that are more frequent together, independently of the user profiles, and clustering show groups of sections that define similar Web usage. Clustering of Boolean and numerical data leads to different results. While for Boolean data results are similar in both kinds of sessions and clustering approaches, it is obtained different reading patterns for numerical data, or no reading patterns at all. We detected that a very significant number of sessions consist of a single page-view referred from a site external to the online newspaper. This may explain why we were able to identify patterns in Boolean data and had more difficulty when dealing with numerical data. This suggests that to find more interesting patterns it is necessary to remove these sessions from the repository. The clustering results on numerical data may also be an outcome of the Euclidean distance-based similarity measures that are not adequate for mining our Web access data. Commonly used clustering algorithms such as K-means, were developed for data samples from Gaussian populations.

In order to exploit the rapid growth of digital documents in the internet, there is an urgent need to efficiently determine and extract relevant information from these documents. An approach was discussed to extract citation information from digital documents (Ying Ding, Gobinda Chowdhury & Schubert Foo, 1999). Four templates are produced by using template mining techniques, one for extracting information from citing articles and the other three for extraction information from cited articles (citations). These are subsequently applied to the chosen domain of Library and Information Science (LIS). The sub-languages of citations are examined, and the flowcharts of the four templates are described in detail. This study further describes the evaluation that was carried out manually, the results obtained and the limitations of this study. It was also proposed two approaches for automatically building up universal citation database: standardized template mining and universal Web authoring tool based on metadata and markup language.

Natural Language Processing (NLP) systems that are more frequently used in clinical study, methods for interpreting the output of these programs become increasingly important (Adam Wilcox, M.A., George Hripcsak & M.D, 1998). These methods require the effort of a domain expert, who must build specific queries and rules for interpreting the processor output. Knowledge discovery and data mining tools can be used instead of a domain expert to automatically generate these queries and rules. A decision tree generator was used to create a rule base for a natural language understanding system. A general-purpose natural language processor using this rule base was tested on a set of 200 chest radiograph reports. When a small set of reports, classified by physicians, was used as the training set, the generated rule base performed as well as lay persons, but worse than physicians. When a larger set of reports, using ICD9 coding to classify the set, was used for training the system, the rule base performed worse than the physicians and laypersons. It appears that a larger, more accurate training set is needed to increase performance of the method.

Given the rate of growth of the Web, scalability of search engines is a key issue, as the amount of hardware and network resources needed is large, and expensive. In addition, search engines are popular tools, so they have heavy constraints on query answer time.

So, the efficient use of resources can improve both scalability and answer time. It was suggested a tool to achieve these goals is called Web mining (Ricardo Baeza-Yates, 2004). Web mining has three branches: link mining, usage mining, and content mining. One important analysis in all these cases is the dynamic behavior. Here examples are given for link and usage mining related to search engines, as well as the related Web dynamics.

The knowledge acquisition bottleneck problem, well known to the Knowledge Management community, is turning the weaving of the Semantic Web (SW) into a hard and slow process. Nowadays' high costs associated with producing two versions of a document – one version for human consumption and another version for machine consumption – prevent the creation of enough metadata to make the SW realizable. There are several potential solutions to the problem. The use of automated methods for semantic markup was advocated, i.e., for mapping parts of unstructured text into a structured representation such as ontology (José Iria & Fabio Ciravegna, 2005). In this paper, initial work on a general software framework is described for supervised extraction of entities and relations from text. The framework was designed so as to provide the degree of flexibility required by automatic semantic markup tasks for the Semantic Web.

Data Quality Mining (DQM) is a new and promising data mining approach from the academic and the business point of view (Jochen Hipp, Ulrich G untzer & Udo Grimmer, 2001). The goal of DQM is to employ data mining methods in order to detect, quantify, explain and correct data quality deficiencies in very large databases. Data quality is crucial for many applications of Knowledge Discovery in Databases (KDD). So a typical application scenario for DQM is to support KDD projects, especially during the initial phases. Moreover, improving data quality is also a burning issue in many areas outside KDD. That is, DQM opens new and promising application fields for data mining methods outside the field of pure data analysis.

A method was presented for mining the Web in order to extract lexical patterns that help in discriminating the senses of a given polysemic word (R. Guzmán-Cabrera, P. Rosso, M. Montes-y-Gómez3 & J. M. Gómez-Soriano, 2005). These patters are defined as sets and sequences of words strongly related to each sense of the word. To discover the patterns, the method first determines the different senses of the word from a reference lexical database, and then it uses the set of synonyms from each sense as search patterns on the Web. The purpose is to create a corpus of usage cases per sense, downloading snippets via fast search engines. Finally, it applies a well-known association discovery data mining technique to select the most relevant lexical patterns for each word sense. The preliminary results indicate that making sense out of the Web is possible and the discovered patters should be of great benefit in tasks such as information retrieval and machine translation.

The rapid expansion of the Web is causing the constant growth of information, leading to several problems such as an increased difficulty of extracting potentially useful knowledge. Web content mining confronts this problem gathering explicit information from different Web sites for its access and knowledge discovery. Its current methods focus on analyzing static Web sites and cannot deal with constantly changing Web sites, such as news sites. A method for mining online news sites was proposed (A. Méndez-Torreblanca, M. Montes -y-Gómez & A. López-López, 2002). This method applies dynamic schemes for exploring these Web sites and extracting news reports, and uses domain independent statistical analysis for trend analysis. The overall method is an application of Web mining that goes beyond straightforward news analysis, trying to understand current society interests and to measure the social importance of ongoing events.

With the phenomenal growth of the Web, there is an ever-increasing volume of data and information published in numerous Web pages. The research in Web mining aims to develop new techniques to effectively extract and mine useful knowledge or information from these Web pages. Due to the heterogeneity and lack of structure of Web data, automated discovery of targeted or unexpected knowledge is a challenging task. It calls for novel methods that draw from a wide range of fields spanning data mining, machine learning, natural language processing, statistics, databases, and information retrieval. In the past few years, there was a rapid expansion of activities in the Web mining field, which consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from Web page contents. This special issue was focused on Web content mining (Bing Liu & Kevin Chen-Chuan Chang, 2004). The objectives of this special issue are two-fold:

1. To bring together and to present some of the latest research results in the field.
2. To encourage more research activities in the field. With the huge amount of data/information already on the Web and more to come, the next big thing is naturally how to make best use of the Web to mine useful data/information and to integrate heterogeneous data/information automatically.

Web structure mining has been a well-researched area during recent years. However, It was observed that data on the Web is changing at any time in any way, even though there are some incremental data mining algorithms that are proposed to update the mining results with the corresponding changes, none of the existing Web structure mining techniques is able to extract useful and hidden knowledge from the sequence of historical Web structural changes (Qiankun Zhao Sourav Saha Bhowmick & Sanjay Kumar Madria, 2003). In this paper, it was proposed a novel research area of Web structure mining named Web structural delta mining. The distinct feature of research is that the mining object is the sequence of historical changes of Web structure (Web Structural Deltas). For Web structural delta mining, the major aim is to extract useful, interesting, and novel Web structures and knowledge considering their historical, dynamic, and temporal properties. Three major issues of Web structural delta mining were also proposed like identify useful and interesting structures, extract association from changes of Web structure, and cluster Web structure based on patterns of changes. Also, the challenges of Web structural delta mining corresponding to different issues are discussed in this paper.

A methodology for bootstrapped learning was proposed in order to extract information from Web sites, using very limited amount of user input (Alexiei Dingli, Fabio Ciravegna, David Guthrie & YorickWilks, 2003). That methodology was exemplified by using a specific application, but the methodology is generic and can be safely extended to a number of other tasks by specifying different Web resources. In the specific application, the only user input is a number of examples of the information to be extracted. In other tasks, some limited manual annotation of examples could be the right way. What is important that the amount of user input can be dramatically reduced and when compared with fully supervised methodologies. The described methodology is applicable to cases here the information is likely to be highly redundant and where regularities in documents can be found. This is often the case of many repositories used for knowledge management and of Web pages belonging to specific communities (e.g. Computer Science Web sites, e-commerce sites, etc.). Other authors have shown that similar (but less sophisticated) methodologies can be successfully applied to retrieve very generic relations on the whole Web. Recent advances on wrapper induction systems show that the regularity required to induce wrappers is not as rigid as it used to be in the past. Current wrapper induction systems can very often be used on free texts making the methodology quite generic. Qualitative analysis of results from preliminary experiments is satisfying. When Armadillo was run on a number of sites (such as *www.nlp.shef.ac.uk* and *www.iam.ecs.soton.ac.uk*), it managed to find most information using just a user-defined list of projects for the first site.

Key phrases are useful for a variety of purposes, including summarizing, indexing, labeling, categorizing, clustering, highlighting, browsing, and searching. The task of automatic key phrase extraction is to select key phrases from within the text of a given document. Automatic key phrase extraction makes it feasible to generate key phrases for the huge number of documents that do not have manually assigned key phrases. A limitation of previous key phrase extraction algorithms is that the selected key phrases are occasionally incoherent. That is, the majority of the output key phrases may fit together well, but there may be a minority that appears to be outliers, with no clear semantic relation to the majority or to each other. Some enhancements were presented to the Kea's key phrase extraction algorithm that was designed to increase the coherence of the extracted key phrases (Peter D. Turney, 2003). The approach is to use the degree of statistical association among candidate key phrases as evidence that they may be semantically related. The statistical association is measured using Web mining. Experiments demonstrate that the enhancements improve the quality of the extracted key phrases. Furthermore, the enhancements are not domain-specific: the algorithm generalizes well when it is trained on one domain (computer science documents) and tested on another (physics documents).

A research was introduced on learning browsing behavior models for inferring a user's information need corresponding to a set of words) based on the actions that was taken during the current Web session (Tingshao Zhu, Russ Greiner, Gerald H¨aubl, Kevin Jewell & Bob Price, 2005). This information is then used to find relevant pages, from essentially anywhere on the Web. The models, learned from over one hundred users during a five weeks user study, are session-specific but independent of both the user and Website. The empirical results suggest that these models can identify and satisfy the current information needs of users, even if they browse previously unseen pages containing unfamiliar words.

The purpose of Web mining is to develop methods and systems for discovering models f objects and processes on the World Wide Web and for Web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data mining, Internet Technology and World Wide Web, and for the more recent Semantic Web. The World

Wide Web has made an enormous amount of information electronically accessible. The use of email, news and markup languages like HTML allows users to publish and read documents at a world-wide scale and to communicate via chat connections, including information in the form of images and voice records. The HTTP protocol that enables access to documents over the network via Web browsers created an immense improvement in communication and access to information. For some years these possibilities were used mostly in the scientific world but recent years have seen an immense growth in popularity, supported by the wide availability of computers and broadband communication. The use of the internet for other tasks than finding information and direct communication is increasing, as can be seen from the interest in “e-activities” such as e-commerce, e-learning, e-government, e-science. Independently of the development of the Internet, Data mining expanded out of the academic world into industry. Methods and their potential became known outside the academic world and commercial toolkits became available that allowed applications at an industrial scale. Numerous industrial applications have shown that models can be constructed from data for a wide variety of industrial problems. The World Wide Web is an interesting area for data mining because huge amounts of information are available. Data mining methods can be used to analyze the behavior of individual users, access patterns of pages or sites, properties of collections of documents. Almost all standard data mining methods are designed for data that are organized as multiple “cases” that are comparable and can be viewed as instances of a single pattern. A “case” is typically described by a fixed set of features (or variables). Data on the Web have a different nature. They are not so easily comparable and have the form of free text, semi-structured text (lists, tables) often with images and hyperlinks, or server logs. The aim to learn models of documents has given rise to the interest in Text Mining methods for modeling documents in terms of properties of documents. Learning from the hyperlink structure has given rise to graph-based methods, and server logs are used to learn about user behavior. The Semantic Web is a recent initiative, inspired by Tim Berners-Lee, to take the World-Wide Web much further and develop in into a distributed system for knowledge representation and computing. The aim of the Semantic Web is to not only support access to information “on the Web” by direct links or by search engines but also to support its use. Instead of searching for a document that matches keywords, it should be possible to combine information to answer questions. Instead of retrieving a plan for a trip to Hawaii, it should be possible to automatically construct a travel plan that satisfies certain goals and uses opportunities that arise dynamically. This gives rise to a wide range of challenges. Some of them concern the infrastructure, including the interoperability of systems and the languages for the exchange of information rather than data. Many challenges are in the area of knowledge representation, discovery and engineering. They include the extraction of knowledge from data and its representation in a form understandable by arbitrary parties, the intelligent questioning and the delivery of answers to problems as opposed to conventional queries and the exploitation of formerly extracted knowledge in this process. The ambition of representing content in a way that can be understood and consumed by an arbitrary reader leads to issues in which cognitive sciences and even philosophy are involved, such as the understanding of an asset's intended meaning.

Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. Web Mining aims at discovering insights about the meaning of Web resources and their usage. Given the primarily syntactical nature of data Web mining operates on, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web resources and navigation behavior are increasingly being used. This fits exactly with the aims of the Semantic Web: the Semantic Web enriches the WWW by machine process able information that supports the user in his tasks. The interplay of the Semantic Web with Web Mining was discussed, having a specific focus on usage mining (Gerd Stumme, Andreas Hotho & Bettina Berendt, 2002). It is also discussed how Semantic Web Usage Mining can improve the results of 'classical' usage mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques. A truly semantic understanding of Web usage needs to take into account not only the information stored in server logs, but also the meaning that is constituted by the sets and sequences of Web page accesses. One important focus is to make search engines and other programs able to better understand the content of Web pages and sites. This is reflected in the wealth of research efforts that model pages in terms of an ontology of the content.

An important approach to text mining involves the use of natural-language information extraction. Information extraction (IE) distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. IE systems can be used to directly extricate abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analyzed with traditional data-mining techniques to discover more general patterns. Different methods and implemented systems for both of these approaches were discussed and results were summarized on mining real text corpora of biomedical abstracts, job announcements, and product descriptions (Raymond J. Mooney & Razvan Bunescu, 2005). We also discuss challenges that arise when employing current information extraction technology to discover knowledge in text.

New XML applications were proposed, XGMML and LOGML (Amir H. Youssefi, David J. Duke & Mohammed J. Zaki, 2004a). XGMML is a graph description language and LOGML is a Web-log report description language. It was generated a Web graph in XGMML format for a Web site using the Web robot of the WWW Pal system (developed for Web visualization and organization). It is also generate Web-log reports in LOGML format for a Web site from Web log files and the Web graph. In this paper the usefulness of these two XML applications with a Web data mining example is also illustrated. Moreover, the simplicity with which this mining algorithm can be specified and implemented efficiently using two XML applications is also shown.

Analysis of Web site usage data involves two significant challenges: firstly the volume of data, arising from the growth of the Web, and secondly, the structural complexity of Web sites. Different Data mining and Information Visualization techniques were applied to the Web domain in order to benefit from the power of both human visual perception and computing; and termed this Visual Web Mining (Amir H. Youssefi, David J. Duke & Mohammed J. Zaki, 2004b). In response to the two challenges, a generic framework is proposed, where Data mining techniques are applied to large Web data sets and use Information Visualization methods on the results. The goal is to correlate the outcomes of mining Web Usage Logs and the extracted Web Structure by visually superimposing the results.

It was proposed using a stratified fuzzy cognitive map (FCM) to amplify inference results of Web mining as a dramatic usage of the Internet for a wide variety of daily management activities (Kun Chang Lee, Jin Sung Kim & Namho Chung, 2004), Web mining becomes one of the intelligent techniques to provide robust decision support. However, conventional Web mining approaches have failed to offer enriched inference results due to the lack of understanding causal knowledge hidden in the Web mining results. In this sense, it is proposed using a stratified FCM to overcome this pitfall of the conventional Web mining approaches.

Data analysis for Web applications requires a number of challenging choices: the technologies for data preparation and cleaning, the Web mining methods that operate on large volumes of user and usage data, and last but not least the evaluation models that assess the quality of the findings with respect to Web design principles, user expectations, and business or other application objectives. First, a general framework for capturing different perspectives on Web mining was described (Bettina Berendt, Ernestina Menasalvas & Myra Spiliopoulou, 2005a). Referring to the CRISP-DM industrial standard for knowledge discovery processes, the central role of evaluation in this framework is elaborated. Then it was presented (Bettina Berendt, Ernestina Menasalvas & Myra Spiliopoulou, 2005b) the two most prominent examples of perspectives to be considered in evaluation for Web mining applications: User-oriented models of goals, goal achievement, and success conceptualize success as the usability of a site, that is, the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in the site environment. Business-oriented models of goals, goal achievement, and success conceptualize success as the ability of a site to turn visitors into (repeat) buyers, and in general as the site's ability to generate revenue and contribute to other strategic goals. Finally, a case study was presented on multi-channel retailing to demonstrate the use of business-oriented evaluation measures for relating the mining results to a Web application's objectives.

Knowledge-intensive methods that can altogether be characterized as Deductive Web Mining (DWM) already act as supporting technology for building the semantic Web. Reusable knowledge-level descriptions may further ease the deployment of DWM tools.

A multi-dimensional, ontology-based framework was developed, and problem-solving methods were designed (Vojt¡ech Sv´atek, Martin Labsk´y & Miroslav Vacura, 2003), which enable to characterize DWM applications at an abstract level. It was shown that the heterogeneity and unboundedness of the Web demands for some modifications of the problem-solving method paradigm used in the context of traditional artificial intelligence.

The popularity of digital images is rapidly increasing due to improving digital imaging technologies and convenient availability facilitated by the Internet. However, how to find user-intended images from the Internet is non-trivial. The main reason is that the Web images are usually not annotated using semantic descriptors. It was presented an effective approach and a prototype system for image retrieval from the Internet using Web mining (Zheng Chen, Liu Wenyin, Feng Zhang, Mingjing Li & Hongjiang Zhang, 2001). The system can also serve as a Web image search engine. One of the key ideas in the approach is to extract the text information on the Web pages to semantically describe the images. The text description is then combined with other low-level image features in the image similarity assessment. Another main contribution of this work is that data mining is applied on the log of user's feedback to improve image retrieval performance in three aspects: First, the accuracy of the document space model of image representation obtained from the Web pages is improved by removing clutter and irrelevant text information; Second, to construct the user space model of users' representation of images, which is then combined with the document space model to eliminate mismatch between the page author's expression and the user's understanding and expectation; Third, to discover the relationship between low-level and high-level features, which is extremely useful for assigning the low-level features' weights in similarity assessment.

The area of Knowledge Discovery in Text (KDT) and Text Mining (TM) is growing rapidly mainly because of the strong need for analyzing the vast amount of textual data that reside on internal file systems and the Web. An overview of this area is provided and feature classification scheme is proposed that can be used to compare and study text mining software (Haralampos Karanikas & Babis Theodoulidis, 2003). This scheme is based on the software's general characteristics, and text mining features. Then feature classification scheme is applied to investigate the most popular software tools, which are commercially available. The major intention is not to conduct an extensive presentation of every tool in the area; only representative tools for each of the main approaches identified are considered.

Web mining is used to automatically discover and extract information from Web-related data sources such as documents, log, services, and user profiles. Although standard data mining methods may be applied for mining on the Web, many specific algorithms need to be developed and applied for various purposes of Web based information processing in multiple Web resources, effectively and efficiently. An abstract Web mining model was proposed for extracting approximate concepts hidden in user profiles on the semantic Web (Yuefeng Li a & Ning Zhong, 2004). The abstract Web mining model represents knowledge on user profiles by using an ontology which consists of both "part-of" and "is-a" relations.

## 2.2 Web Ontology

Ontology is a systematic computer oriented representation of the world. Each ontology provides the vocabulary (or names) for referring to the terms in a subject area, as well as the logical statements that describe what the terms are, how they are related to each other, how they can or cannot be related to each other, as well as rules for combining terms and relations to define extensions to the vocabulary. Hence, ontologies represent a common machine-level understanding of topics that can be communicated between users and applications, i.e., domain semantics independent of reader and context. Ontology is an explicit specification of a set of objects, concepts and other entities that are presumed to exist in some area of interest and the relationship that hold them. As implied by the above general definition, an ontology is domain dependent and it is designed to be shared and reusable.

Ontologies can be applied in the following two general approaches:

1. When both ontology and the instances of ontology entities are known, this usually applies in cases where instances of ontology have been identified among the input Web data. With this additional data semantics, Web mining techniques can

discover knowledge that is more meaningful to the users. For example, ontology based Web clustering can use HTML elements corresponding to concept instances as features to derive more accurate clusters. If Web pages are concept instances ontology based Website structure mining can derive linkage pattern among concepts from Web pages for Website design improvements.

2. When only ontology is available as input semantic structures. Ontologies can also be used as background semantic structures for Web mining. For example instead of categorizing Web pages, ontology based Web page classification may classify Web pages as concept instances and Web page pairs as relationship instances. This allows Web pages to be searched using more expensive search queries involving search conditions on concept and relationships. In ontology based Web extraction, one may address the problem of extracting both HTML elements as concept instances and finding related pairs of HTML elements.

One of the key conditions for transforming large quantities of text into effective repositories of knowledge is to allow a user to "search by an idea" Text search tools cannot still match an expert looking for relevant information in a document collection. It was described that text processing can be profitably improved by the integration of advanced knowledge representation tools (Miriam Baglioni, Micro Nanni & Emiliano Giovannetti, 2004). Even literary texts, whose structure is inherently complex due to subtle elaboration of every linguistic level, can be interpreted from a semantic view point, once an adequate representation of their domain is specified current knowledge representation languages are able to account for the structure & relationships of the world of text and its external context. Moreover they are able to discover inherent conceptual chunks hidden in the representation. The system is then able to provide a user with useful answers with respect to all data at its disposal. As a test bed the author chose the electronic version of Dante's Inferno, manually tagged using XML, enriched with a domain ontology describing the historical, social and cultural context represented as a separate XML document.

World Wide Web is the most excited society in the last 20 years. Web has turned to be largest information source available in the planet. It is the huge, explosive diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information and also raise the complexity of how to deal with the information from the different perspectives of view-users, Web service providers, business analyst. The user wants to have effective search tools to find relevant information easily and precisely. Web mining is the term of applying data mining techniques to automatically discover extract useful information from Web. An automatically generated ontology was presented for a semantic Web search system using data mining techniques (K.R. Reshmy & S.K. Srivatsa, 2005). This may improve the query process and may get better semantic results. Ontology serves as metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined meaning. Ranking algorithm used here is the hyper textual ranking algorithm that scans both the contents of the documents and also the reciprocally linked documents. This technique has several advantages that include providing better semantic notion during the search. It also serves for multiple frame documents.

The World Wide Web today provides user access to extremely large number of Web sites many of which contain information of education and commercial values. Due to unstructured and semi structured nature of Web pages and the design idiosyncrasy of Web sites, it is a challenging task to develop digital libraries for organizing and managing digital content from the Web. An otology was presented as a set of concepts and their inter relationships relevant to some knowledge domain (Ee-Peng Lim & Aixin Sun, 2005). The knowledge provided by ontology is extremely useful in defining the structure and scope for mining Web content.

Over the last years, it is observed that an explosive growth in the information available on the Web. It is very difficult task to analyze and classify the Web documents to several major categories in a given using domain ontology. It was proposed a new ontology based methodology for classification of Web documents to main categories according to the user “Information Needs” (Marina Litvak, Mark Last & Slava Kisilevich, 2005). The main contribution of this work is using domain based multi lingual ontology in the conceptual representation of documents.

Some Websites on the internet are being changed dynamically due to this data extraction from these sites is very difficult. A mechanism was proposed that the original data are structured in different ways (Hicham Snoussi, Laurent Magnin & Jian-Yun Nie, 2002), it is more suitable to structure then according to common model that is independent of the information sources. Thus extracting and combining data from different sources will be much easier and more reliable. The present study uses a modeling of ontology close to object oriented (OO) modeling with the OO paradigm. One can express ontology in an explicit way and generate software elements that are easily exploitable by other applications. Author proposed a design of ontology that uses a 3-level model: basic objects, model and meta model.

The notion of ontology is very ambiguous and the term is interpreted differently in various communities, the nature as well as role of ontologies in Web information extraction may significantly vary from one project to another. Three main types of ontologies were described (Martin Labsky, Vojtech Svatek & Ondrej Svab, 2004) that are used for Web information extraction. They are domain ontologies, presentation ontologies and terminological ontologies. Domain ontologies are used for the target application in particular domain, presentation ontologies are used in heuristic in heuristic template filling and terminological ontologies used in text annotation. Domain ontologies are suitable for reasoning over real world objects. Presentation Ontologies are restricted to a smaller portion of original domain. Terminological ontologies are centered around human language terms without direct reference to real world. Their main structures are synonym set and hierarchies.

A better approach is required to facilitate the automation of Web services tasks including automated Web server discovery execution composition and mediation by using XML based metadata and ontology. It was proposed a front end agent system for ontology management and semantic Web services management (Kotaro Nakayama, Takahiro Hara & Shojiro Nishio, 2005). The proposed system has four components. First is semantic wrapper which creates a semantic database on the semantic Web standard Resource Description Framework (RDF) from various applications as male client schedulers and so on. Second component is personalize ontology which is a dictionary database prepared for each users. The hierarchical structure of concepts depends upon the user culture such as company, family region, etc. The personalized ontology is used in dialogs with the agent inference and the vocabulary for the semantic wrapper. The personalized ontology is used in dialogs with the agent, inference and the vocabulary for the semantic wrapper.

The third component RDF mining generates ontology by using natural language processing on RDF metadata. It identifies noun words and unknown words by using NLP tools. The registered will be checked to user via dialog interface and will define the relation between other existing words. The fourth and last component is Web Service Description Language (WSDL) wrapper. WSDL wrapper is semantic metadata for Web services which enables agent program to make inferences from grounding data on personalized ontology. Users can search and execute Web services by using the agent interface. The search function uses this semantic data and personal ontology to infer what Web services user wants to execute from request key word.

Ontology construction is a complex process involving different type of users and multiple tasks and it should take place in an integrated environment including other elements of organizational memory such as databases and document basis. It was proposed that various hypertext interfaces can be created easily using a declarative approach that was originally developed to enable the publication of databases (Gilles Falquet & Claire-Lise Mottaz Jiang, 2003).

Automatically generated ontologies were presented for a semantic Web search system using data mining techniques (K.R. Reshmy & S.K. Srivatsa, 2005). This will improve the query process and will get better semantic results. Ranking algorithm is used to search and analyze Web documents in a more flexible and effective way. Hyperlink structure of Web document is utilized to rank the results. Association rule mining is used to find the maximal keyword patterns. Clustering is used to group retrieved documents into distinct sets. This will extract knowledge about query from the Web, populate a knowledge base. The search engine that searches the Web documents so far are syntactic oriented. Here we develop a searching system that semantically searches the documents. The semantics of the terms is achieved using the ontologies. Ontology serves as Meta data schemas, providing a controlled vocabulary of concepts, each with explicitly defined meaning. Ranking algorithm used here is the hyper textual ranking algorithm that scans both the contents of the documents and also the reciprocally linked documents. This technique has several advantages that include providing better semantic notion during the search. It also serves for multiple frame documents. There is a need for automatic generation of ontologies when using the semantic searching system. The paper focuses on how the automatic generation of ontologies could be done for a semantic search system using data mining techniques.

Ontologies in current computer science parlance are computer based resources that represent agreed domain semantics. Unlike data models, the fundamental asset of ontologies is their relative independence of particular applications, i.e. ontology consists of relatively generic knowledge that can be reused by different kinds of applications/tasks. The first part of this paper concerned some aspects that help to understand the differences and similarities between ontologies and data models. In the second part an ontology engineering framework was presented that supports and favors the generosity of ontology (Peter Spyns, Robert Meersman & Mustafa Jarrar, 2002). Peter Spyns introduced the DOGMA ontology engineering approach that separates “atomic” conceptual relations from “predicative” domain rules. A DOGMA ontology consists of an ontology base that holds sets of intuitive context-specific conceptual relations and a layer of “relatively generic” ontological commitments that hold the domain rules.

A query system on texts was described and literary material with advanced information retrieval tools (Miriam Baglioni, Mirco Nanni & Emiliano Giovannetti, 2004). As a test bed Miriam Baglioni1 chose the electronic version of Dante’s Inferno, manually tagged using XML, enriched with a do-main ontology describing the historical, social and cultural context represented as a separate XML document. One of the key conditions for transforming large quantities of texts into effective repositories of knowledge is to allow a user to “search by an idea”. Text search tools cannot still match an expert looking for relevant information in a document collection. Answers of an expert are intelligent as he has, at his disposal, a lot of data and he is able to compute all the available information using a sophisticated reasoning process. To match the answers of a human being, the system should be able to reply the queries such as: which are Dante’s attitudes towards holders of feudal power? And is there a statistical correlation between the belonging to a feudal system and salvation, or between belonging to city system and damnation? Answering to these queries requires the evaluation of different types of knowledge: text content knowledge and context knowledge. Some queries will be solved using only one of the available sources of information while others will require a comparison between them.

Only recently the use of intelligent techniques in humanistic fields has been receiving attentions from researchers thanks to the design of new tools for text representation and the application of advanced markup tools for conceptual manipulations. The markup of several aspects of a text is the goal of the Text Encoding Initiative. By using SGML texts can be stored obtaining a rich meta-representation of their multilevel information. Hypertext representation can be used for the realization of a theory of narrative evolution, typical of certain literary trends. In “The World of Dante” Project a hypermedia environment for the study of Inferno of Divina Commedia has been implemented. New techniques, aimed at processing the meaning of a text, require the treatment of the knowledge “not only of the text itself, but of the world”. About data mining applications to literary texts the majority of the efforts have been devoted to the discovery of frequent expressions (patterns) within texts of particular authors without taking into account implicit knowledge information. In replying to queries like the ones reported above, there is the need of added in-formation that describes both the text and the context. To do this we have manually tagged the text and defined an a priori domain ontology containing the contextual information. The chosen test bed has proven to be extremely structured both from the textual and the knowledge point of view, and for this reason ideal to demonstrate the power of querying systems like the one presented in the paper.

## 2.3 Intelligent Agents related to Web Mining

Agents are defined as software or hardware entities that perform some set of tasks on behalf of users with some degree of autonomy. In order to work for some body as an assistant, an agent has to include a certain amount of intelligence, which is the ability to choose among various courses of action, plan, communicate, adapt to change in the environment and learn from experience. In general an intelligent can be described as consisting of a sensing element that can receive events, a recognizer or classifier that determines which event occurred.

In the context of intelligent agents, an event is defined as anything that happens to change the environment or anything of which the agent should be aware. For example, an event could be the arrival of a new mail, or it could be a change to a Web page.

The list of attributes often found in agents is listed as: autonomous, goal oriented, collaborative, flexible, self starting, temporal continuity, character communicative, adaptive mobile. On the other hand, the main characteristics of the tasks, where the agent technology is found suitable for, include complexity, distribution and delivery, dynamic nature, information retrieval, high volume of data handling, routine, repetitive, time critical etc. Some examples in which agent paradigms are frequently used include:

1. Taking the advantages of distributing computing resources such as multiprocessors applications and distributed artificial intelligence problems.
2. Coordinating teams of interacting robots where each robot necessarily has physically separate processor and is capable of acting independently and autonomously.
3. Increasing system robustness and reliability in situations where an agent is destroyed, other can still carry out the tasks.
4. Assisting users by reducing their work and information loads.
5. Modeling groups of interacting experts, as in concurrent engineering and other joint decision-making processes.
6. Simplifying modeling very complex processes as a set of interacting agents.
7. Modeling processes that are normally performed by multiple agents, such as economic processes involving groups of buying & selling agents.

Several types of agents have been defined, based on their abilities and more often the task, they are designed to perform.

- Filtering agents
- Information agents
- User interface agents
- Office or workflow agents
- System agents
- Brokering or commercial agents

In recent years, agents become a very popular paradigm in computing because of their flexibility, modularity and general applicability to a wide range of problems. Technological developments in distributed computing, robotics and the emergence of object orientation have given rise to such technologies to model distributed problem solving. A short survey of agent paradigm was presented in the context of information retrieval, filtering, classification and learning and possible use in data mining tasks (Ayse Yasemin Seydim, 1999). Agent based approaches are becoming increasingly important because of their generally, flexibility, modularity and ability to take advantage of distributed resources. Agents are used for information retrieval, entertainment, coordinating system of multiple robots and modeling economic system. They are useful in reducing work and information overload in complex tasks such as medical monitoring and battlefield reasoning. Agents provide an efficient framework for distributed computation where the retrieval of only documents minimizes the duration of the expensive network connection.

With the rapid development of WWW, to find a valuable knowledge in the immensity resources becomes a hard job. The search engines use the simple world matching algorithm or manual search, so they cannot satisfy the requirements of users. There are three research categories in Web mining, based on Context Mining (WCM), Structure Mining (WSM) and Usage Mining (WUM). As a semi structure, the document of Web contains Web data including wave, image and text, thus making the Web data become multi dimension, heterogeneous. The mining model available focuses much on simple pure text matching, which will use crawler searching in the net or manual finding. These methods low efficient becomes a focus in the mining area. A mining model was proposed to provide the underlying framework for efficient Web mining (Li Zhan & Liu Zhijing, 2003). In this model the knowledge caroler based on mobile features used as an effective object to collect the interesting information, which will send back the collected data to ANS (Agent Naming system). After getting these data, the expectation Machine, Generalization Machine and Analysis Machine will successively extract and cluster the valuable knowledge to stakeholders (users, information provider etc).

Scientists and intelligence analysts are interested in quickly discovering new results from the vast amount of available geographical data. The key issues that arise in this pursuit are how to cope with new and changing information and how to manage the steadily increasing amount of available data. A new agent architecture was described that has been developed and tested to address these issues by combining innovative approaches from three distinct research areas: Software agents, geo-referenced data modeling and content based image retrieval (Paul Palathingal, Thomas E. Potok & Robert M.Patton, 2005). All of the software agents in this system were deployed using the Oak Ridge Mobile Agent Community (ORMAC). The ORMAC framework allows execution of mobile, distributed software agents and establish communication among them.

With the rapid growth of the internet, people are facing the information overload that makes the users spend more time and put more efforts to find the information they need. To resolve this method of constructing navigation agents was proposed that provide more personalized Web navigation by exploiting domain in specific ontologies (Jaeyoung Yang, Hyunsub Jung & Joongmin Choi, 2005). Web pages are converted into concepts by referencing to domain specific ontologies which employee a hierarchical concept structure. This concept mapping makes it easy to handle Web pages and also provide higher level classification information.

The more than enthusiastic success obtained by e-commerce and continuous growth of the World Wide Web has radically changed the way people look for and purchase commercial products. E-retail stores offer any sort of goods through evermore appealing Web pages, sometime even including their own search engines to help customers find their loved products. With all these great mass of information available, people often get confused or simply do not want to spend time browsing the internet, loosing themselves into myriads of available e-shape and typing to compare their offers.

It was proposed that E-retail systems are the natural solution to this problem (Maria Teresa Pazienza, Armando Stellato & Michele Vindigni, 2003). They place at people disposal user friendly interfaces, helping the customers in finding products from different e-shapes that match their desired and comparing these offers for them.

With the development of technology the enterprises are suffering more pressure than ever and facing very difficulties to make decisions. It was proposed a multi agent Web text mining system on the grid to support enterprise decision (Kin Keung Lai, Lean Yu & Shouyang Wang, 2006). This paper first proposes a single intelligent agent to perform text mining. With the rapid increase of Web information, a multi-agent Web text mining system on the grid is then constructed for large scale text mining application. Author presents a framework of the Back Propagation Neural Network (BPNN) based intelligent learning agent for text mining. To scale the computational load for large scale text mining tasks, a multi agent Web text mining system on the grid is proposed.

Common agents are not very much intelligent to mine patterns precisely and efficiently. It was presented that metadata likes rules decisions and classes on test case data are embedded into agents in order to improve the existing intelligence (Andreas L. Symeonidis, Pericles, A. Mitkas & Dionisis D. Kechagias, 1998). The use of intelligent agent technology will make the application dynamically adjustable to a changing environment. A new framework with the name of agent academy has been introduced which constitutes various components as agent factory, agent use repository agent framing module and data miner. When an agent request goes to agent factory it creates new agent on user demands. On the other hand agent use repository developed the required agent tracking tools which monitor the agent to agent transaction and provides a formal description of stored data using a certain metadata. The agent training module develops techniques to enable the ability of learning to an agent and mines the information from the given data repository with the help of data miner module using association and classification rules.

A multi-agent Web text mining system on the grid was developed to support enterprise decision-making (Kin Keung Lai, Lean Yu & Shouyang Wang, 2006). First, an individual intelligent learning agent that learns about underlying text documents is presented to discover the useful knowledge for enterprise decision. In order to scale the individual intelligent agent with the large number of text documents on the Web, it is provided a multi-agent Web text mining system in a parallel way based upon grid technology. Finally, it is discussed how the multi-agent Web text mining system on the grid can be used to implement text mining services.

Data mining is the process of extraction of interesting information or patterns from data in large databases. Agents are defined as software or hardware entities that perform some set of tasks on behalf of users with some degree of autonomy. In order to work for somebody as an assistant, an agent has to include a certain amount of intelligence, which is the ability to choose among various courses of action, plan, communicate, adapt to changes in the environment, and learn from experience. In general, an intelligent agent can be described as consisting of a sensing element that can receive events, a recognizer or classifier that determines which event occurred, a set of logic ranging from hard-coded programs to rule-based inference, and a mechanism for taking action. In several steps through knowledge discovery, which include data preparation, mining model selection and application, and output analysis, intelligent agent paradigm can be used to automate the individual tasks. In the experiment setup, association rules were discovered in a distributed database using intelligent agents (Maria Teresa Pazienza, Armando Stellato & Michele Vindigni, 2003a). An original approach is applied for effective distributed mining association rules: loose-couple incremental methods.

The more than enthusiastic success obtained bye-commerce and the continuous growth of the WWW has radically changed the way people look for and purchase commercial products. E-retail stores offer any sort of goods through evermore appealing Web pages, sometimes even including their own search-engines to help customers find their loved products. With all this great mass of information available, people often get confused or simply do not want to spend time browsing the internet, loosing themselves into myriads of available eshops and trying to compare their offers. E-retail systems were proposed that provide the natural solution to this problem: they place at people’s disposal user-friendly interfaces, helping the customers in finding products from different e-shops that match their desires and comparing these offers for them (Maria Teresa Pazienza, Armando Stellato & Michele Vindigni, 2003b). Inside CROSSMARC, (a project funded by the Information Society Technologies Program of the European Union: IST 2000-25366) different techniques coming from the worlds of NLP, Machine Learning-based Information Extraction and Knowledge Representation have been considered and conjoined to give life to an agent-based system for information extraction (IE) from Web pages, which operates in a wide range of situations involving different languages and domains. This paper describes the main components that realize the CROSSMARC architecture, together with their specific role in the process of extracting information from the Web and presenting them to the user in a uniform and coherent way.

The number of Web pages available on Internet increases day after day, and consequently finding relevant information becomes more and more a hard task. However, when communities of people are considered with common interests, it is possible to improve the quality of the query results using knowledge extracted from the observed behaviors of the single users. An agent-based recommendation system was proposed for supporting communities of people in searching the Web by means of a popular search engine (Alexander Birukov, Enrico Blanzieri & Paolo Giorgini, 2004). Agents use data mining techniques in order to learn and discover users' behaviors, and they interact with one another to share knowledge about their users. The paper presents also a set of experimental results showing, in terms of precision and recall, how agent’s interaction increases the performance of the overall system. The knowledge produced from observations is used in order to suggest links or agents to a group of people and to their personal agents. The main idea is that it is not expressed this knowledge in explicit form but is used for improving the quality of further search sessions, including searches performed by new users. Personal agents produce results by asking another personal agent about links and agent IDs. Each agent has the learning capabilities that help to produce results even without interaction. The experience of community members is exploited by means of interactions when the user performs the search already done by someone else. This feature prevents the user from searching from scratch" and increases the search quality. The SICS architecture as well as implicit Culture concepts allow implicit to be a solution to the problem of finding necessary information on the Web. One of the main advantages of this approach is represented by the use of both search engine results and suggestions produced by community members. The multi-agent system mimics natural user behavior of asking someone who probably knows the answer. Finally, the process of producing suggestions is completely hidden from the user and therefore does not force him/her to perform additional actions.

A Web media agent was presented, which is an intelligent system to automatically collect semantic descriptions of multimedia data on behalf of users whenever and wherever they access these multimedia data from the Web and provide necessary suggestions when users want to use these multimedia data again (Zheng Chen, Liu Wenyin, Rui Yang, Mingjing Li & Hongjiang Zhang, 2001). It also shown in the experiments, the Web media agent is effective in gathering relevant semantics for media objects and is able to help users to quickly find relevant media objects.

Chapter Three

# 3 Methodology and Approaches

Web structure mining is the art of studying the anatomy of the hyperlinks that have been used to link and connect pages with each other. These links can be used to study the hierarchal structure of a Website. The carried out study of the links placed inside a Web page can be significant for identifying and analyzing the surfing complexity of various component Web pages a Website. The easiness of surfing through various pages of a Website depends upon the structural composition of a Website. The ratio of finding particular information present on a Website also depends upon the hierarchal constituents of the Web pages through hyper-links.

The conventional Web structure mining techniques are adequately efficient in reading a Web page and studying the available hyper-links and track the various available routs of surfing of that particular Website. These techniques are effective only for static Web data. Links on the static Web pages are tracked and the hierarchal structures of the target Website are drawn. The problem associated to the conducted research is that the dynamically generated Web pages, user forms and Web portal like Web information is not accessible for the purpose of mining and this hidden data is 60% of the current Web repository (Magdalini Eirinaki & Michalis Vazirgiannis, 2003b). The retrieved results and outcomes of Web structure mining become quite doubtful and unreliable after missing more than half of the available Web repository.

A certain mechanism was required to address this important issue. A framework has been designed which is sufficiently intelligent to automatically process not only the static Web pages but also the dynamic Web pages. The major issues in processing the dynamic Web pages were to handle the dynamic content generation. The major problem was to get links of the dynamically generated Web pages which are accessed in real time by filling the necessary required information by a user. These links can be retrieved by automatically filling the available fields on these Web pages and submitting the request. An automated mechanism was required that would be able to read all the input fields of a user form intelligently and fill those fields with appropriate answers. The available fields can be text boxes, check boxes, password, option buttons, command buttons, combo boxes, etc. Each identified input field is filled with appropriate answers intelligently.

The chose methodology is composed of multiple steps to accomplish the desired task. The conspired methodology bases on a newly designed algorithm that has adequate ability of reading the contents. The responsibility of the designed system is the read the user form and identifies the available input fields and filling those fields with appropriate values. After filling all the fields with suitable values, the '*doPost*' or '*doGet*' methods are invoked automatically to access the next page and ultimately the links of the next pages are retrieved and ultimately becomes the part of the hierarchal structure of the Website.

## 3.1 Algorithm

The designed algorithm uses basic guidelines of natural language processing methods. The responsibility of the designed algorithm is to read the target Web-page and point out the input tags as 'text', 'submit', 'radio', password, etc. The design and decision steps of the designed algorithm have been taken in the following sequence.

***Step – 01***

An HTML file is acquired first of all by using a string of path of the file.

***Step – 02***

The acquired HTML file is read by the designed system character by character by using various string functions.

***Step – 03***

These characters are concatenated to form words, symbols or lexicons to perform the lexical analysis.

***Step – 04***

The designed system searches for ‘<’ character to find out the HTML tags. The designed system searches for static hyper-link tags and also searches for the links of the dynamic contents.

***Step – 05***

If ‘<a’ string is found then the string ‘*href*=“’ is searched if found then the preceding string is stored as a static link.

***Step – 06***

To get the links of the dynamic Web contents, <*input*> tags and <o*ption*> tags are searched.

***Step – 07***

If ‘<input’ string is found, its type parameter is checked. If its type is ‘*text*’ or ‘*textarea*’ then a string ‘*value=“string*” ’ is added before ‘>’ character.

***Step – 08***

If the value of type is ‘*password*’ then a string ‘*value=“string*” ’ is added before ‘>’ character.

***Step – 09***

If the value of type is ‘*checkbox*’ or ‘*radio*’ then a string ‘*checked*’ is added before ‘>’ character.

***Step – 10***

If the ‘<*option*’ string is detected then a string ‘*selected*’ is added before the ‘>’ character.

***Step – 11***

If the value of type is ‘submit’ then the ‘*doPost*’ or ‘*doGet*’ methods are executed directly and link of next page is stored.

***Step – 12***

Dummy values are transferred to the server and acquired the links of all related hidden pages.

***Step – 13***

Build a tree of all target pages related to current dynamic page.

The flow of the designed algorithm is shown in the Figure 3.1.

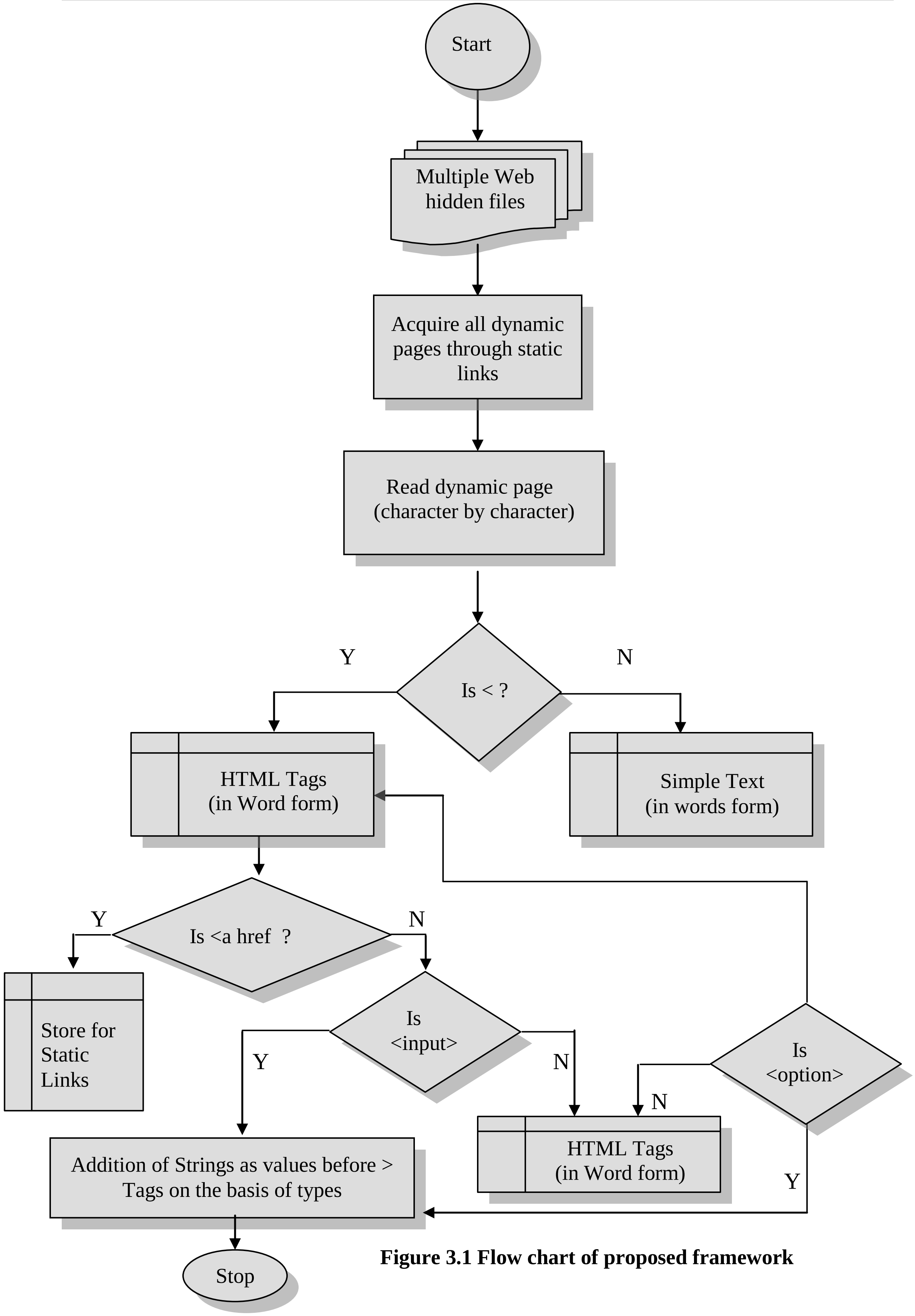

Start
Multiple Web hidden files
Acquire all dynamic pages through static links
Read dynamic page (character by character)
Y
N
Is < ?
HTML Tags (in Word form)
Simple Text (in words form)
Y
Is <a href ?
N
Store for Static Links
Is <input>
Y
N
Is <option>
N
HTML Tags (in Word form)
Y
Addition of Strings as values before > Tags on the basis of types
Stop

**Figure 3.1 Flow chart of proposed framework**

## 3.2 Functional Areas

Web structure mining is one of the three major types of the Web mining and it is used to discover useful knowledge from the structure of hyperlinks. This type of structure mining can be used to find and expose the structure and anatomy of the Web-pages, this would be good for navigation purpose and make it possible to compare or integrate the Web page schemes and improve the overall hierarchal structure of the Websites to facilitate the user for efficient and quick searching. During the design and development of the proposed system following functional areas were considered.

### 3.2.1 Collecting Data

Collecting information about the various types of the fields associated in the conventional and orthodox forms as First Name, Father Name, age, address, Phone No, Fax No, Address, Country etc. A metadata is maintained which keeps information of various types of field and what type of information is filled. Information about the various types of forms is also stored there. This information is used during the process of automatic filling of the various field of a form.

### 3.2.2 Analyzing HTML Files

Design the algorithms which can read and understand an HTML file and explain the different parts of the HTML file as finding HTML tags specifically the link tags as <*a href*= *“- - - ”* >, <*form*= *“- - - - -”* >, <*input type*= *“- - -”* > and other related tags. Design the algorithms which can automatically fill the text fields, selects check boxes, radio buttons and process the other related fields of the user form. After processing the user forms execute the dynamic pages to get the links of the target pages.

### 3.2.3 Design Hierarchal Tree

After getting the links of all the static and dynamic Web page associated in a particular Website, a hierarchal tree is designed to show the structure of that target Website. As conventional mining techniques cannot cover the links of the dynamic pages, therefore generated tree is more accurate and concise because it contains all the links both in static and dynamic form.

## 3.3 Proposed Framework

In current age, internet based Web repository is a vast source of information for the persons relating to every field of life. The modern Web consists of an ever-growing set of pages contributed by people from contrary cultures, interests, and education levels. Conventionally, Web spiders visit these Web pages and index the surfed pages to provide search engines. Information finding ratio on Web depends upon the link structure of a Website. Link information in Web pages is used to identify how many pages point to a Web page, and how many pages a Web page points to. Web crawlers can search information using these links.

The Web spiders or Web crawlers can search information only from static Web pages sue to the fact that dynamic Web pages do not provide direct links as in the form of static pages. A framework has been designed that has ability to read the links not only from static pages but also from the dynamic ones as shown in Figure 3.2. Following are the major phases of the designed framework.

### 3.3.1 Reading HTML File

This module helps to acquire the HTML file and read its contents. This module reads the input text in the form characters and generates the words by concatenating the input characters. These words or lexicons in the later phases are used to identify the tags and simple words inside a Web page. This module is the implementation of the lexical phase. Lexicons and tokens are generated in this module.

### 3.3.2 Identifying Tags

This is the second module and reads the input from first module in the form of words. These words are categorized into simple words or tags. Tags are distinguished on the basis of symbols like ‘<’ and ‘>’. These tags are further divided into simple tags or user form tags as <*form*> tags, <*input*> tags, <*list*>tags, etc. Static links are also identified in this module and are stored for further processing.

### 3.3.3 Filling Form Fields

In this module, the user forms are filled with appropriate values after consulting the user form metadata. The available fields can be text boxes, check boxes, password, option buttons, command buttons, combo boxes, etc. Each identified input field is filled with appropriate answers intelligently by the designed system automatically.

### 3.3.4 Reading Dynamic Links

After filling all the fields with suitable values, this module help to automatically invoke the '*doPost*' or '*doGet*' methods to access the next page and ultimately the links of the next pages are retrieved and ultimately they are used to draw the structural tree of the Website.

### 3.3.5 Generating Structural Trees

All the links retrieved in the previous fields as static fields and dynamic fields are used to draw the hierarchal tree of the Website to demonstrate the overall hieratical structure of the Website based on the hyper-links. On the top there is main page as shown in the Figure 3.2 and the link pages has been shown as the sub pages of the main page.

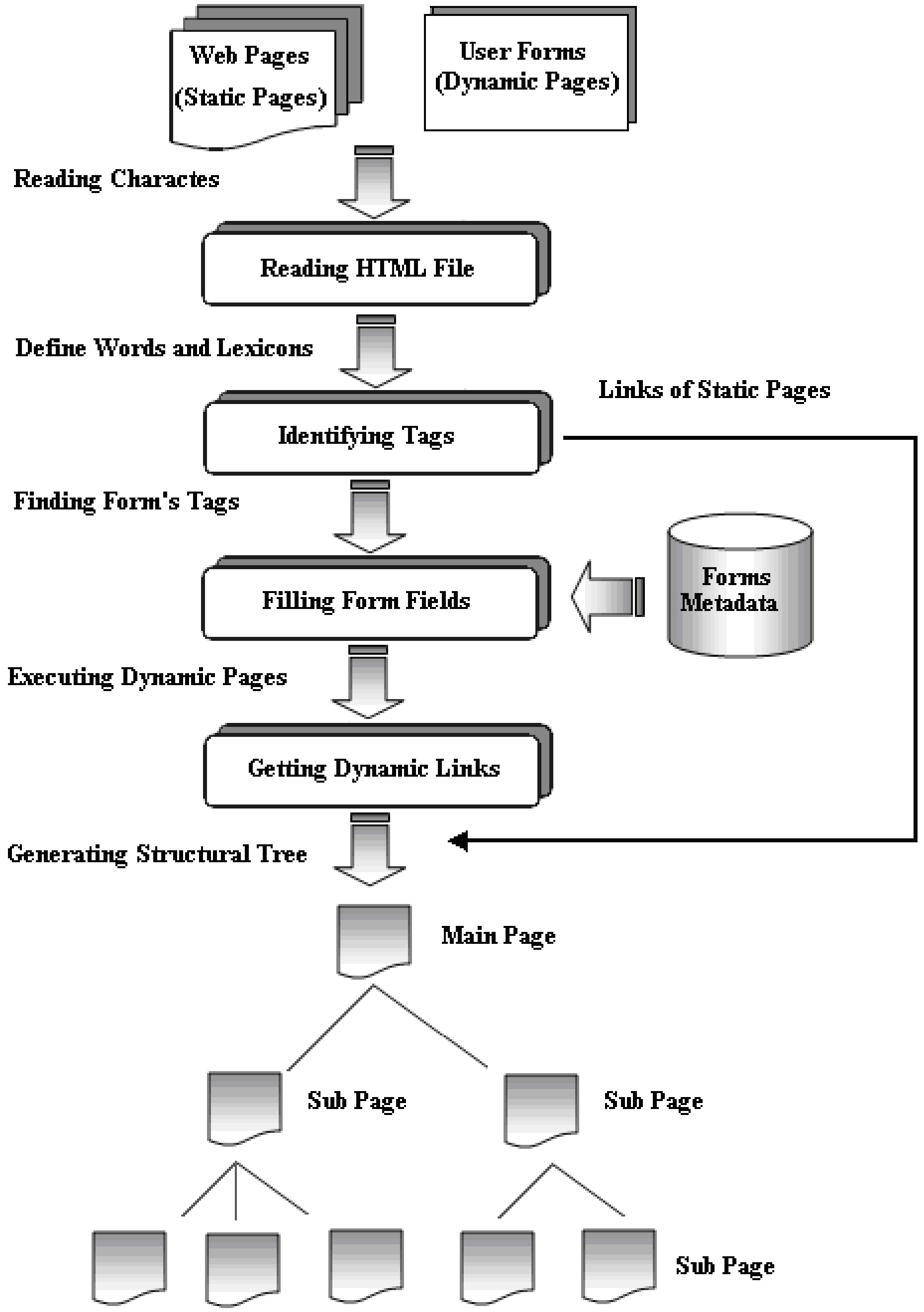


**Figure 3.2 Architecture of the Web structure mining of dynamic pages**

# Chapter Four

## 4 Design of the Experiment and Analysis

Web mining is an essential research area that helps to address problems related to search information on World Wide Web by applying techniques from data mining and machine learning to Web documents and Web pages. Hyper-links are the main source of generating relationships and linking various Web pages on the Web. To identifying these relationships and the patterns of occurrences of these relationships can be performed by reading these links and finding their hierarchy. The study and analysis of these hyperlinks helps to search and explore precious and valuable information available in the Web in hidden form.

The designed system has robust ability to read the links from these Web pages rather they are static links or dynamic links. To study, tack and analyze the static links is a quite easy job and can be performed with facing in problems and hurdles. The major problem was to find and explore the links situated on the dynamic Web pages. Links located on the dynamic Web pages are not in the form of conventional links. These links can be only be retrieved by actually executing those pages. The major issues here is that these dynamic Web pages are constituted by various forms that are the combination of various types of input fields as text boxes, check boxes, password, option buttons, command buttons, combo boxes, etc to get various types of information from the user as First Name, Father Name, Age, Address, Phone No, Fax No, e-mail, Address, Country etc. These dynamic Web pages can not be executed if these input fields are not filled with appropriate values according to their requirement.

The major issue in this problem was to fill these text fields with suitable values automatically. To fill these fields automatically, it was important to read the nature of the input field. This information is stored in a user form metadata component especially designed to fill these forms accurately. This user form metadata is automatically updated and has ability to learn. After successfully filling the fields with proper values, the Web page is executed. Now the page is executed and the link of the next page is retrieved easily. The retrieved link from the target page is stored so that it may be used to draw the hierarchal tree structure of the Website. A tree is also drawn which includes both the static links and the dynamic links form the target Website.

To perform all these experiments and provide the proof of the designed issue, a practical demo has also been designed. This practical demo is pertinently efficient to read the target HTML file, find the desired HTML tags, fill the user forms various fields, execute the dynamic Web pages, draw the hierarchal tree composition of the Website and ultimately generate an analysis report related to the compositional structure of the Website. The details and particulars of this demo are stated below.

## 4.1 Technology

To provide the concrete proof of the conducted research, Web structure mining (WSM) software has been designed to practically prove the characteristics and abilities of designed algorithm. The design and implementation details of this software system have been described later in this chapter. In the implementation of the designed algorithm, software system has been coded which is combination of various modules. Visual Basic language provides a File System Object (FSO) to read an HTML file and segregate its contents into tags and simple words.

The designed system practical demo actually basis on the various functions used for the file handling and string handling. Visual Basic provides a very strong interface to manipulate files and strings as `System.IO.File` and `FileStream` classes for file handling and so many functions and `FileSystemObject` class for String handling. It also features many utility classes for handling such things as lists, arrays, times and dates, graphics and mapped collections. By using these classes, user gain extra power over data in his programs and simplify many operations involved in using complex data structures can be performed easily. The detail of various string handling functions and classes and some functions and classes used for file manipulation are given below:

### 4.1.1 String Handling Functions

Functions available for string handling in Visual Basic are as `Left()`, `Mid()`, `Right()`, `Chr()`, `ChrW()`, `UCase()`, `LCase()`, `LTrim()`, `RTrim()`, `Trim()`, `Space()`, `String()`, `Format()`, `Hex()`, `Oct()`, `Str()`, Error. These are the dreaded variant functions. They take a variant and return a variant. These functions are quite efficient to perform various string level functions.

### 4.1.2 FileSystemObject Component

In Visual Basic 6.0, a string manipulation component is available with the name `FileSystemObject` which gives access to the file system. It typically allows creating, manipulating, deleting, and obtaining information about drives, folders, and files. To use `FileSystemObject` in Visual Basic 6.0 code, it needs to be declared in the following way:

```
Dim fso As New FileSystemObject
```

This `FileSystemObject` class can be used by using its distinctive properties and methods. The important methods used in this practical demo are as `CreateTextFile`, `GetFolder`, `CopyFile`, `MoveFile`, `GetParentFolderName`, `GetFile`, etc. All these methods are used to get a file from a specified location that is usually a folder and after opening this folder a particular file is retrieved and further processed. `FileSystemObject` allows creating ASCII or Unicode text files. When reading the HTML files, `FileSystemObject` reads in only one direction and only line by line. A file can’t be open for reading and writing functions simultaneously. In the designed system, a file is opened only in `ForReading` mode using `OpenTextFile`, but to make changes to the file later on the file is opened as a `TextStream` object.

### 4.1.3 System.IO.File Class

`System.IO.File` Class Implementations of this class are required to preserve the case of path strings. Implementations are required to be case sensitive if and only if the platform is case-sensitive. Various File Methods are available as `FileAccess` used to specify read and write access to a file, `FileShare` for specifying the level of access permitted for a file that is already in use and `FileMode` used to specify whether the contents of an existing file are preserved or overwritten, and whether requests to create an existing file cause an exception.

### 4.1.4 FileStream Class

`FileStream` class supports both synchronous and asynchronous read and write operations. Use the `FileStream` class to read from, write to, open, and close files on a file system, as well as to manipulate other file-related operating system handles including pipes, standard input, and standard output. You can specify read and write operations to be either synchronous or asynchronous.

```
Dim instance As FileStream
```

`FileStream` buffers input and output for better performance. `FileStream` objects support random access to files using the Seek method. Seek allows the read/write position to be moved to any position within the file. This is done with byte offset reference point parameters. The byte offset is relative to the seek reference point, which can be the beginning, the current position, or the end of the underlying file, as represented by the three properties of the `SeekOrigin` class.

## 4.2 Major Modules

In conventional Websites, a Web page may be linked to another Web page directly, or the Web pages are neighbors, relationships among those Web pages are significant. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient. To perform these analytical procedures, a software system has been designed in Visual Basic 6.0 language and this software system is composed of various modules.

The designed system provides an automatic way of generating hierarchal tree structure of a Website by reading the HTML files. This system performs the described function in four phases: detection, extraction of related attributes, comparison of these attributes and finally classification as shown in Figure 4.1.

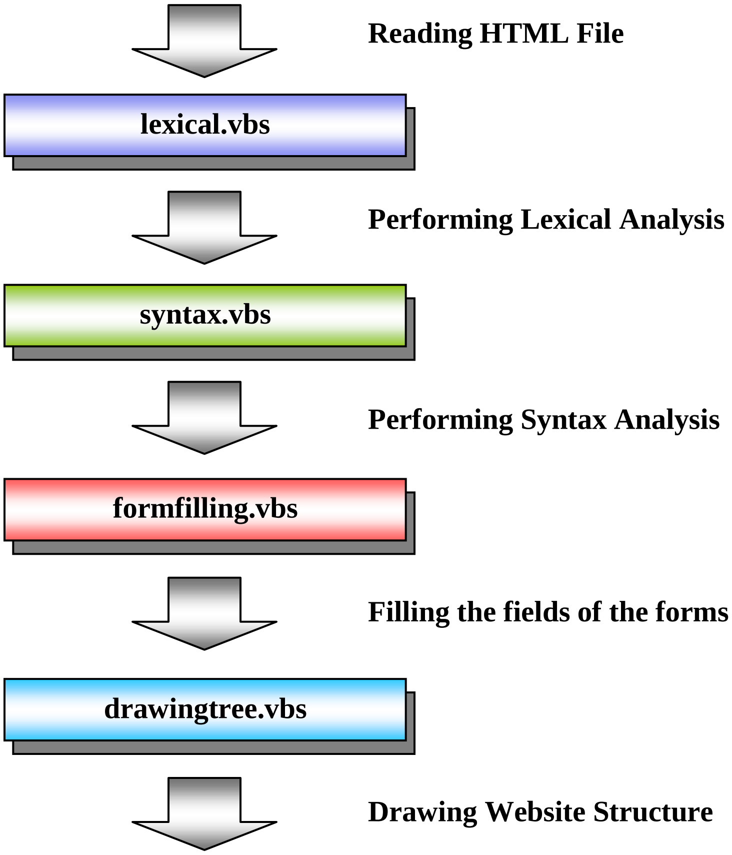

Reading HTML File
lexical.vbs
Performing Lexical Analysis
syntax.vbs
Performing Syntax Analysis
formfilling.vbs
Filling the fields of the forms
drawingtree.vbs
Drawing Website Structure

**Figure 4.1 Major implementation modules of proposed architecture**

A demo system is designed and then implemented in Visual Basic 6.0 using the newly designed algorithm based on rule based technique. Demo system consists of four classes as follows.

- lexical.vbs
- syntax.vbs
- formfilling.vbs
- drawingtree.vbs

### 4.2.1 lexical.vbs

In this class, `FileSystemObject` class and its methods has been used to read an HTML file character by character. This module is basically responsible for forming simple words and tags using combination of characters. This module basically searches for two characters, first for '<' character then searches for '>' character like <*html*> etc. This module trims the unnecessary spaces to make the process more efficient and precise, which may exist inside tags.

### 4.2.2 syntax.vbs

This class is typically responsible for searching and comparing the values inside the tags. Different string operations are used here for identifying and comparing HTML tags. Moreover different parts of tags are also identified so as to accomplish the automatic filling of text boxes and other input fields as radio buttons, check boxes, etc. This class characteristically helps to find out the form tags as <*a*> tags, <*form*> tags, <*input*> tags, <*list*>tags, etc. Static links are also identified in this module and are stored for further processing.

### 4.2.3 formfilling.vbs

This responsibility of this class is to actually fill the various fields of the user forms as text boxes, radio buttons etc. The purpose of this module is to create a new HTML file having all fields filled with dummy values. This class identifies different types of input fields on the basis of the tags. A parameter 'type' is used in the input tag, which helps to determine the subsequent type of the input field and this parameter also helps to fill the respective type of a text field. These input fields are filled with dummy values by adding string ‘ *value= “name”* ’ inside the tags. For example if an input tag is

```
<input  type= “text” name= “text1”>
```

The filled form of this tag will be as following, where a string ‘*value= “name”* ’ is concatenated.

```
<input  type= “text” name= “text1”  value= “name”>
```

### 4.2.4 drawingtree.vbs

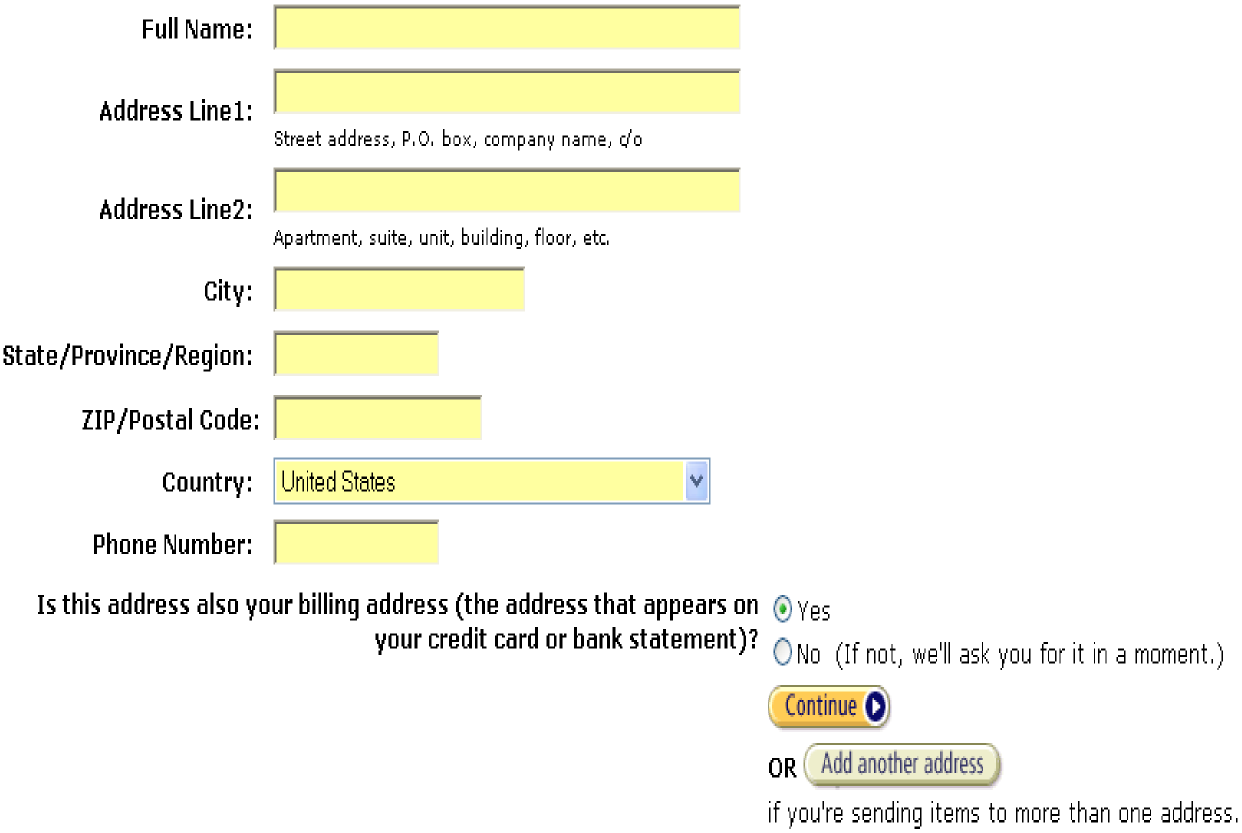
This class practically draws the hierarchal tree of the input Website to show the hierarchal structure of that Website that is ultimately used to draw a Web table. The structural information generated from the Web structure mining includes the information measuring the frequency of the local links in the Web row in a Web table; the information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document; the information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites; the information measuring the frequency of identical Web tuples that appear in the Web table or among the Web tables.

## 4.3 Selected Screens of the System

The designed classifier software system has been implemented in Visual Basic 6.0 and the power of Visual Basic’s string and file handling libraries has made the program easy to build and use. There are three major areas of the system in terms of its interface. Some major features of the user interface are that, it is very simple and easy to use. Interface of the designed system completes its task in three phases. In first phase, the designed system gets the input HTML file, reads it and in the second phase the target HTML file is analyzed and filled with suitable values and at the end the Website structural tree is drawn on the basis of retrieved static and dynamic hyper-links.

### 4.3.1 Reading HTML File Window

This is the first phase of designed interface, here the input html file is provided to the designed system for further processing. Figure 4.2 shows an input page before processing.



Full Name:

Address Line1:

Street address, P.O. box, company name, c/o

Address Line2:

Apartment, suite, unit, building, floor, etc.

City:

State/Province/Region:

ZIP/Postal Code:

Country: United States

Phone Number:

Is this address also your billing address (the address that appears on your credit card or bank statement)?

● Yes

○ No (If not, we'll ask you for it in a moment.)

Continue

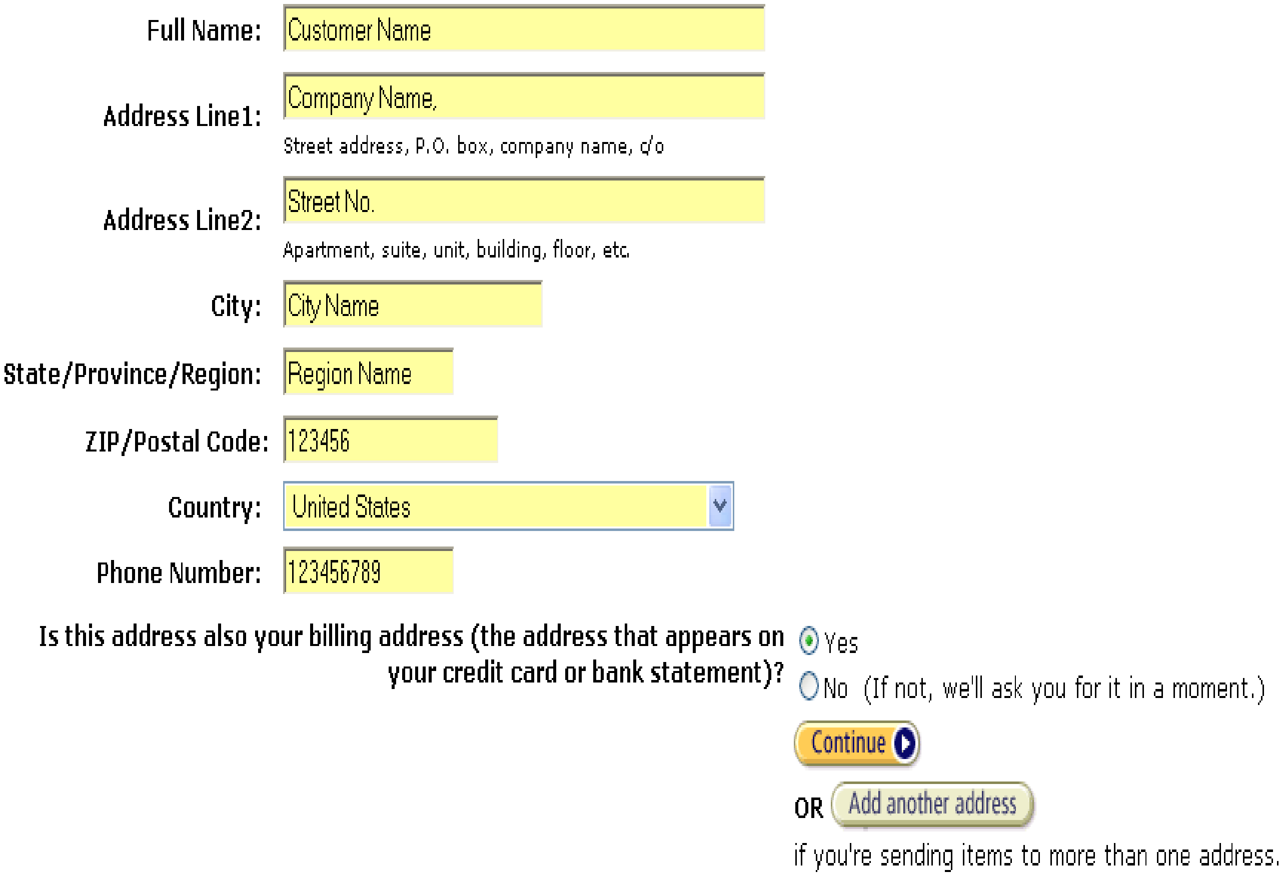
OR Add another address

if you're sending items to more than one address.

**Figure 4.2 A Web page with an empty form**

### 4.3.2 Start Training Window

This is the second phase where the various input fields of the target form are tracked and filled automatically. Figure 4.3 is an example of automatically filled form.



Full Name: Customer Name

Address Line1: Company Name,
Street address, P.O. box, company name, c/o

Address Line2: Street No.
Apartment, suite, unit, building, floor, etc.

City: City Name

State/Province/Region: Region Name

ZIP/Postal Code: 123456

Country: United States

Phone Number: 123456789

Is this address also your billing address (the address that appears on your credit card or bank statement)? ● Yes ○ No (If not, we'll ask you for it in a moment.)

**Figure 4.3 A Web page with dummy values**

### 4.3.3 Tree Generation Window

This is the third and last phase of designed interface. In this phase the ultimate outcome of the designed system is produced. Here the tree structure of the target Website has been generated by using all the links from static pages and also from dynamic pages. Following is the example of an output tree structure of a news Website. All the links from static and dynamic Web links has been drawn. The links with dynamic Web contents also have been shown with circles in Figure 4.4.

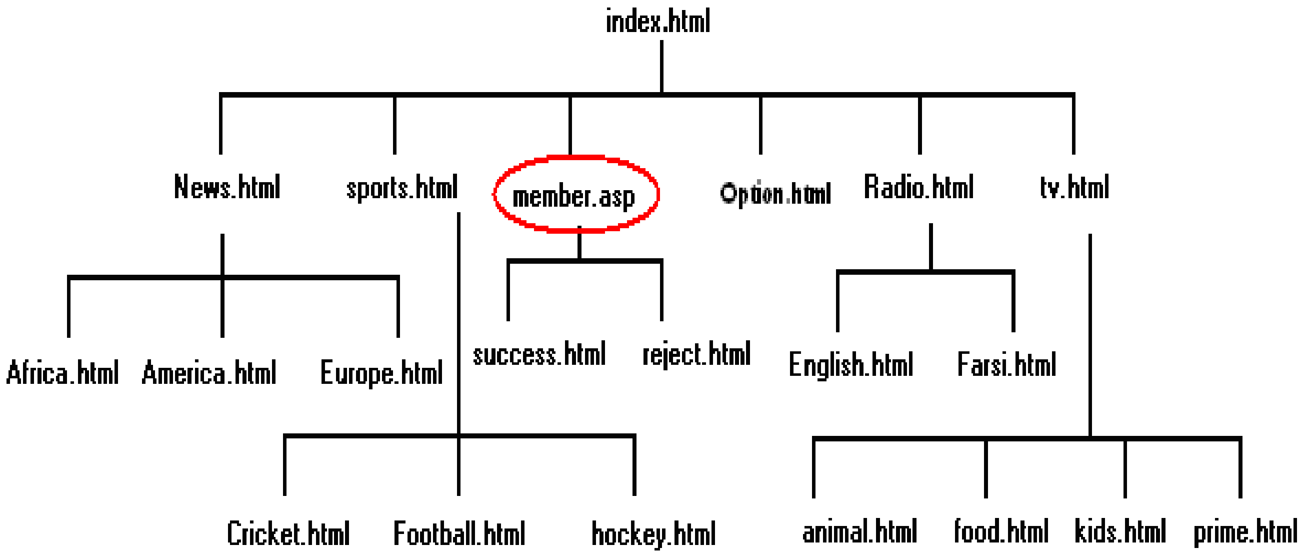


**Figure 4.4 Tree structure of a news Website**

## 4.4 Testing Data

To test the result of the designed system Web Structure Mining in hidden Web data, various Websites have been used. Three types of testing data have been used to verify the accuracy ratio of the designed system. These three types of data is as following:

- Websites with only static pages
- Websites with only Dynamic Pages
- Websites with both static and dynamic pages.

### 4.4.1 Websites with Only Static Pages

This group of data belongs to the websites that are composed of only static Web-pages. These are websites which do not have the dynamic Web content generated pages. There are so many examples; some of them are *www.apnanumltan.com*, *www.hamaralahore.com*, *www.apnaorg.com*, etc.

### 4.4.2 Websites with Only Dynamic Pages

In this group of testing data, all those Websites lie that contains majority of dynamic Web-pages. These are Websites which have most of the pages that are generated dynamically. There are so many examples as *www.yahoo.com*, *www.hotmail.com*, *www.amazon.com*, *www.ebay.com*, *www.showbiz.com*, etc

### 4.4.3 Websites with both Static and Dynamic Pages

This group of data belongs to the Websites that are composed of static as well as dynamic Web-pages. Static pages contain only text information while dynamic pages contain user input forms. There are so many examples as *www.bbc.co.uk*, *www.cnn.com*, *www.sciencemag.org*, etc.

## 4.5 Experiment Results

The designed system was checked for various types of data described in the section 4.4. The designed system showed various behaviors with different sets of data with slight variations. Comparative values results are shown in Table 4.1 for 100 different iterations. All three classes of data showed very high accuracy.

**Table 4.1: Accuracy ratio comparison among various data sets**

| **Classes** | **Static Websites** | **Dynamic Website** |
|---|---|---|
| **Static Websites** | **98** | **92** |
| **Dynamic Websites** | **92** | **87** |

Total Accuracy = 92.5 %

Figure 4.5 shows the accuracy ratios among various classes of data types. It is very obvious that the Websites with dynamic Web pages are also showing the appropriate accuracy ratio up to 87 %, where the aggregate of all three classes is 92.5%.

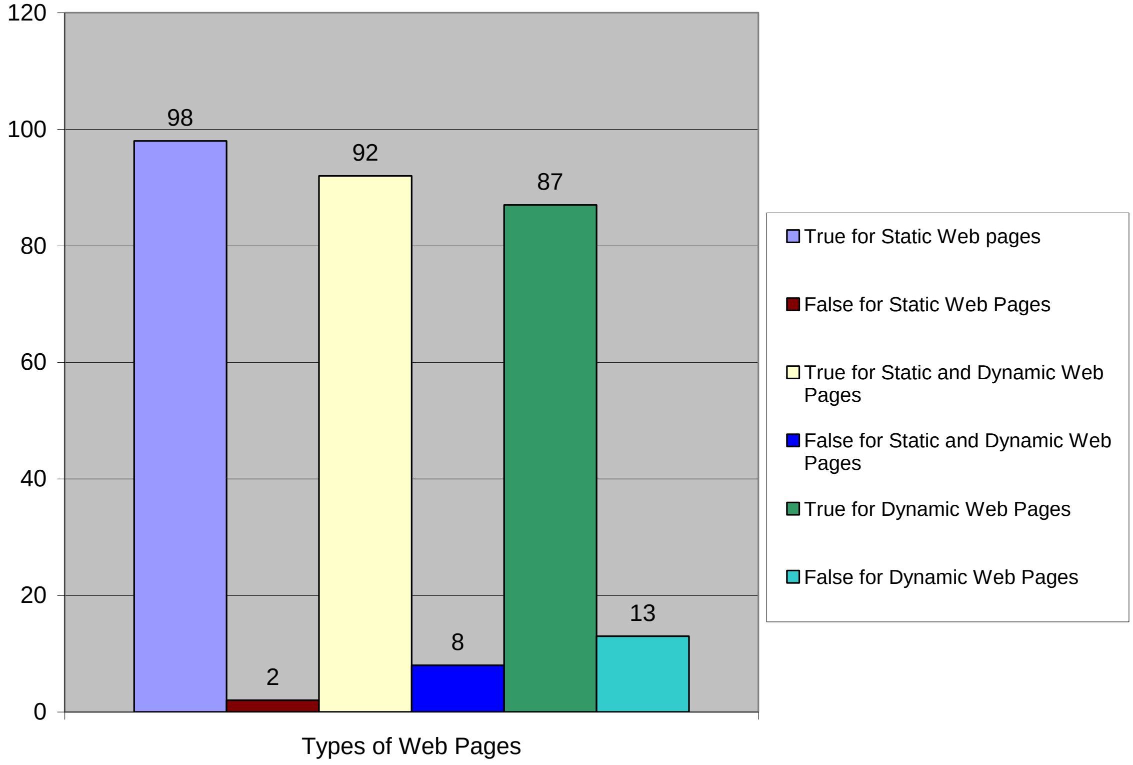


**Figure 4.5: Graph showing the comparison of accuracy ratios**

The accuracy ratio is very optimal and this ratio adequately supports the Web mining results. As the main objective of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domains as query processing, form filling, information searching, etc.

Chapter Five

# 5 Conclusions and Future Work

The major aim and objective of Web structure mining is to generate structural description about the Web pages in a Website. Principally, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

## 5.1 Conclusions

In this research thesis, a new rule based algorithm has been introduced that helps to improve the results of Web structure mining in terms of accuracy and reliability. The problem associated to the conducted research is that the dynamically generated Web pages, user forms and Web portal like Web information is not accessible for the purpose of mining and this hidden data is two third of the current Web repository. The retrieved results and outcomes of Web structure mining become quite doubtful and unreliable after missing more than half of the available Web repository.

A certain mechanism was required to address this important issue. A framework has been designed which is sufficiently intelligent to automatically process not only the static Web pages but also the dynamic Web pages. The major issues in processing the dynamic Web pages were to handle the dynamic content generation. The major problem was to get links of the dynamically generated Web pages which are accessed in real time by filling the necessary required information by a user. These links can be retrieved by automatically filling the available fields on these Web pages and submitting the request. An automated mechanism was required that would be able to read all the input fields of a user form intelligently and fill those fields with appropriate answers. The available fields can be text boxes, check boxes, password, option buttons, command buttons, combo boxes, etc. Each identified input field is filled with appropriate answers intelligently.

## 5.2 Advantages of Designed Approach

The designed approach has been quite useful to solve a problem that was a major reason of degrading the accuracy and reliability aspect of the mining results for Web structure mining. Following are some major advantages.

1. The designed approach not only considers the static pages links but also covers the links from dynamic Web pages.

2. This approach automatically fills the forms in the dynamic Web pages.

3. These forms are automatically processed and executed by the designed system.

4. Links of dynamically generated Web pages are also retrieved.

5. Tree structure of a Web page is automatically generated.

6. Meaningful information may be retrieved rapidly from designed tree structure.

## 5.3 Shortcomings of Designed Approach

There are few shortcomings of the newly designed approach besides the benefits stated in section 5.2 Following are few shortcomings in the addressed approach.

1. The intended approach only covers the dynamic Web pages like user forms but there are also other types of dynamic Web pages as Web portals, etc.

2. The designed approach only draws the hierarchal tree to show the structural format of a Website, not showing the analytical report and doesn’t mention the suggestions for betterment and enhancements.

## 5.4 Future Enhancements

According to the shortcomings stated in section 5.3 following improvements can be devised. They are not addressed in this research thesis as they are not related to the scope of this research but are quite useful to improve the results of the research. Following can be the further improvements in the currently conducted research.

As the planned approach only covers the dynamic Web pages that have user forms as there are other types of dynamic Web pages as Web portals, etc. So a vigorous mechanism should be defined that should be able to also address these parts of the problems i.e. to also process the Websites contain Web portals, and the Web contents designed in java script, VB script, DHTML, etc

In further enhancements the system should also generate an appropriate analytical report that may support in for future decisions in case of improvements the structure and anatomy of the whole Website for better access of data.

# 6 Appendix

The following is the source code for designed system which is written in Visual Basic 6.0. Given source code is successfully compiled and executed for a number of dummy tests. Designed program takes the name of dynamic page as an input, finds all blank entries like text fields, checkbox, combo box, lists and radio buttons, fills all these entries automatically with some dummy values and generates a required temporary dynamic page with filled entries.

## Source Code, Implemented in Visual Basic 6.0

```
Public mystr As String

Public mystaticstr As String

Dim les As Boolean

les = False

Dim varin As String

Dim FS As FileSystemObject

Dim TS As TextStream

Dim TSS As TextStream

Set FS = CreateObject("Scripting.FileSystemObject")

SetTS=FS.OpenTextFile("d:\source_code\inputfile.html",ForReading,False,TristateUseDefault
)

While Not TS.AtEndOfStream

    buffer = TS.Read(1)

    If Asc(buffer) = 60 Then                                    'Checking HTML Tag <

    les = True

    End If

    If les = True Then

        If Asc(buffer) = 32 Then                                'Checking Space
```

```
              If (UCase(varin) = "<INPUT") Then
              varin = varin + " Value=abc CHECKED"
              End If
             varin = varin + buffer
          Else
            varin = (varin) + (buffer)
             If UCase(varin) = "<A HREF=" Then                   'Checking static pages
                  buffer = TS.Read(1)
                  While Asc(buffer) <> 62
                        mystaticstr = mystaticstr + buffer
                         buffer = TS.Read(1)
                    Wend
              mystaticstr = mystaticstr + "<Br>"                 ' To write a static page

SetTSS=FS.OpenTextFile("d:\source_code\staticpage.htm",ForWriting,True,TristateUseDefault
)
                   TSS.Write (mystaticstr)
                   TSS.Close
            End If
               If Asc(buffer) = 62 Then                          'Checking HTML Tag >
              les = False

              mystr = mystr + varin

              varin = ""

           End If

    End If

   Else

      mystr = mystr + buffer

  End If

Wend

TS.Close

Dim fil As File                                                ' To write a dynamic page

SetTS=Fs.OpenTextFile(“d:\source_code\dynamicpage.htm",ForWriting,True,TristateUseDefault)

TS.Write (mystr)

TS.Close
```

# 7 References

A. Méndez-Torreblanca, M. Montes -y-Gómez & A. López-López, [2002], "*A Trend Discovery System for Dynamic Web Content Mining*", Puebla, Pue., 72000. México.

Adam Wilcox, M.A., George Hripcsak & M.D, [1998], "*Knowledge Discovery and Data Mining to Assist Natural Language Understanding*", Columbia University, New York, NY.

Alexander Birukov, Enrico Blanzieri & Paolo Giorgini, [2004], "*Implicit: An AgentBased Recommendation System for Web Search*", birukou@dit.unitn.it.

Alexiei Dingli, Fabio Ciravegna, David Guthrie & YorickWilks, [2003], "*Mining Web Sites Using Adaptive Information Extraction*", International Conference on European Chapter of the Association for the Computational Linguistics, Volume 2, pp 75-78, S1 4DP Sheffield, UK.

Amir H. Youssefi, David J. Duke & Mohammed J. Zaki, [2004], "*Visual Web Mining*", Proceeding of 13th International World Wide Web Conference on Alternate Track Papers & Posters, pp 394-395, Troy, NY 12180, U.S.A.

Andreas L. Symeonidis, Pericles, A. Mitkas & Dionisis D. Kechagias, [1998], "*Mining Patterns and Rules for Improving Agent Intelligence Through an Integrated Multi Agent Plateform*", asymeon@ee.ayth.gr, mitkas@eng.auth.gr, diok@iti.gr.

Armstrong R., Freitag D., Joachims T. & Mitchell T., [1995], "*Webwatcher: A learning apprentice for the World Wide Web*", Proceedings of the AAAI-95 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, pp6-12.

Ayse Yasemin Seydim, [1999], "*Intelligent Agents: A Data Mining Perspective*", Southern Methodist University, Dallas, TX 75275, USA.

Baoyao Zhou, Siu Cheung Hui & Alvis C. M. Fong, [2005], "*Web Usage Mining for Semantic Web Personalization*", Workshop on Personalization on the Semantic Web, pp 66-72.

Battina Berendt, Andreas Hotho & Gerd Stumme, [2001], "*Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space*", Workshop on Representation of and Analysis of Web Spaces, D-10178 Berlin, D-34121 Kassel.

Bettina Berendt, Ernestina Menasalvas & Myra Spiliopoulou, [2005], "*Evaluation for Web mining applications*", Berlin, Germany

Bing Liu & Kevin Chen-Chuan Chang, [2004], "Editorial: Special Issue on Web Content Mining", Chicago, IL 60607-7053.

Brin S. & Page L., [1998], "*The anatomy of a large-scale hypertextual Web search engine*", Proceedings of the 7th World Wide Web Conference.

Chen H. [2001], "*Knowledge management systems: A text mining perspective*." Tucson, AZ: University of Arizona.

Cooley R., Mobasher B. & Srivastava J., [1997], " *Web Mining: Information and Pattern Discovery on the World Wide Web*", Minneapolis, MN 55455, USA.

Cristian Aflori & Florin Leon, [2004], "*Efficient Distributed Data Mining using Intelligent Agents"*, in Proceedings of the 8th International Symposium on Automatic Control and Computer Science, Iaşi, ISBN 973-621-086-3.

*"Data Mining: What is Data Mining?*" (URL:http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm).

Ee-Peng Lim & Aixin Sun, [2005], "*Web Mining-The Ontology Approach*", aseplim@ntu.edu.sg, aixinsun@cse.unsw.edu.au.

Gerd Stumme, Andreas Hotho & Bettina Berendt, [2002], "*Usage Mining for and on the Semantic Web*", D-76128 Karlsruhe, Germany.

Gilles Falquet & Claire-Lise Mottaz Jiang, [2003], "A Framework to Specify Hypertext Interfaces for Ontology Engineering", Geneve 4, Switzerland.

Haralampos Karanikas & Babis Theodoulidis, [2003], "*Knowledge Discovery in Text and Text Mining Software*", Manchester, M60 1QD, UK.

Hicham Snoussi, Laurent Magnin & Jian-Yun Nie, [2002], "*Toward an Ontology-based Web Data Extraction*", Montreal, Canada H3A 1B9.

Jaeyoung Yang, Hyunsub Jung & Joongmin Choi, [2005], "*Building Web Navigation Agents Using Domain-Specific Ontologies*", Kt Corp., Dajeon, Korea.

Jimmy Lin & Boris Katz, [2003], "Question answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques", Twelfth International Conference on Information and Knowledge Management, New Orleans, Louisiana.

Jochen Hipp, Ulrich G untzer & Udo Grimmer, [2001], "*DATA QUALITY MINING*", DaimlerChrysler AG, Research & Technology, Ulm, Germany.

John R. Punin, Mukkai S. Krishnamoorthy & Mohammed J. Zaki, [2001], "*Web Usage Mining - Languages and Algorithms*", Troy NY 12180.

José Iria & Fabio Ciravegna, [2005], "*Relation Extraction for Mining the Semantic Web*", University of Sheffield, UK.

K.R. Reshmy & S.K. Srivatsa, [2005], "*Automatic Ontology Generation for Semantic Search System Using Data Mining Techniques*", Asian journal of Information Technology, 4 (12), pp1187-1194, Chennai, India.

K.R. Reshmy & S.K. Srivatsa, [2005], "*Automatic Ontology Generation for SemanticSearch System Using Data Mining Techniques*", Asian Journal of Information Technology, 4 (12), pp 1187-1194.

Karsten Winkler & Myra Spiliopoulou, [2001], "*Extraction of Semantic XML DTDs from Texts Using Data Mining Techniques*", Jahnallee 59, D-04109 Leipzig, Germany.

Kin Keung Lai, Lean Yu & Shouyang Wang, [2006], "*Multi-agent Web Text Mining on the Grid for Enterprise Decision Support*", Beijing 100080, China, Kowloon, Hong Kong.

Kin Keung Lai, Lean Yu & Shouyang Wang, [2006], "*Multi-agent Web Text Mining on The Grid for Enterprise Decision Support*", Chinese Academy of Sciences, Beijing 100080, China.

Kok-Leong Ong, Wee-Keong NG & EE-Peng Lim, [2003], "*A Web Mining Plateform For Enhancing Knowledge Management on the Web*", N4-B3C-14, Singapore 639798.

Kosala R. & Blokeel H., [2000], "*Web Mining Research: A survey*," SIGKDD Exploration, 2(1).

Kotaro Nakayama, Takahiro Hara & Shojiro Nishio, [2005], "*An Agent System for Ontology Sharing on WWW*", Osaka University, Japan.

Kun Chang Lee, Jin Sung Kim & Namho Chung, [2004], "A Fuzzy Cognitive Map-Driven Inference Amplification Approach to Web Mining", Seoul 110-745, Korea.

Li Zhan & Liu Zhijing, [2003], "*Web Mining based on Multi Agents*", Xidian University, Xi'an 710071, China.

Magdalini Eirinaki & Michalis Vazirgiannis, [2003], "*Web Mining for Web Personalization*", Athens University of Economics and Business.

Maria Teresa Pazienza, Armando Stellato & Michele Vindigni, [2003], "*Purchasing the Web:an Agent based E-retail System with Multilingual Knowledge*", Disp-University of Rome "Tor Vergata", Italy.

Maria Teresa Pazienza, Armando Stellato & Michele Vindigni, [2003], "*Purchasing theWeb: an Agent based E-retail System with Multilingual Knowledge*", DISP- University of Rome "Tor Vergata", Italy.

Marina Litvak, Mark Last & Slava Kisilevich, [2005], "Improving Classification of Multi Lingual Web Documents Using Domain Ontologies", Beer-Sheva 84105, Israel.

Martin Labsky, Vojtech Svatek & Ondrej Svab, [2004], "*Types and Roles of Ontologies in Web Information Extraction*", {labsky, svatek, xsvao06}@vse.cz.

Minos N., Garofalakis, Rajeev Rastogi & Seshadri S., [2001], "*Data Mining and the Web: Past, Present and Future*", minos@bell-labs.com.

Miriam Baglioni, Micro Nanni & Emiliano Giovannetti, [2004], "*Mining Literary Texts by Using Domain Ontology*", Via Buonarroti 2,56100, Pisa, Italy.

Miriam Baglioni, Mirco Nanni & Emiliano Giovannetti, [2004], "*Mining Literary Texts by Using Domain Ontologies*", Via Buonarroti 2, 56100, Pisa, Italy.

Osmar R. Zaïane, [1999], "*Introduction to Data Mining*", CMPUT690 Principles of Knowledge Discovery in Databases.

Paul Palathingal, Thomas E. Potok & Robert M.Patton, [2005], "*Agent Based Approach for Searching, Mining and Managing Enoromous Amounts of Spatial Image Data*", Oak Ridge, TN 37831-6085.

Paulo Batista & M´ario Silva J., [2002], "*Mining Web Access Logs of an On-line Newspaper*", 1749-016 Lisboa Portugal.

Peter D. Turney, [2003], "*Coherent Keyphrase Extraction via Web Mining*", 18th International Joint Conference on Artificial Intelligence, Ontario, Canada, K1A 0R6.

Peter Spyns, Robert Meersman & Mustafa Jarrar, [2002], "*Data modelling versus Ontology engineering"*,Brussel, Belgium.

Qiankun Zhao, Sourav Saha Bhowmick & Sanjay Kumar Madria, [2003], "*Research Issues forWeb Structural Delta Mining*", Singapore, 639798.

R. Guzmán-Cabrera, P. Rosso, M. Montes-y-Gómez3 & J. M. Gómez-Soriano, [2005], "*Mining the Web for Sense Discrimination Patterns*", Universidad Politécnica de Valencia, Spain.

Raymond J. Mooney & Razvan Bunescu, [2005], "*Mining Knowledge from Text Using Information Extraction*", Austin, TX 787120233.

Ricardo Baeza-Yates, [2004], "*Web Mining in Search Engines*", Blanco Encalada 2120, Santiago, Chile.

Tingshao Zhu, Russ Greiner, Gerald H¨aubl, Kevin Jewell & Bob Price, [2005], "*Using Learned Browsing Behavior Models to Recommend Relevant Web Pages*",18th International Joint Conference on Artificial Intelligence, Edmonton, Alberta T6G 2E8.

Vojt¡ech Sv´atek, Martin Labsk´y & Miroslav Vacura, [2003], "*Knowledge Modelling for Deductive Web Mining*", W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic.

W. Frawley, G. Piatetsky-Shapiro & C. Matheus, [1992], "*Knowledge Discovery in Databases: An Overview. AI Magazine*", pp. 213-228.

Y. Elovici, A.Kandel, M.Last, B.Shapira & O.Zaafrany, [2001], "*Using Data Mining Techniques for Detecting Terror Related Activities on the Web*", 84105, Israel, Tampa,Fl,33620,USA.

Yan Wang, [2000] "Web Mining and Knowledge Discovery of Usage Patterns", CS 748T Project thesis.

Ying Ding, Gobinda Chowdhury & Schubert Foo, [1999], "*Template mining for the extraction of citation from digital documents*", Nanyang Technological University, Nanyang Avenue, Singapore 639798.

Yongjian Fu & Ming-Yi Shih, [2002], "*A Framework for Personal Web Usage Mining"*, International Conference on Internet Computing, Rolla, MO 65409-0350.

Yuefeng Li a & Ning Zhong, [2004], "*Web Mining Model and Its Applications for Information Gathering*", Brisbane, OLD 4001, Australia.

Zheng Chen, Liu Wenyin, Feng Zhang, Mingjing Li & Hongjiang Zhang, [2001], "*Web Mining for Web Image Retrieval*", Microsoft Research China.

Zheng Chen, Liu Wenyin, Rui Yang, Mingjing Li & Hongjiang Zhang, [2001], "*A Web Media Agent*", Beijing 100080, P.R.China.