CHAPTER 1

INTRODUCTION

Introduction to Web Engineering

These days WWW has become a major distribution platform for a diversity of complex and sophisticated enterprise applications in many areas. Additionally it has inherent multifaceted functionality, the web applications have complex behaviour and have distinctive requirements on their performance, ability, security and usability to emerge and grow. Web applications are developed in an ad-hoc basis and face the problem of maintainability, reliability, usability and quality. However web development has certain distinguishing features that demand special consideration and it can benefit from established practices from other related area. In the last few years, there have been some developments towards addressing these problems and requirements.

1.1 Web Engineering

As an emerging discipline, web engineering is an emerging discipline actively encourages significant, systematic, quantified and disciplined approaches for the successful development of good-quality, throughout usable web-based systems and applications. Web application development has special features that make it differentiable from traditional software system and applications. Web engineering is important contributions from various domains: analysis and design, software engineering and requirements engineering. Web engineering uses software engineering discipline; it comprises new approaches, practices, tools, techniques, and guidelines to achieve the special requirements of web-based applications.

1.1.1 Scope of Web Engineering

Web engineering is the application of systematic, disciplined and quantifiable approaches to development, operation, and maintenance of web-based applications. It is both a pro-active approach and a growing collection of theoretical and empirical research in web application development.

Web applications were collections of hyper textual documents, so they weren't like traditional software, because presentation aspects were predominant. In this period (when scripting languages hadn't a great diffusion), web engineering was considered as a separate discipline, whereas today web engineering is considered as a specialization of software engineering: web engineering processes, models and techniques are adaptations of the software engineering ones. A discussion about the nature of web engineering can be found in [3].

1.1.2 Need of Web Engineering

The need for web engineering is felt according to perceptions of the developers and managers, their experiences in creating applications made feasible by the new technologies, and the complexity of web applications. In the early stages of web development [9] identified and emphasized the need for engineering as in web document engineering and web site engineering. Web engineering, more generally, explicitly recognises the fact that good web development requires multidisciplinary efforts and does not fit neatly into any of the existing disciplines.

1.2 Motivation of the Work

The maintainability is one of the critical aspects of a WA (web application): WAs have to be modified and evolve in a very fast way, then those features affecting it should be defined, identified and evaluated in order to improve/reduce the ones that have a positive/negative impact on the maintainability both during the development and maintenance process of a WA. Unfortunately, there are very few works in the literature addressing the problem of assessing the WA maintainability.

1.2 Aim of Thesis

The aim of the thesis is to propose and realize an approach for the reengineering engineering of web applications by extracting information and abstracting documentation describing the physical and conceptual structure of the application. Reengineering web applications is a complex task, due to the heterogeneity and diversity of languages, practices, methodologies, tools and technologies used together. The availability of documentation at different level of abstraction can be used will in maintenance interventions, migration and reengineering processes and it may contribute in decreasing their related risks and costs and enhancing their effectiveness

These metrics will accurately assess the efforts in the web based applications .Web page metrics is one of the basic units in measuring various particularity of web site. Metrics gives the explicit values to the attributes of web sites that can be used to assess different web pages.

An understanding of quality attributes is relevant for the web application to deliver more reliable web application. An empirical assessment of metrics to predict the quality attributes is essential in order to gain insight about the design attributes of web sites.

1.3 Main Contribution

a) Web Engineering and Maintainability

Web engineering is way of developing and organising knowledge about web application development and applying that knowledge to develop web applications, or to address new requirements or challenges. It is also a way of managing the complexity and diversity of web applications.

Web maintenance, even more than software maintenance, is a continuous activity. Depending on the nature of the application, the maintenance can become quite complex and does not solely reside in the technical domain. Content generation, and hence its update and maintenance, will necessarily vary across organisations and applications. The allocation of responsibilities for content may be carried out by human resources or other, general management units.

b) Web Reengineering

When maintenance cost is not feasible, we go for reengineering the software system. Reengineering makes the software system new. It includes both the concept of forward engineering and reverse engineering. The thesis proposed STAR paradigms of reengineering and V model for reengineering which is extension of software V model.

c) Quality Prediction of Web Application

Quality is a critical field in web engineering and important facet of web quality. Every organisation wants to determine the quality of web product at early stage so that development cost and time can be saved. This would lead to early detection of poor website design responsible for the poor quality product.

Hence, the main contributions of this section are:

(1) In this section relationship between web metrics and status of website (good or bad) has been established. There are a number of website design metrics such as total number of links, web page size, total number of embedded text and complexity of a web page etc. but all the metrics cannot predict the quality of a web site. Thus, it is very important to understand the relationship of web metrics and status. In other words, we must find out which of the metrics are significant in predicting the status (good and bad) of a website. Then, these significant metrics can be combined into one set to build the multivariate prediction models for predicting quality. Identified metrics will help web developer to predict goodness of a website.

(2) The research analyzes machine learning methods .Nowadays, machine learning is widely used in various domains .There are various machine learning methods available. This work has used seven machine learning methods to predict the accuracy of the model predicted. These seven machine learning methods have been widely used in literature and have shown good results. Amongst the various models predicted, one must determine one of the models to be the best model, which can be used by researchers in further studies to predict the goodness of a website.

1.4 Thesis Outline

The thesis is divided in six chapters.

Chapter 1 is the introduction part. It describes web engineering, need of web engineering motivation of work, aim of thesis, main contribution and the structure of thesis. Chapter 2 presents the related work.

In the chapter 3 a general background is provided. This chapter reports a discussion about the nature of web application. This chapter gives the brief introduction of maintainability and quality of web application and its importance.

Chapter 4 presents the web reengineering and its various approaches. It reports the description of proposed V model of web reengineering and STAR paradigm of reengineering for web applications. It also presents the comparative analysis of reverse engineering and reengineering.

In chapter 5 quality prediction of web sites is presented. In this chapter, relationship between web metrics and quality of web page is established using statistical learning methods. This chapter used machine learning methods for the model prediction.

Chapter 6 provides result analysis. It includes result of univariate logistic regression and model evaluation using the ROC curve.

Finally thesis work is concluded along with future remarks.

CHAPTER 2

LITERATURE REVIEW

LITERATURE REVIEW

A web application is a software system that exploits the WWW infrastructure to provide its users the opportunity to alter the status of the system and of the enterprise it supports [1]. A different kind of hierarchies has been proposed to classify web application. Tilley and Huang [2] proposed an interesting taxonomy for web applications. According to which there are three classes of web applications with growing complexity can be differentiated.

Web-based applications and system provide complex functionalities to different groups of users. Over the year, our dependence on the web increased drastically, hence their reliability, maintainability; performance and quality have become crucial. As a result, the development of web applications has become more complex and challenging in this era. It is different from traditional software development. However development and maintenance of many web applications exhibit chaos and is undisciplined. Web developers need to adopt a disciplined development process and a sound methodology to build and maintain complex web based application and system successfully. The growing discipline of web engineering encourages disciplined approach to develop web applications [3]. The maintainability is one of the critical aspects of a WA (web application): WAs have to be modified and evolve in a very fast way, then those features affecting it should be defined, identified and evaluated in order to improve/reduce the ones that have a positive/negative impact on the maintainability both during the development and maintenance process of a WA. It has been measured that in the maintenance phase software professionals spend at least half of their time analyzing software to understand it [4]. Web engineering processes, models and techniques are adaptations of the Software Engineering ones. A discussion about the nature of web engineering can be found in [3]. In the early stages of web development [9] identified and emphasized the need for engineering as in web document engineering and web site engineering.

The software maintenance as defined in IEEE standards [5] is: The reconstruction and updating of a software system after delivery to correct bugs, to enhance performance or other attributes or organise the product to a modified environment. According to Basili and Mills [6] the software maintenance may be looked as: Most software systems are complex, and modification requires a deep understanding of the functional and non-functional requirements, the mapping of functions to system components and the interaction of components. Chikofsky and Cross define reengineering as 'the examination and alteration of a software system to reform and reorganise it in a new form and subsequent implementation of that form' [7].According to IEEE Std. 1998 'A system changing activity that results in creating a new system that either retains or does not retain the individuality of the initial system' [8]. Techniques of static and dynamic analysis of the source code and dynamic data were taken into account.

In research papers [35][36][37] transaction based model has been described that the Reengineering transfigure a final user interface into a logical representation that is allow forward engineering to port a UI from one computing platform to another with maximum flexibility and minimal effort. Reengineering is used to modify a UI to another format. The re-organisation is accomplished by the adapting the code into another computing platform and the redesigning of the user interface into target platform incorporating the concerned constrained in a better way.

Data reengineering process is described in [22] according to which data reengineering achieves a very simple objective that enhances the value of the data which values to your business. In some ways it is analogous to the transformation that often occurs when you extract data from your operational databases into a data warehouse.

UIML [46] allows the developer to specify the presentation and dialog components of a UI in a device independent language, which can then be used to produce several UI for different computing platforms. Web Revenge [45] is a tool that analyzes web site code in order to automatically reconstruct the underlying logical interaction design. Such a design is represented through task models that describe how activities should be performed to reach users goals. Teresa [43] follows a task-based approach for the generation of UI for multiple devices. The process of generation is decomposed into 4 parts: 1) the creation of a unique task model. For each task, the designer specifies the interaction objects needed to accomplish the task, and the platforms on which the task is available. 2) A separate task model is then automatically generated for each platform. 3) Based on this task model, a logical UI is then created for each platform, by identifying the enabled task sets (tasks that should appear on the same screen) and the interactors needed to fulfil the tasks. 4) The final UI code is generated. The user has the freedom to modify the heuristics and models at each step of the process to fine-tune the UI.

Reweb.s reengineering process [39] restructures web applications in order to avoid their inevitable degradation. It uses a set of transformation rules aiming at the improvement of maintainability, usability and portability. It also restructures the design thanks to a web application model and by incorporating frame-based navigation. Similarly some transcoding tools [19], [20], [21], [22] automatically transform a UI code from the original platform to a target platform.

Izzat Alsmadi et. al described that the evaluation of the characteristics of websites can take several methods. Some of these methods depend upon the users and others are on website itself. Authors have described that there are many tools and websites available that measures attributes of web sites such as vulnerability, navigability, performance and structure etc. Their work focused on website structural and related metrics that can be used as indicator of a websites. Websites structural metrics can be used to predict maintainability requirement [86]. Melody Y. Ivory et. al provided empirical evidence that important metrics, including page composition, page formatting, and overall page characteristics, differ among web site categories such as education, community, living, and finance. These results provide an empirical foundation for web site design guidelines and also suggest which metrics can be most important for evaluation via user studies [79].

Webby awards website [65] in every category there are two honours presented in the Webby awards- the webby awards and the webby people's voice award that exhibits five category of entry type i.e. Interactive Advertising & Media, Online Film & Video, Mobile & Apps and Social websites. For each category, the nominee and the winner of the both awards is selected by the member of international academy of digital arts and sciences. However webby people's voice winner award is selected by voting process for best nominee in each category. From all over the world, The Webby People's Voice Awards are garnered by millions of vote [65].

Full Bayesian Network Classifiers [74]: In their paper, Jiang Su et. al have described the concept of full Bayesian Network classifier. This includes the use of variable independence in learning condition probability tables instead of in learning structure. The main advantage is that learning decision trees for CPTs capture essentially both variables independence and context-specific independence. Moreover the reduced efforts help in minimization of time-complexity. Multivariate Logistic Regression Prediction of Fault-Proneness in Software Modules[83] the authors have applied three different techniques of logistic regression, i.e. forward stepwise logistic regression, backward stepwise logistic regression and one without stepwise logistic regression over nasa promise data for fault prediction. Later on they have concluded that the backward stepwise logistic regression gives the best result.

Hybrid Version of MLP Neural Network for Transformer Fault Diagnosis System [71][72] [73] The authors have used a hybrid version of standard multilayer perceptron, aka Hybrid Multilayer Perceptron (HMLP). Further they have used the Modified Recursive Prediction Error (MRPE) algorithm to train their neural network. They used three different algorithms to analyse the performance where they found out HMLP to be the best one. Implementation of Breiman’s Random Forest Machine Learning Algorithm [70] the author has demonstrated the use of Breiman’s Random Forest Machine Learning Algorithm to improve the classification of diverse data and minimize the chances of misclassification that occur in other classification tools. They have used Weka tool in their experimentation [66].

Ensemble methods in Machine Learning [9] the paper aims to promote ensemble of Machine Learning methods. Different ensemble methods (i.e. using different machine learning and then taking their average or weighted vote of their prediction) have been used such as Bayesian averaging, bagging and boosting. The paper finally recommends that ensemble sometimes prove to be better than any single classifier. A Short Introduction to Boosting [67] this paper illustrates the concept of Adaboost algorithm that improves the accuracy of the classifiers using some examples. The power of Decision Tables, this paper [75] illustrates the concept of tree algorithm.

CHAPTER 3

WEB APPLICATION: MAINTAINABILTY AND QUALITY

Web Application: Maintainability and Quality

In this chapter some basic definitions are reported. In particular, a definition of web application by distinguishing between web applications and web sites is provided. Moreover, the main architectures and technologies typically used for implementing web application are described as well as short description of maintainability and quality of web application.

3.1 Web Application

A web application is a software product designed to be executed in the World Wide Web environment. A web application can be considered as an extension of a web Site. A web Site is a collection of hyper textual documents, located on a web server and accessible by an Internet user. Unlike a web site that simply provides its users the opportunity to read information through the World Wide Web (WWW) window, a web application can be considered as a software system that exploits the WWW infrastructure to offer its users the opportunity to modify the status of the system and of the business it supports [1].

3.1.1 Classification

A large number of taxonomies have been proposed to classify web application. Tilley and

Huang [2] proposed an interesting taxonomy for web applications. According to this taxonomy, three classes of web applications with increasing complexity can be distinguished.
Class 1 applications are primarily static applications implemented in HTML, and with no user interactivity.
Class 2 applications provide client-side interaction with Dynamic HTML (DHTML) pages, by associating script actions with user-generated events (such as mouse clicking or keystrokes).

Finally, class 3 applications contain dynamic content, and their pages may be created on the fly, depending on the user interaction with the application. A class 3 applications is characterised by a large number of employed technologies, such as Java Server Pages (JSP), Java Servlets, PHP, CGI, XML, ODBC, JDBC, or proprietary technologies such as Microsoft's Active Server Pages (ASP).

Web applications have some peculiarities that influence their life cycle and differentiate them from traditional application. An indicative list of these peculiarities could be the following:

- The main purpose of a web application usually consists in data storing and browsing;
- Web applications are always interactive applications: usability is a fundamental quality factor for them;
- Web applications are always concurrent applications and the number of contemporary users may vary in unpredictable way: scalability is another fundamental quality factor;
- Web application developers are usually low-skilled people, subject to a frequent turnover;

- Many technologies doesn’t encourage separation between logic layers: often peoples with different skills must work together (i.e. programmers and graphic artists);
- Web applications need a continue evolution, for technological and marketing reasons, too;
- Web applications must be developed in a very short time, due to the pressing short time-to market.

These factors give an idea of the problems related to web application developing. Life cycles commonly adopted for web applications are incremental ones, and all phases follow an iterative developing.

3.2 Maintainability of Web Based Systems

The maintainability is one of the critical aspects of a WA (web application): WAs have to be modified and evolve in a very fast way, then those features affecting it should be defined, identified and evaluated in order to improve/reduce the ones that have a positive/negative impact on the maintainability both during the development and maintenance process of a WA. Unfortunately, there are very few works in the literature addressing the problem of assessing the WA 0aintainability.

Definition: - The ease with which repair may be made to the software as indicated by the following sub attributes: analyzability, changeability, stability and testability (ISO 9126).

3.2.1 Importance of Maintainability

It has been measured that in the maintenance phase software professionals spend at least half of their time analyzing software to understand it [4]. The cost of software maintenance accounts for a large portion of the overall cost of a software system. Thus malfunctions of a critical software system can cause serious damages. For example, a problem in the Amazon.com web site in 1998 put the site down for several hours which cost the company an estimated $400,000. Also the relationship between the company and its customers can be greatly affected by such down time. Quantitative metrics and models for predicting web applications' maintainability must be used to control the maintenance cost.

After a web-based system is developed and deployed online for use, it needs to be maintained. As outlined earlier, content maintenance is a continual process. We need to formulate content maintenance policies and procedures, based on the decision taken at the system architecture design stage on how the information content would be maintained, and then we need to implement them. Further, as the requirements of web systems grow and evolve, the system needs to be updated and also may be redesigned to cater to the new requirements.

It is important to periodically review web-based systems and applications regarding the concurrency of information content, potential security risks, performance of the system, and usage patterns (by analysing web logs), and take suitable measures to fix the shortcomings and weaknesses, if any. It is a well established fact that the web applications require frequent maintenance because of cutting– edge business competitions.

The software maintenance as defined in IEEE standards [5] is: The modification of a software product after delivery to correct faults, to improve performance or other attributes or to adapt the product to a modified environment. According to Basili and Mills [6] the software maintenance may be looked as: Most software systems are complex, and modification requires a deep understanding of the functional and non-functional requirements, the mapping of functions to system components and the interaction of components.

Maintainability is an important attribute in all the software applications as it is learned that only 25% to 33% of the total effort put in during the complete life cycle of a software system goes in actually building the system [6]. The rest is consumed by effort expended towards the operational maintenance of this system. This figure clearly indicates that maintenance takes more efforts as compared to the development of the software. The maintainability of software system has always been a problem with software professionals. Since the third-party maintenance is now becoming a reality as more and more organizations are opting for third-party maintenance of their web applications. It is the high time that software maintenance be looked in the right perspective so that a realistic cost estimates are prepared for the software maintenance. Most web applications involve critical business assets which promote their services through internet. Because of globalization and cut-throat business competition, these web applications evolve continuously during their life-cycle. Lehman et.al. [5] gave two laws of software evolution that affect the evolution of web applications. They are

1. The law of continuing change: A program used in real world must change or eventually it will become less useful in the changing world.
2. The law of increasing complexity: As a program evolves it becomes more complex and extra resources are needed to preserve and simplify its structure.

Web applications are different from traditional software systems in the sense that they involve heterogeneous technologies in hardware as well as software. For successful development of large web applications, we need a team of people with wide ranging knowledge and skills. We need graphic designers to develop the look and feel, we need people with library science background to organize, navigate and search information. We need database designers and programmers to develop code, network security and other security aspects. We often involve architects to get better aesthetics in the web applications. The code development will involve hypertext structures, JSP, Servlet, scripting languages, etc. It is a common practice that web applications are hosted and maintained by third party. Because of heterogeneity of such web applications, the maintenance becomes a cumbersome process and becomes impossible to predict maintenance cost using traditional models and metrics.

When we talk about the web application maintenance the structure of a web application should be considered. Web applications are different from traditional software systems in the sense that they involve heterogeneous technologies in hardware as well as software. These are built up of different items coded with different programming languages. Any web application is an arrangement of web pages. These pages can be static as well as dynamic. Dynamic pages are generated at run-time. The static pages are normally written in HTML.

The dynamic pages are normally written using the scripting languages, and the back end is typically written using the database management languages.

3.3 Quality Evaluation of Web Application

Quality Assurance in the web is a topic of increasing interest and concern, especially if we take into consideration the expansion that this media has experienced in the last few years and the negative economic impact that rolling out a low quality web system can cause, very specially when we refer to the e-commerce area. Web based applications (WAs) are the multi tier, distributed applications accessed through client side browsers in the heterogeneous environment by unlimited number of users with varied experience. While the complexity of WAs increases each day, the resources and testing times decrease [24]. As was become complex, there is a growing concern about their quality. Evaluating website quality is essential, but there are few ways to analyze and evaluate the quality of the website in quantitative form.

3.3.1 Important Aspects of Quality of Web Application

The websites are becoming a competitive tool for business applications. The websites exposes products and services to its users and can create more sales by communicating the characteristics of the products and services. Credibility of a website is directly proportional to its quality. The need is to create a method which can guide the internet users to evaluate a website in the least possible time. Quality test is based on frequency of update, load time, response time, page rank, traffic, design, size, number of items, accessibility error, validation, and broken link etc. Good design should make a site easy to use and should be able to accomplish the site’s objectives and goals. In order to create a new website quality evaluation method effectively, some limitation has to be considered according to existing website evaluation methods.

- It is necessary to create a comprehensive website evaluation method that is applicable to all the websites
- A new website evaluation method needs to involve the all identified new software technologies as the numbers of new criteria.

- Metrics for website quality are characterized by uncertainty, subjectivity, imprecision and vagueness with perception of response. When consumers make decisions, a more realistic approach may use linguistic assessments to express thinking and subjective perception.

CHAPTER 4

Web Maintainability and Web Reengineering

Web Maintainability and Web Reengineering

4.1 Introduction

Most web applications are developed under proper schedules and in a rapidly evolving environment. The development is often ad-hoc in nature and the applications are poorly structured and poorly documented. Maintenance of such applications becomes problematic and increases the complexity of the web application grows. Creating appropriate design and architecture models is the solution to managing this complexity and supporting evolution of web applications. Researchers have identified the need to reengineer the system already existing web applications into abstract design models. The diverse and dynamic nature of elements and techniques used to develop web application, due to the lack of testing technique and effective programming principles which are used for implementing basic software engineering principles.

4.2 Relationship between Web Maintainability and Web Reengineering

Maintenance and reengineering terms are closely coupled with each other. These terms came from the world of hardware objects. It is difficult to draw a clear cut line between these two terms. Many a times these are used interchangeably. Reengineering of software systems is a topic of importance and in coming time it will be gaining more attention in the world of software systems. Software managers are often confused over maintenance and reengineering.

Maintenance is one of the stages in the software development life cycle. It starts after the deployment of software in the working field. It is to remove the defects and deficiencies which encounters while starts actually working in the field.

According to IEEE Std. 'Software maintenance is the process of modifying a software system or component after delivery to correct faults, improve performances or other attributes, or adapt to a changed environment'.

Reengineering is the analysis of existing software system and modifying it to constitute into a new form. Chikofsky and Cross define reengineering as 'the examination and alteration of a subject system to reconstitute it in a new form and subsequent implementation of that form' [7].

The technological evolution of the last year has made the web service of the ideal platform for the appropriate support for their delivery and the development of web based applications. According to research [1] and [2] the development of a web application is a multi-faceted activity, involving not only technical but also organizational, managerial and even social and artistic issues. Web application development refers a set of activities which applied in order to develop a web application of high quality having awaited characteristics, and to accomplish this development efficiently and coherently. Web engineering is an important topic in these days and is gaining more attention. It is fast developing area and not existing from centuries. Web maintenance and web reengineering both falls in the scope of web engineering. The World Wide Web has ability to ubiquitously provide and gather information to the economy globalization together with the need of new marketing strategies has enormously boosted the development of web applications (WA). Software application is the backbone of the WWW infrastructure.

Most web applications are developed under proper schedules and in a rapidly evolving environment. The development is often ad-hoc in nature and the applications are poorly structured and poorly documented. Maintenance of such applications becomes problematic and increases the complexity of the web application grows. Creating appropriate design and architecture models is the solution to managing this complexity and supporting evolution of web applications. Researchers have identified the need to reengineer the system already existing web applications into abstract design models.

Reengineering

Reengineering is the analysis of existing software system and modifying it to constitute into a new form. Chikofsky and Cross define reengineering as 'the examination and alteration of a subject system to reconstitute it in a new form and subsequent implementation of that form' [7].According to IEEE Std. 1998 'A system changing activity that results in creating a new system that either retains or does not retain the individuality of the initial system' [8].

4.1.1 Nature and Scope of Reengineering

When maintenance cost is not feasible, we go for reengineering the software system. Reengineering makes the software system new. Reengineering has the following three stages.

1. Reverse engineering
2. Transformations or Transfiguration
3. Forward engineering

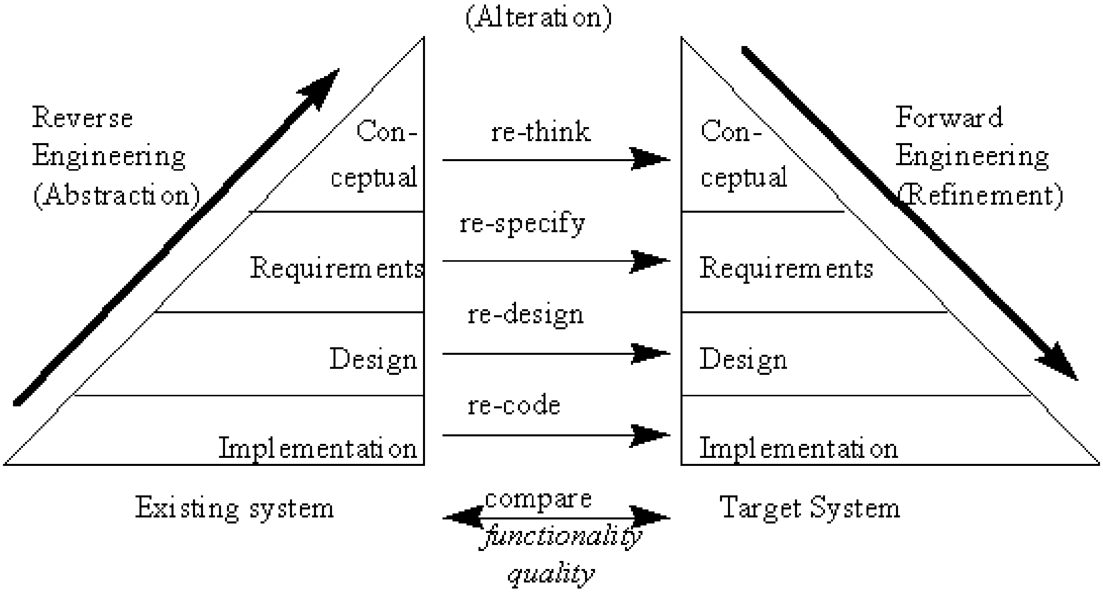


Figure 1 General Model for Software Reengineering [85].

Reverse Engineering

Reverse engineering is the process of analysing the system which helps in recovering its design and specification. Reverse engineering is a process of analysis to determine the relationship of the system component and create the components of the system in another form or in a higher level of abstraction. Reverse engineering is used to produce a better system and it is a part of the reengineering process. However propose of a reverse engineering process for web encompassing the following phases:

1. Static Analysis

2. Dynamic Analysis

Transformations and Transfiguration

This phase involves the transition, alteration, modification, reformation, reconstruction, remodelling of the web application system. Web architecture is altered. It is modified, improved to cope with the new technology and new environment. It is the architecture designing stage. The reorganisation is accomplished by the adapting the code into another computing platform and the redesigning of the user interface into target platform incorporating the concerned constrained in a better way.

4.1.2 Approaches for Web Reengineering

With the advancement of technology continuous changes are introducing in web industry and hence web application need to be cope with the latest technologies and competing in market. The need to fulfil the market and additional requirement may lead to need for web reengineering where the web application system is transformed from one state to another state.

Reengineering of Web Pages

Reengineering of web pages can be accomplished by detecting and analysing the interaction of objects and then transforming these objects for the adaption of new platform itself and generating the source code into new language. There are several presentation models that can be used to transform into another model for different context supporting flexible reverse engineering process. The detection and transformation phases of the reengineering process can be governed as

- *Percolating the objects*, tags and elements of the web pages that include selection of any HTML item, with given properties that require to keep all control mechanism and discarding the unwanted tags and elements from web pages.
- *Transformation in the layout options* and relationships that include alignment, balance (horizontal or vertical balancing), centric which depend upon the position of the objects on the page.
- *Content updating* of the web pages according to the requirement changes, market evolution, usage and owner of website.
- *Clustering of web pages* allow for the information obtained by analysis (static and dynamic) of reengineering process can be used to produce a graph whose nodes represent the set of web application objects, and whose links specify the interaction between these objects. In [9], this kind of graph is called WAG; *Web Application connection Graph.* WAG' analysis may support the comprehension of the application. However, since this graph may be large (in terms of the number of nodes and edges) even in the case of small size web applications, in order to simplify the analysis of large WAG graphs, some kind of automatic clustering [10] can be used to decompose this graph into smaller cohesive parts. In the third step of the reverse engineering process, the automatic clustering approach proposed in [9] is applied, in order to group software items of a WAG into meaningful (i.e. highly cohesive) and independent (i.e. loosely coupled) clusters. This clustering approach evaluates the degree of coupling between entities of the application (such as server pages, client pages and client modules) that are interconnected by *Submit, Build, Link, Load in Frame*, *Redirect,* and *Include* relationships.

  Among several cluster obtain by clustering algorithm we choose the most optimal cluster by evaluating the degree of i*ntra-connectivity* and degree of *inter-*

connectivity(minimizes intra-connectivity and maximizes the intra-connectivity)The 'optimal' configuration is considered the most suitable for including clusters implementing functions at higher levels of abstraction than that of the cluster's single items. Validation of the clusters, based on a Concept Assignment Process [11] has to be carried out.

- Grouping of objects that are close to each other because they are semantically related this process is called *association* and ungroup objects that are isolated without any connection when they are unrelated which is called *dissociation*.

Transaction Reengineering

In a transaction oriented web site, the user executes a series of activities in order to carry out a specific task. One of the reasons for the success of e-commerce business today is the transactional behaviour that the web offers. Business processes are realized by means of transactions, which in this context can be interpreted as high-level workflows corresponding to user tasks (e.g. purchasing an airplane ticket). The process is a revised version of the UWA Transaction Design Model [12], which is the portion of the UWA framework that focuses specifically on the design of web application transactions.

The UWA design framework provides a complete design methodology for ubiquitous web applications that are multi-channel, multi-user, and context-aware. The UWA design framework organizes the process of designing a web application into four main activities [13]. (1) Requirements elicitation (2) Hypermedia and operation design (3) Transaction design and (4) customization design [14] [15] .Using the UWA methodology, the transaction design process produces two conceptual models: the Organization Model and the Execution Model. The organization model describes a transaction from a static point of view. It uses a particular UML class diagram [17] in which the activities involved in the transaction are represented by class stereotypes, which are arranged to form a tree. The activity represented by the root of the tree corresponds to the entire transaction; component activities and sub activities are intermediate nodes and leaves of the tree that represent sub transactions and elementary activities, respectively. The Execution Model of a transaction defines the possible execution flow among its component activities and sub activities. It is a customized version of the UML Activity Diagram [18]. The sequence of activities is described by UML Finite State Machines, in which activities and sub activities are represented by states (ovals), and execution flow between them is represented by state transition (arcs).

Application Migration Reengineering

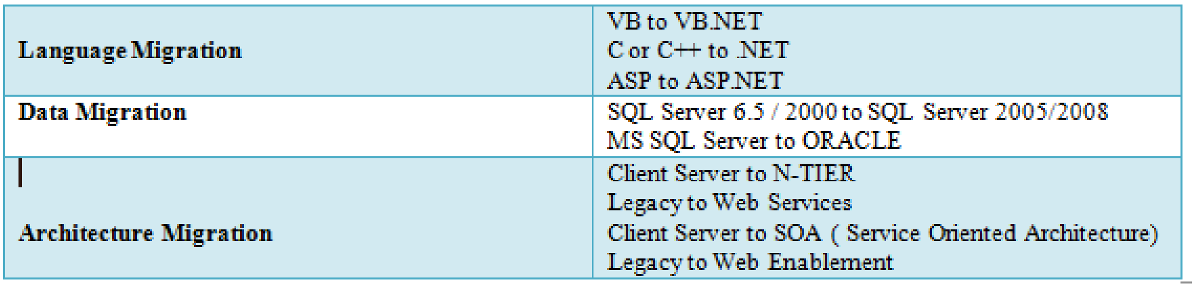
Migrating applications to the newer technologies can give business a leading edge by removing inefficient workflow and processes while preserving original objectives, model and investment. We can help enterprises in migration of the legacy systems from old technologies to present day platforms. Reengineering must keep into consideration the strategically designed to overcome the cross platform compatibility challenges.

Due to upcoming advance technology and growing business states, there is need for the migration of legacy software systems to new technologies and environments. There are different kind of legacy system reengineering services that includes language and database migration, platform-to-platform porting and system redevelopment.

A web application must follows the enterprises standard and rules implemented in a legacy application, while transforming those to new business and architecture requirements, to produce a flexible, tested or validated modified system. Reengineering and Migration Benefits are the saving time and effort, Enhancements in operational efficiency, Benefits of the latest technologies and platforms

Web application migration can include following services:

Legacy application and reusable component analysis
New technology and platform inspection
Platform, language, database and architecture migration
Design, development and integration
Version rendering
Functionality enhancement
Application and process organising



Language Migration	VB to VB.NET C or C++ to .NET ASP to ASP.NET
Data Migration	SQL Server 6.5 / 2000 to SQL Server 2005/2008 MS SQL Server to ORACLE
Architecture Migration	Client Server to N-TIER Legacy to Web Services Client Server to SOA (Service Oriented Architecture) Legacy to Web Enablement

Table 1 Application Migration for Web Application

Graphic Design Reengineering

Reengineering is used to modify a user interface into new context. To change user interface it is not mandatory to start developing it from scratch. Some transcoding tools [19][20][21][22] automatically transform a UI code from the original platform to a target platform. Portability and transcoding exhibits some limitation as they do not need to consider constraints imposed by the target platform such as: operating system, programming language, screen resolution, interaction capabilities To overcome these shortcomings UI reverse engineering process can be combined with UI forward engineering process to produce not only more usable UIs in a logical way, but also to benefit from the reverse engineering to port a UI to any other target platform.

The Cameleon Reference Framework [23] locates UI development steps for context-sensitive interactive applications. A context is defined an element of the environments set considered for the interactive system, element of the platforms set considered for the interactive system and an element of the users set for the interactive system. A simplified version (Fig. 1) structures development for two contexts of use, here for two platforms: the one on the left represents the source and the one on the right represents the target. The development process can be decomposed into four steps:

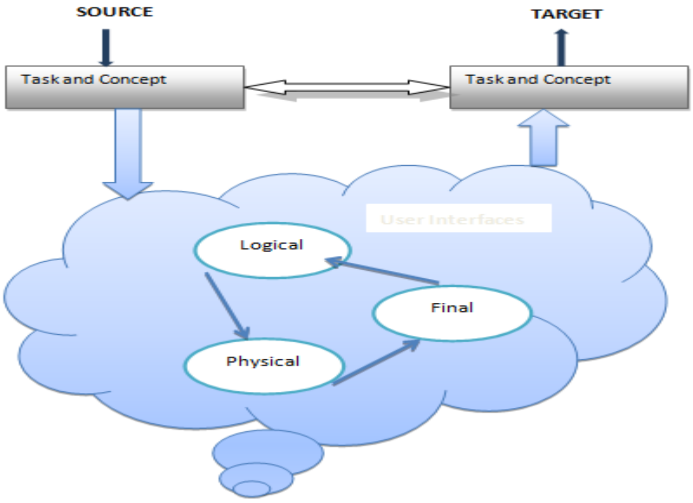


Figure2 UI Development Steps

Task and concepts describe the various tasks to be performed and the application oriented notions according to tasks to be performed. Logical UI is a basic symbol and notation for manipulation of the application concepts and routine in a way that does not dependent upon the perspective interacts present on the targets. The elements used in the logical UI are abstractions of existing product. Physical UI represents a logical UI into real Interaction objects to define product layout and interface navigation scheme. This interface is now composed of existing UI product. Final UI produced at the very last step of the objectification supported by a multi-target development environment and represented as source code.

4.2 V Model for Web Reengineering

The reengineered product goes through a complete web development life cycle and therefore it becomes mandatory for it to pass through complete testing cycle. The legacy system or product is transformed in new form by various means. The below figure illustrates the V model for the reengineering process. A V model as described below is proposed for designing the testing strategies for this category. Similar to the traditional V model, left side of the reengineering V model describes the stages of the design and coding and right side defines the corresponding stages of validation process. It has following phases:

a) Requirement Gathering for New Web Application

The first step involves the collection of the new requirements. This will list out the key points why reengineering is required for the software under consideration. Client get start discussing with the web development team about the newly generated requirement due to market evolution, technology changes and for the product improvement for better performance. System is reengineered in order to incorporate the new business requirements which involve functional and non-functional requirement. In this phase developer may make check list that deals with the various reason of reengineering is required.

b) Analysis of Existing Legacy System/Specification Building

The second stage is the study of the legacy system functionality and underlying design and come out with the difference with new functions. The nature of reengineering is to improve or transform existing system so it can be understood, controlled & reused as new system. Web reengineering is vital to restore & reuse the things inherent in the existing system, put the cost of system maintenance to the lowest in the control & establish a basis for the development of system in future

c) Reengineering of Application Migration

When existing systems become redundant, business switch from legacy systems to modern and new systems built on the latest technology / platforms. This switch is usually time consuming and expensive. A cost effective alternative to such scenarios is to reengineer, migrate or port the legacy systems into the latest technology / platforms.

d) Test Planning & Strategizing

The stage will include the test planning & test cases preparation if required as per the new requirements and strategizing the test execution for functional and non-functional areas. Test strategizing play an important role in carrying out the entire test execution program and involvement of high business risk, huge investments and mission critical systems make it important to strategizing the test phase. The best way is to identify the risky areas and the failure rate and then develop a test strategy.

e) Test Execution

This stage carry out the functional test execution as per plan defined in the previous stage. Test execution carry out performance testing, if the major design changes are there or the new requirements are related to improvement of performance. It test all the links in web pages, database connection, forms used in the web pages for submitting or getting information from user, cookie testing, test for navigation, content checking, interface testing include web server and application server interface, application server and database server interface.

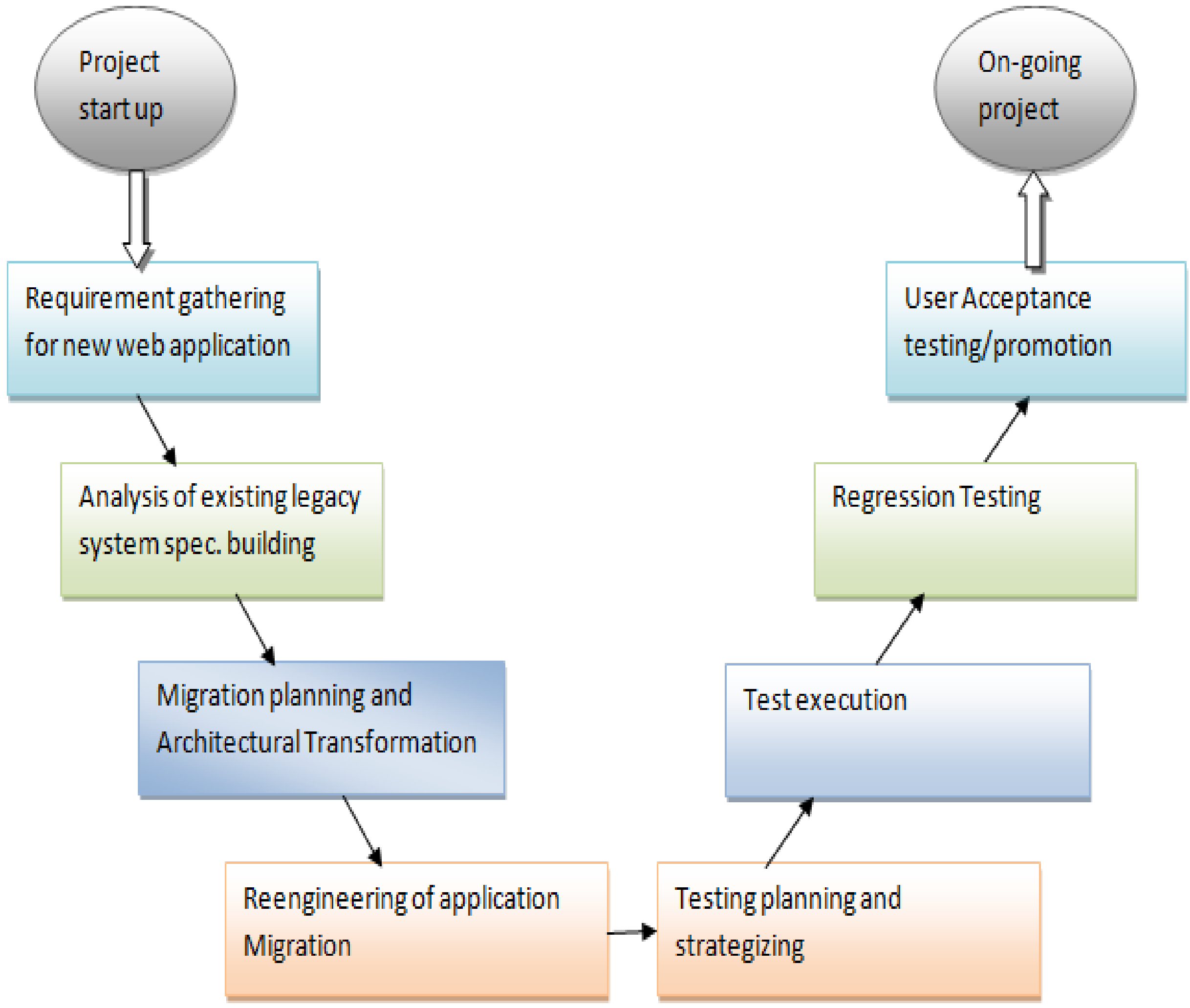

Project start up
Requirement gathering for new web application
Analysis of existing legacy system spec. building
Migration planning and Architectural Transformation
Reengineering of application Migration
Testing planning and strategizing
Test execution
Regression Testing
User Acceptance testing/promotion
On-going project

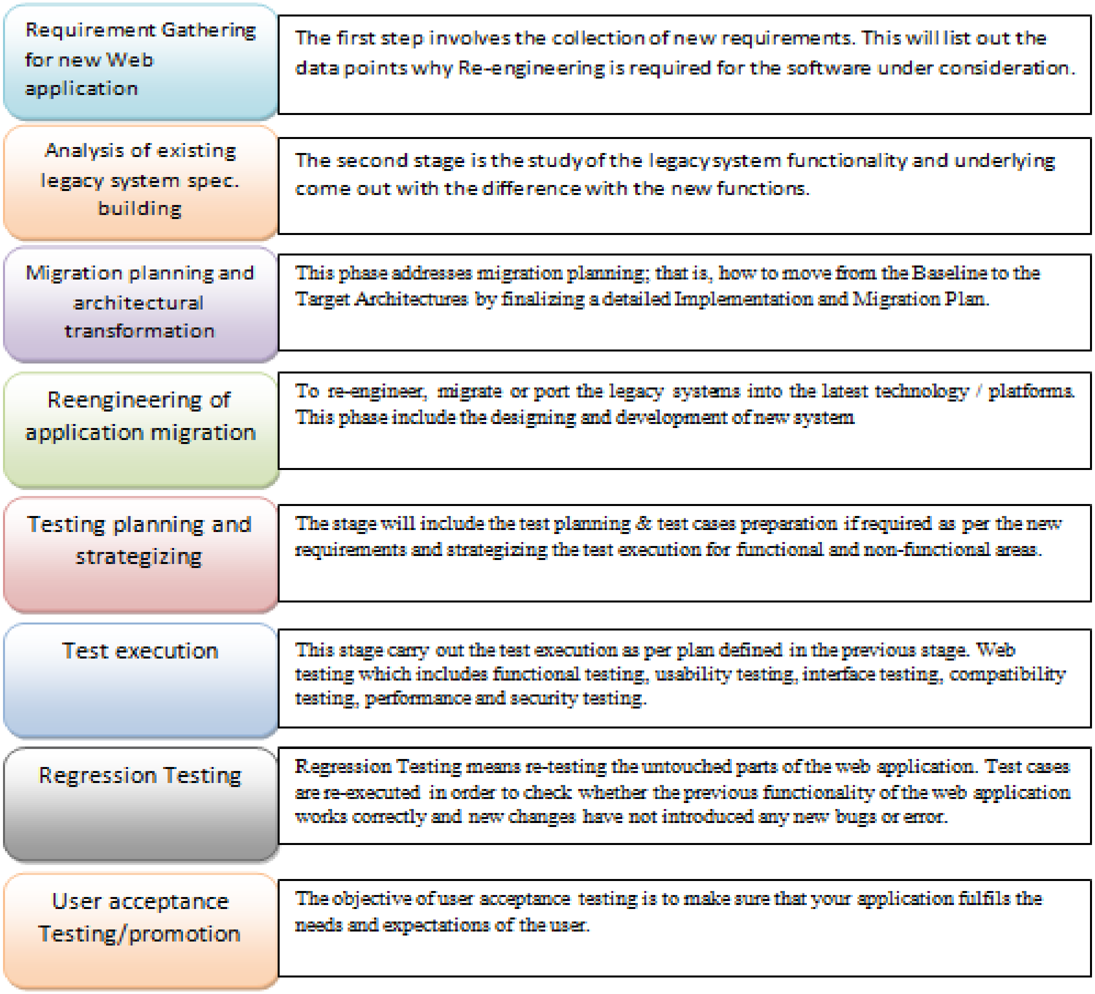
Figure 3 V Model for Web Reengineering

f) Regression Testing

Regression means retesting the effect of change in other parts of the web application. Test cases are executed again and again in order to check whether previous functionality of application work appropriately and changes made have not introduced any new bugs or error. This test can be performed on a new reconstructed system when new functionality added to it. A regression testing plan covers the updated functionalities. Many automated tool for regression testing is available for web application.

g) User Acceptance Testing

The purpose of user acceptance testing is to make sure your application meets the user’s expectations. It ensures that the application is ready to deploy services and change has been done effectively. The activities for user acceptance testing ensure browser compatibility, make sure that mandatory fields are given data in forms, check for time outs and field widths, and make sure that proper control is used to feed data.

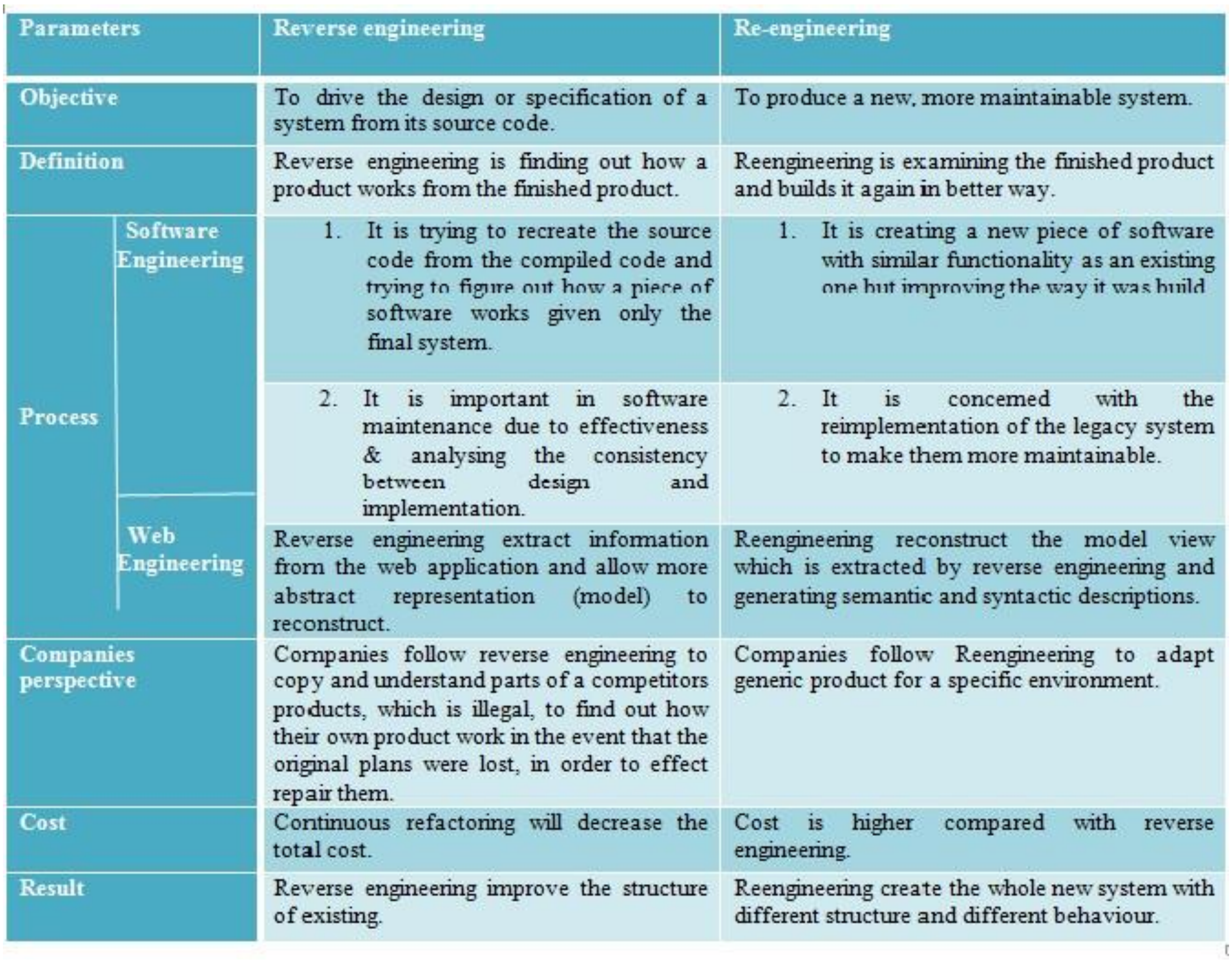


Requirement Gathering for new Web application	The first step involves the collection of new requirements. This will list out the data points why Re-engineering is required for the software under consideration.
Analysis of existing legacy system spec. building	The second stage is the study of the legacy system functionality and underlying come out with the difference with the new functions.
Migration planning and architectural transformation	This phase addresses migration planning; that is, how to move from the Baseline to the Target Architectures by finalizing a detailed Implementation and Migration Plan.
Reengineering of application migration	To re-engineer, migrate or port the legacy systems into the latest technology / platforms. This phase include the designing and development of new system
Testing planning and strategizing	The stage will include the test planning & test cases preparation if required as per the new requirements and strategizing the test execution for functional and non-functional areas.
Test execution	This stage carry out the test execution as per plan defined in the previous stage. Web testing which includes functional testing, usability testing, interface testing, compatibility testing, performance and security testing.
Regression Testing	Regression Testing means re-testing the untouched parts of the web application. Test cases are re-executed in order to check whether the previous functionality of the web application works correctly and new changes have not introduced any new bugs or error.
User acceptance Testing/promotion	The objective of user acceptance testing is to make sure that your application fulfils the needs and expectations of the user.

Figure 4 Descriptions of Stages for Web Reengineering V Model.

4.3 Reengineering Process Vs Reverse Engineering Process

As from earlier study, that we know that the reverse engineering is the part of reengineering. Therefore, there are some differences in their process. The section focuses on some intrinsic and extrinsic differences of reengineering and reverse engineering [24]. Intrinsic comparisons like web page processing were designed to be stateless but in software processing language make use of states and extrinsic parameter like cost and results. Listed below some differences:



Parameters		Reverse engineering	Re-engineering
Objective		To drive the design or specification of a system from its source code.	To produce a new, more maintainable system.
Definition		Reverse engineering is finding out how a product works from the finished product.	Reengineering is examining the finished product and builds it again in better way.
Process	Software Engineering	1. It is trying to recreate the source code from the compiled code and trying to figure out how a piece of software works given only the final system.	1. It is creating a new piece of software with similar functionality as an existing one but improving the way it was build
		2. It is important in software maintenance due to effectiveness & analysing the consistency between design and implementation.	2. It is concerned with the reimplementation of the legacy system to make them more maintainable.
	Web Engineering	Reverse engineering extract information from the web application and allow more abstract representation (model) to reconstruct.	Reengineering reconstruct the model view which is extracted by reverse engineering and generating semantic and syntactic descriptions.
Companies perspective		Companies follow reverse engineering to copy and understand parts of a competitors products, which is illegal, to find out how their own product work in the event that the original plans were lost, in order to effect repair them.	Companies follow Reengineering to adapt generic product for a specific environment.
Cost		Continuous refactoring will decrease the total cost.	Cost is higher compared with reverse engineering.
Result		Reverse engineering improve the structure of existing.	Reengineering create the whole new system with different structure and different behaviour.

Table 2 Comparison of Reverse Engineering v/s. Reengineering

Table 2 shows that reverse engineering is used during the software reengineering process for recovery of the structure of program that engineers use for better understanding of a program before re-organized its structure. The activities of reengineering process involves *source code translation* which converted program form an old programming to modern version and in data reengineering the data processed by the program is change to reflect program changes [25], reverse engineering is used to analyse the program and information extracted from the program which helps in document its organization and functionality and to improve program structure that controls the structure of program and modified to make it easier for understanding, program modularization is a part of reverse engineering that used to grouped together related parts of program.

4.4 Overview of Reengineering

There are many definitions to describe the reengineering process, each are slightly different but all have same overarching theme,” the radical redesign of business process to achieve dramatic improvement in productivity and performance”. The two terms are radical which means getting rid of existing processes, procedure and inventing new ways and dramatic improvement means a quantum leap in performance [26]. Reengineering is the main process in which organization becomes more modernize and efficient and to meet demands for quality service, flexibility and low cost, process must made simple, it transform an organization in way that directly affect performance.

4.4.1 Software Reengineering

The objective of software reengineering process is to improve maintenance, improve reliability, adaptation of new platform and enhance the functionality of the software system. Some problems are faced due to lack of well-designed structure through which any change will affects the entire system which lead to complex and expensive software system. The organization do not ruin the system because it was built for many subsidiaries of group which, if destroyed will result in the application process to be lost.

Therefore reuse the logic and component so that initial cost of developing logic and component of the software system are not wasted. The challenges of software reengineering are to use existing system, add the good features and attributes. Developers created new software system where goal is to maintain required function in the application of new technologies. Reengineering is used to prepare for advance functionality rather than to enhance existing function.

There are 3 levels of software reengineering [27]:

a) Domain Level,

b) System Level and

c) Component Level

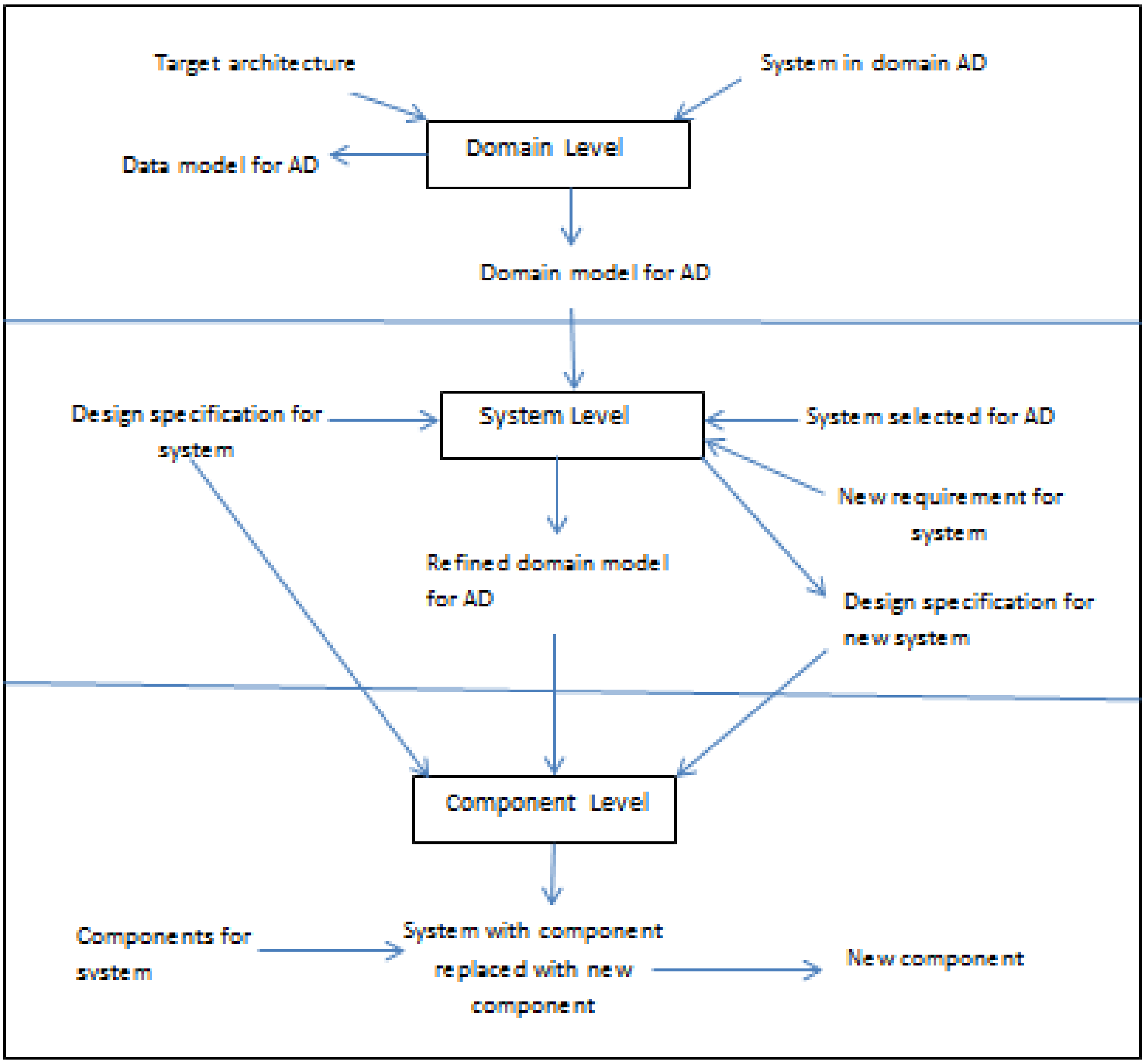


Figure 5 Stages of Software Reengineering.

Domain Level objective is to understand the system in application domain, prepare architecture framework for new system, analysis of domain [28]. *System Level* produces the complete logical design specification for a selected system in application domain, these design specification are produced for both the original and target system. *Component level* is the incremental reengineering the selects system, component by component approaches of software reengineering: To access the software reengineering there are three different methods [25]:

1. *Big bang Approach:* This approach is used when we want to replace the entire system at a time. This method is often used for projects that require being resolved immediately. It consumes too many resources and large amount of time before generating the desire system. The risk of this approach is very high.
2. *Incremental approach:* In this approach the part of the system is reengineered and incremental updates have been made as the new version of the system required to meet the new goals. The advantages of this approach is that component of the system are produced faster and easily control the errors when new components are clearly defined. This approach has lower risk than the big bang approach.
3. *Evolutionary approach:* In this approach part of the original system is replaced by the newly redesigned system and the parts are selected based on their function instead of existing system structure.

4.4.2 Web Reengineering

Web reengineering service helps to re-conceptualize and redesign your existing website and application service. Its objective is to re-structure the web application and its pages are modified to control and access standard and policies [29]. There are also 3 levels of the web reengineering [30]:

a) Design Recovery

b) Analysis and Evaluation

c) Redesign

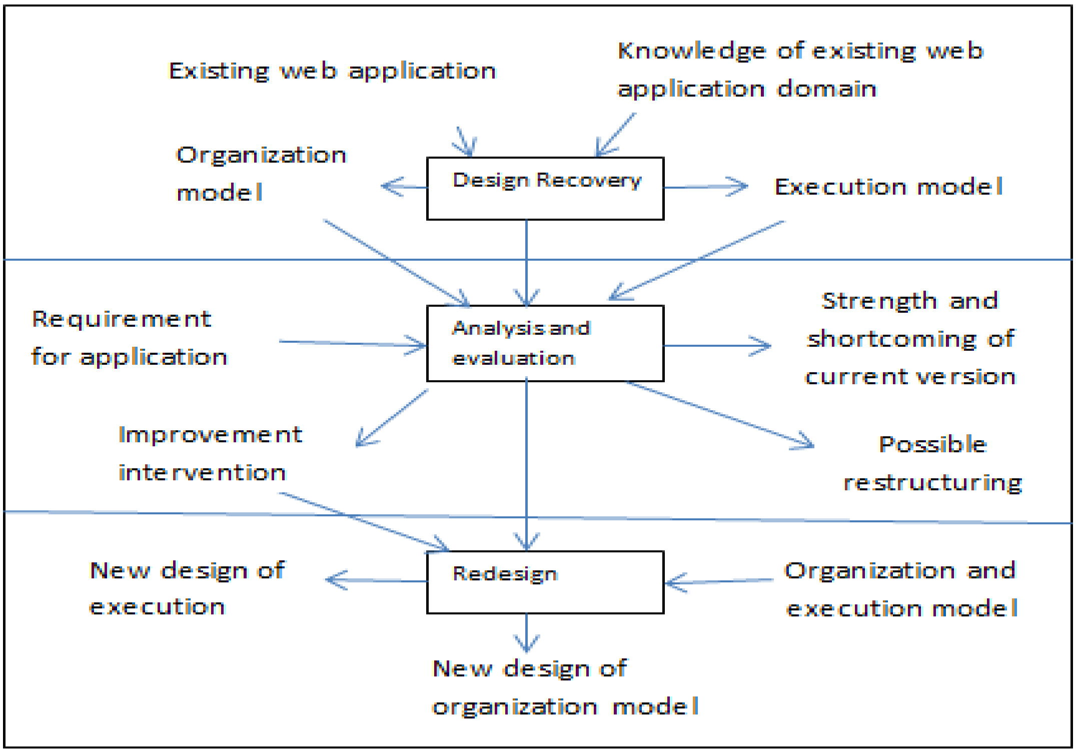

Existing web application
Knowledge of existing web application domain
Organization model
Design Recovery
Execution model
Requirement for application
Analysis and evaluation
Strength and shortcoming of current version
Improvement intervention
Possible restructuring
New design of execution
Redesign
Organization and execution model
New design of organization model

Figure 6 Stages of Web Reengineering

Design Recovery: For existing web site goal is to revise the models with data obtain by direct inspection and analysis of site's content and structure [31]. Model can be recreated using three steps of design recovery procedure:

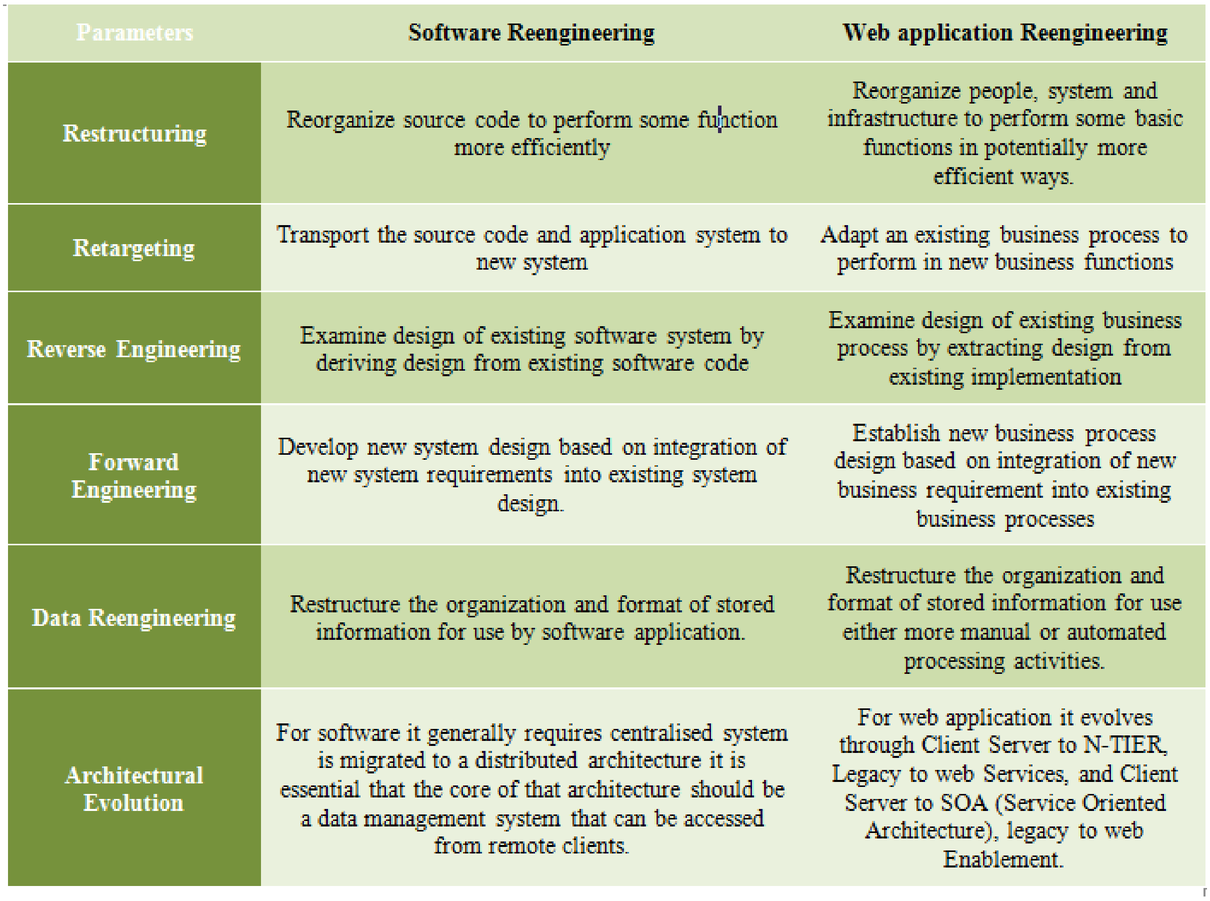
1) Formalization the transaction (execution and organization model);

2) Creating the execution model;

3) Construction of organization model. Use the automated reverse engineering process to improve the efficiency of the process.

Analysis and Evaluation: The result of design recovery procedure is to model a web application transaction using revised version of the execution and organization model. The next step of this process is to perform the user-oriented analysis and evaluation of the recovered execution and organization model. The objective of this step is to define the set of possible restructuring for the current design and implementation of the considered transaction addressing the strength and shortcoming highlighted by this level. *Redesign*: The objective of this level of reengineering is to redesign the execution and organization model and introduces the changes defined during analysis phase into recovered the transaction design model produces new design.

As we already discussed that reverse engineering and forward engineering subset of the reengineering process, so that development community of model based approach has shown the interest for forward engineering with many models such as domain model, dialog model, presentation model and application model [32], [33]. In table 3 has compared software reengineering with web reengineering with respect to the various perspectives like restructuring, retargeting, reverse engineering, forward engineering, data reengineering, business process reengineering, and architectural evolution.



Parameters	**Software Reengineering**	**Web application Reengineering**
Restructuring	Reorganize source code to perform some function more efficiently	Reorganize people, system and infrastructure to perform some basic functions in potentially more efficient ways.
Retargeting	Transport the source code and application system to new system	Adapt an existing business process to perform in new business functions
Reverse Engineering	Examine design of existing software system by deriving design from existing software code	Examine design of existing business process by extracting design from existing implementation
Forward Engineering	Develop new system design based on integration of new system requirements into existing system design.	Establish new business process design based on integration of new business requirement into existing business processes
Data Reengineering	Restructure the organization and format of stored information for use by software application.	Restructure the organization and format of stored information for use either more manual or automated processing activities.
Architectural Evolution	For software it generally requires centralised system is migrated to a distributed architecture it is essential that the core of that architecture should be a data management system that can be accessed from remote clients.	For web application it evolves through Client Server to N-TIER, Legacy to web Services, and Client Server to SOA (Service Oriented Architecture), legacy to web Enablement.

Table 3 Comparison of Software Reengineering v/s Web Application Reengineering

Web application must cope with an extremely short development evolution life cycle: A high level of flexibility, maintainability, and adaptability are actually necessary to compete and survive to market inflation. Unfortunately, to accomplish tight timing schedules to deliver web services, web applications are usually directly implemented without producing any useful documentation for their maintenance and evolution, and so those requirements are never be satisfied. In order to satisfy a growing market request for web applications and to deal with their increased technological complexity, we require specific methods and techniques able to support a disciplined and more effective development process. However, the high time pressure often forces the developers to implement the code of the application directly, without using disciplined development process, and this may have black effects on the delivered quality and documentation of the web application. This situation same as one occurring for traditional software produced in a short time, without respecting software engineering principles and using no disciplined development process. Poor quality and poor documentation must be considered the main factors essentially abortive and expensive maintenance, unattainability of applying more structured and documentation-based approaches.

4.5 STAR: Using the Situation/Tools/Application/Restructuring Paradigm to Define a Reengineering Process for Web Application

Reengineering process is usually run to abstract and extract data, document them from existing software, and to unified these documents and data with expert knowledge and previous experiences that cannot be instinctively reconstructed from software. According to the STAR paradigm, a reengineering process is characterized by situation, tools, application and restructuring. *Situation* defines a set of views of the applications to be reverse engineered. *Tools* include techniques and tools to support the information recovery process. *Application* describes which particular types of applications require to migrate from one technology to another. Lastly *Restructuring* specifies the actual implementation of reengineering process and also decides which type of reengineering process (transaction reengineering, data reengineering, graphic design reengineering etc.) is suitable for above paradigm .Possible situation, tools, application and renovating characterizing web application reengineering processes are presented below.

Situation

In the field of web applications, situation explains the possible scenario in which requirement of reengineering has emerged. A reverse engineering process may aid assessment of the characteristics of an existing application, in order to be able to evaluate its quality attribute, including reliability, security or maintainability [34].

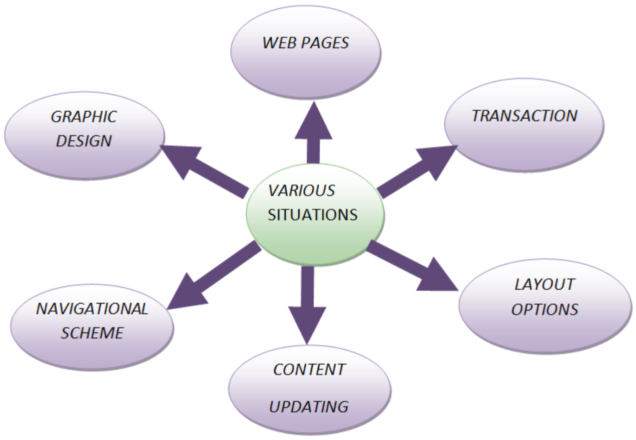


Figure 7 Various Situations for Web Reengineering

Several situations and scenarios are there to reengineer a web application such as change in web pages by filtering the entity, tags and elements of the web pages that involve assemblage of any HTML objects with given properties that require keeping all control mechanism and eliminating the unwanted tags and elements from web pages. Reengineering of web pages can be accomplished by identifying and analysing the intercommunication between elements and then altering these elements for the modification of new platform itself and generating the source code into suitable programming language. *Transformation in the layout options* and relationships that include alignment balancing which depend upon the position of the objects on the page and *content updating* of the web Pages according to the requirement changes, market evolution, usage and owner of website. Reengineering is used to modify a user interface to new context and to alter a navigational scheme that tells how web elements such as server page, client page, form, frame, email etc. are linked.

Tools

The recovery of information from an existing web application and the production of models, documentation of its relevant features that cannot be effectively accomplished without the support of suitable techniques and *Tools* that automate the web application analysis. However, the diverse and dynamic nature objects producing the application, and the lack of effective mechanisms for implementing the basic software engineering principles in web applications, that makes analysis process more complex and make it necessary to address specific methodological and technological problems. Some transcoding tools [35], [36], [37], [38] automatically transform a UI code from the original platform to a target platform.

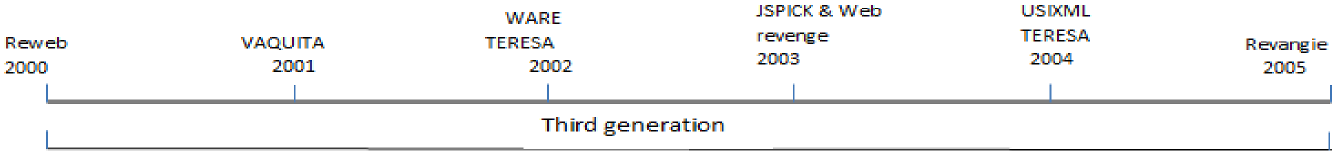


Figure 8 Evolutions of Reverse Engineering Tools and Methodologies

More precisely, heterogeneous software components developed with different technologies and languages require techniques and tools for multi-language analysis. The existence of ‘dynamic software components’ in a web application, such as pages created at runtime depending on user input, will impose the application of dynamic analysis techniques and static analysis of the code in order to obtain more precise information about the web application behaviour. In addition, the absence of effective mechanisms for implementing the software engineering principles of modularity, encapsulation, and separation of concerns, will make the use of suitable analysis approaches, such as program slicing which is necessary in order to localize more cohesive parts in the web application code. Finally, on the basis of the situations, tools / technique and application identified the sequence of activities composing the reengineering process, their input/output and responsibilities will be precisely set out. The reverse engineering process can be executed with the support of various engineering tools proposed as reweb [39], vaquita [40], WARE [41], [42], TERSA [43], JSPICK [44], web revenge [45], UIML [46] and revangie tool.

Applications

Web development within an organisation depends upon several factors. The quantity and the radiance of web applications have growing rapidly over year by year. Since the quantity and impact of security vulnerabilities in such applications have grown as well [47] .The motivation depends upon the initial purpose of web usage, the customer’s expectations and the competitive environment. The drive to systematise development is subject to overall view of the web and conscious policy decisions within the organisation. For example, a low level view of the web is lead to ad hoc. Initially we need to understand the problem domains that currently addressed by web. Table 4 presents categories of web applications. Organizations that started their web development early may also have followed a similar order in the past. Although, it is possible to start web development with applications in any category, this table has been useful to explain to organisations with modest presence on the web how they might improve or benefit from incremental exposure, thus keeping the risks to the minimum.

Category	*Examples*
Workflow	Planning and scheduling systems, inventory management, status monitoring
Collaborative work Environments	Distributed authoring systems, collaborative design tools
Informational	Online newspapers, product catalogues, newsletters, service manuals, classifieds, e-books
Interactive ▪ **User-provided information** ▪ **Customized access**	Registration forms, customized information presentation, games
Online communities, Marketplaces	Chat groups, recommender systems, marketplaces, auctions
Web Portals	Electronic shopping malls, intermediaries

Transaction	E-shopping, ordering goods and services, banking
Web Services	Enterprise applications, information and business Intermediaries

Table 4 Categories of Web Application

Migration of applications to the newer technologies can give business a leading edge by removing inefficient workflow and processes while preserving original objectives, model and investment. We can help enterprises in migration of the legacy systems from old technologies to present day platforms. Reengineering strategically designed to overcome the cross platform compatibility challenges. Due to upcoming advance technology and growing business states, there is need for the migration of legacy software systems to new technologies and environments. There are different kind of legacy system reengineering services that includes language and database migration, platform-to-platform porting and system redevelopment.

A web application must follows the enterprises standard and rules implemented in a legacy application, while transforming those to new business and architecture requirements, to produce a flexible, tested or validated modified system. Reengineering and Migration benefits are the saving time and effort, enhancements in operational efficiency, benefits of the latest technologies and platforms

Restructuring/Re-Conceptualisation

Reengineering is the analysis of existing software system and modifying it to constitute into a new form. Chikofsky and Cross define reengineering as 'the examination and alteration of a subject system to reconstitute it in a new form and subsequent implementation of that form' [7].According to IEEE Std. 1998 'A system changing activity that results in creating a new system that either retains or does not retain the individuality of the initial system' [8]. Techniques of static and dynamic analysis of the source code and dynamic data were taken into account.

Additional techniques for analysing the web application structure and identifying relevant subsets of its components were also considered where clustering techniques were defined to carry out this analysis. Finally, the specifications of the tools required to support these analyses could be defined.

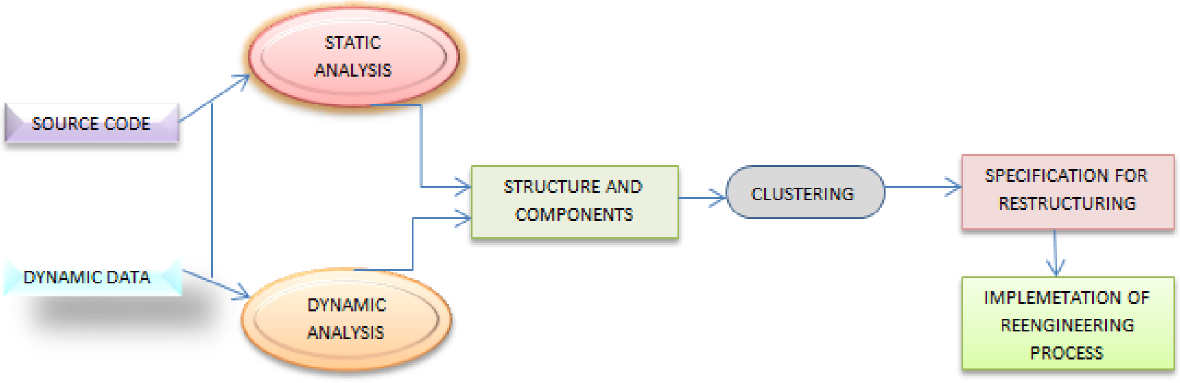


Figure 9 Steps of Reengineering Process

Static Analysis

Web application security is accomplished by static analysis and runtime analysis. Web application security has been great challenge for this static analysis tool such as ASPWC is used to detect the attack and vulnerabilities based on taint analysis [48]. Static analysis does not require the execution of the application.

It recovers web application architecture components and the static relations among them. HTML files, directory structure, scripting language sources as well as any other static information (e.g., database connections, use of applets/servlets) are processed. HTML pages and page sub elements (frames, forms, widgets) composing the given page are localized, classified and recorded in an intermediate representation. Central to the reverse engineering process is the mapping between web application elements and object oriented entities, according to Conallen proposals [49] [50][51].

Dynamic Analysis

In a dynamic web application, the set of entities making up the application can be significantly modified at run-time, thanks to the facility offered by script blocks, of producing new code that is enclosed in the resulting client pages, or exploiting the possibility of producing dynamic results offered by active web objects (such as Java applets or ActiveX objects). Therefore, in the second step of the process, dynamic analysis is executed with the aim of recovering information about the WA Components and Direct Relations.

The dynamic analysis phase relies on the static analysis results. The web application is executed and dynamic interactions among the components are recorded. It is performed observing the execution of the web application, *tracing* to source code any event. Traced events are those observed by the user or related to components external to the web application *(e.g.,* third party databases or web sites). Events are the HTML pages/frames/forms visualization, the submission of forms, the processing of data, a link traversal, or a database query, etc. All elements responsible of these actions (typically links, scripts, applets) are localized.

Web Application Connection Graph

The information about the web application obtained by static and dynamic analysis can be used to produce a graph whose nodes represent the set of web application entities, and whose edges describe the different relationships among these entities. In [27], this kind of graph has been named as WAG; *Web Application Connection Graph.* WAG analysis may support the comprehension of the application. However, since this graph may be large (in terms of the number of nodes and edges) even in the case of small size web applications, in order to simplify the analysis of large WAG graphs, some kind of automatic clustering [67] can be used to decompose this graph into smaller cohesive parts. This clustering approach evaluates the degree of coupling between entities of the application (such as server pages, client pages and client modules) that are interconnected by *Link, Submit, Redirect, Build, Load in Frame*, and *Include* relationships. After getting more abstract specification from clustering algorithms actual implementation of reengineering process is executed according to selected situation and application.

Automatic Clustering of the Web Application

In the third step of the reverse engineering process the problem to group together set of Components collaborating to the realization of a functionality of the web application are addressed. An automatic algorithm partitioning the components of the web application in a set of clusters, on the basis of the information extracted during the first two steps of the reverse engineering process, has been defined and is described in the following chapter.
The obtained clusters are analysed by a human expert in order to identify the functionalities that they realize.

Clustering Methodology for Web Applications

The goal of the proposed clustering method is to group software components of a web application into meaningful (highly cohesive) and independent (loosely coupled) clusters. According to Anquetil et al. [52], three issues must be considered to do clustering. The first issue is to build a model in which the components to be clustered are adequately described. The second one consists of defining when a set of components should be clustered into a cohesive unit, and the third issue consists of selecting the clustering algorithm to be applied. There are many different clustering algorithms in the literature [53]. Some of them have been exploited in the field of software remodularization [54][55], to support reverse engineering [56], [57] or program comprehension [58, 59]. A possible taxonomy distinguishes between hierarchical and non-hierarchical ones. Anquetil et al. [60] experimented with several clustering algorithms, and their results show that hierarchical clustering provides as good results as other ones. A hierarchical clustering has been adopted in this approach, since it can be used to obtain different partitioning of a system at different levels of abstraction.

Different type of reengineering processes are generated based upon the above three perspective are listed in table 5.

Data Reengineering	The process of analysis and re-organising the data structure and to understand data values in asystem.
Transaction Reengineering	In transaction oriented website, the user executes a series of activities in order to carry out a specific task. Business processes are realized by means of transaction, in which this context can be interpreted as high level work flows corresponding to user tasks.
Application Migration	Migration application to the new technologies can give business leading edge by removing inefficient workflow and process while preserving original objectives, model and investment.
Graphic Design Reengineering	It is used to modify a user interface to different context. To change user interface developers do not necessarily want to start from scratch to design a UI for a new platform since UI already exists.
Reengineering of Web pages	Reengineering of web pages can be accomplished by detecting and analysing the interaction of object and transforming these objects for the adoption of new platform itself and generating the source code into new language.
Business Process Reengineering	It is the analysis and redesign of workflow within and between enterprises. It defines all the process in an organization and prioritizes them in order to redesign urgency.

Table 5 Different types of Reengineering Processes

Let us take an example illustrating real motive of STAR paradigm, suppose there is situation of graphic design need to be alter then we can have transcoding tool TERESA[36] that automatically transform a graphic interface code from the original platform to a target platform with any of the application suppose Interactive websites (User-provided information Customized access)

CHAPTER 5

WEB APPLICATION QUALITY PREDICTION

Web Application Quality Prediction

5.1 Introduction

Web applications are different from traditional software system in terms of the involvement of diverse technologies in software as well as in hardware. Basic building blocks of web application are web pages which are of two type static and dynamic web pages. Static web pages are written in HTML and dynamic web pages normally written using scripting languages such as PHP. Dynamic pages are generated at run time and the back end typically written using database management language. Web development can benefit from traditional application and practices from other associated areas, it has certain discriminate attributes that wish remarkable considerations. An influential entity of any web engineering process is metrics. Web metrics are used for better understanding of any web page attributes and elements that we develop.web metrics is used most importantly to determine the quality of web application. Since metrics are important source of information for decision making, a large number of web metrics have been proposed in the last decade to compare the structural quality of a web page [61].

5.2 Web Metrics Description Selected for Study

There are several researchers who proposed many web metrics [62] [63]; from them we collected various metrics for our research work. Web metrics are used to identify the attributes of the web page that we create. Most importantly we used we use web metrics to assess the quality of web page. Since web metrics are key source of information many web metrics have been proposed in last year to determine structural quality of a website.
A precise description of web metrics that we have used in our study is text below:

1) *Total word length*

Total number of words on a page including special character *, $, # is taken.

2) *Body text length*

This metrics calculates the total words on a body of a web page; it also counts the special symbol in a body.

3) Title text length

This metrics tells the total number of words in title of a page and calculated by counting the number of words present in a title of web page.

4) Total links

There total number of links present in a web page including internal links, external links, embedded links, wrapped links etc. This can be calculated by counting the total links present in a web page.

5) Internal links

There are links that are internally linked in web site and can be calculated by counting the total internal links.

6) Size of page in KB

It refers to the size of a web page in kilobytes.

7) Emphasize text

This metric is calculated by counting total emphasized text that is the text which are in bold, capital and italic etc.

8) Total no. of !'s

It shows total occurrences of exclamation sign on a page and calculated by counting all exclamatory signs on a web page.

9) Average Number of Lines

This metrics shows number of lines for all functions or methods on a web page.

10) HTML Lines

This metric refers to number of all html lines.

11) *Java Script Lines*

This metric refers to number of all Java Script lines.

11) Blank Lines of Code

This metric can be calculated by counting all the blank lines presents in the code of a web page.

12) Source Lines of Code

This metric tells number of lines containing source code.

13) Statements

It counts the number of statements on a web page.

14) Cyclomatic Complexity

The structural complexity of website is determined with Mc. Cab's cyclomatic complexity metric [64]. This metric is used to know navigation path for a desired web page. The cyclomatic complexity metric is derived in graph theory.

15) Nesting

This metric tells maximum nesting level of control constructs.

16) Frames

This metric can be calculated by counting total number of frames used on a page.

17) Total list

This metric refers to total number of list on a web page and can be calculated by analysing the entire ordered and un-ordered list present in a web page.

18) Number of tables

This metrics tells the number of tables is used in making a web page.

Total words length	Total words on a page
Body text length	Total words in body that are displayed.
Title text length	Words in title of the page
Total links	Total links including external, internal, embedded links etc.
Internal links	Total internal links of a web page
Size of page in KB	Total size of a page in kilo byes
Emphasize text	Total emphasized text
Total no. of !'s	Exclamation sign on a page
Average Number of Lines	Number of lines for all functions or methods.
HTML Lines	Number of all html lines.
JavaScript Lines	Number of all java Script lines.
Blank Lines of Code	Number of blank lines
Source Lines of Code	Number of lines containing source code
Lines with Comments	Number of lines containing comment
Statements	Number of statements
Cyclomatic Complexity	Cyclomatic complexity
Nesting	Maximum nesting level of control constructs.
Frames	Total number of frames used on a page
Total list	Total list on a page
Number of tables	Number of tables present on web
Graphic	Total number of images

Table 6 List of Metrics

5.3 Research Methodology

This section has described the descriptive statistics for all the metrics that have been studied. The section also explains the methods used for the analysis of metrics chosen and also for the model prediction. This section also have quantitative analysis of web metrics like total number of links, complexity of web page, embedded text, total graphics etc. These metrics have been calculated from the websites downloaded from the webby awards site. The webby awards site contains the 70 number of different categories of websites like news, entertainment, sports, kids, education, fashion etc.

5.3.1 Data Collection Process

The study has selected nearly 200 websites having different categories from webby award site and downloaded the home page of each website and the process these web pages using the tool that has made using MATLAB. For evaluation there have been selected 94 good data points and 101 bad data points. The process of data collection is explained in block diagram.

To honour excellence on the Internet, there is a leading International award named, The Webby Award, which was established in 1996 during the Web's infancy. It is presented by the International Academy of Digital Arts and Sciences (IADAS), which includes more than thousand members for judging including executive members of leading web experts, creative celebrities, business figures, visionaries, luminaries and Associative members.

From all over the world, The Webby People's Voice Awards are garnered by millions of vote [65].

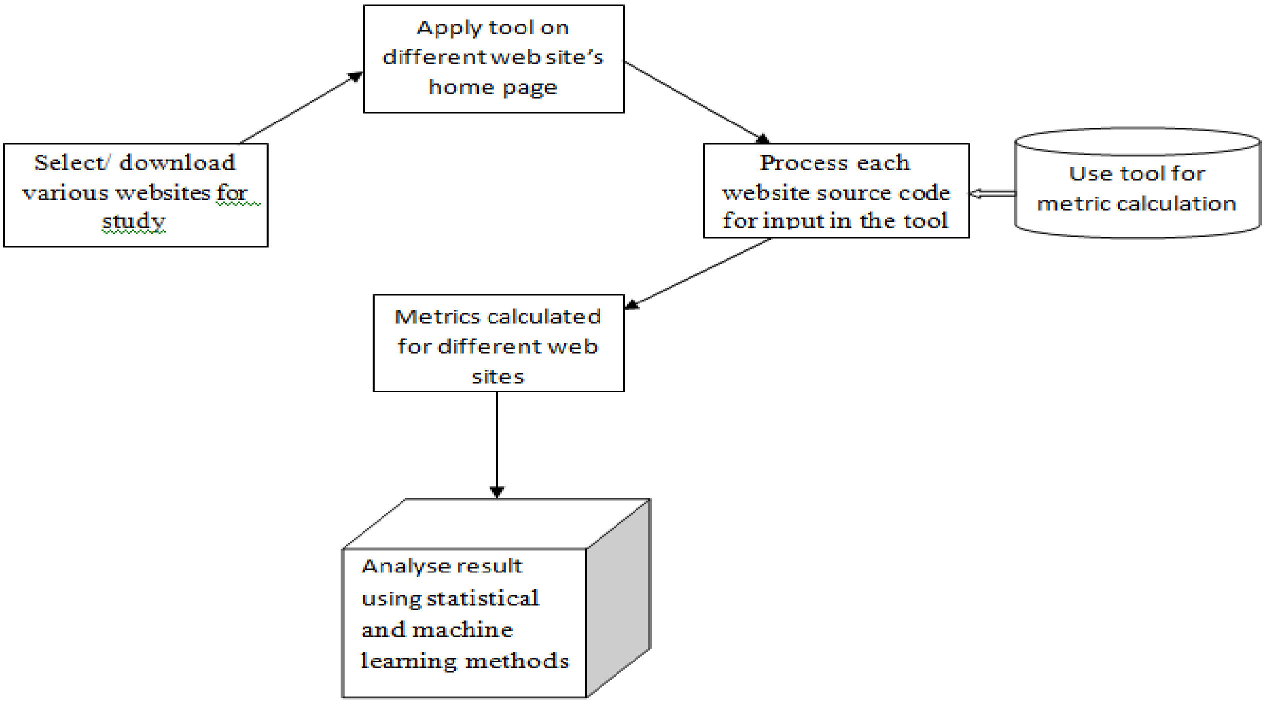
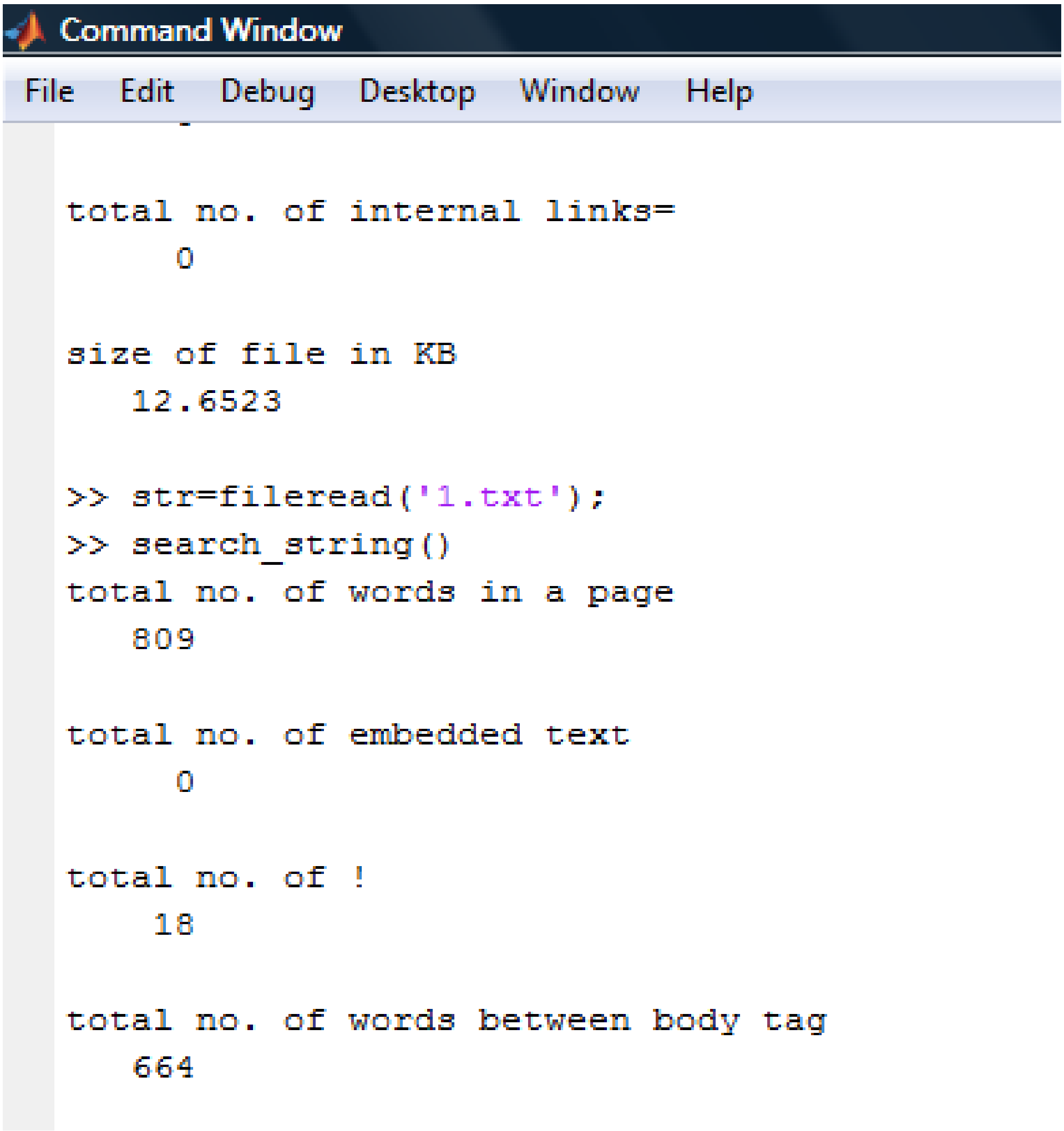

Apply tool on different web site's home page
Select/ download various websites for study
Process each website source code for input in the tool
Use tool for metric calculation
Metrics calculated for different web sites
Analyse result using statistical and machine learning methods

Figure 10 Block Diagram of Research Methodology

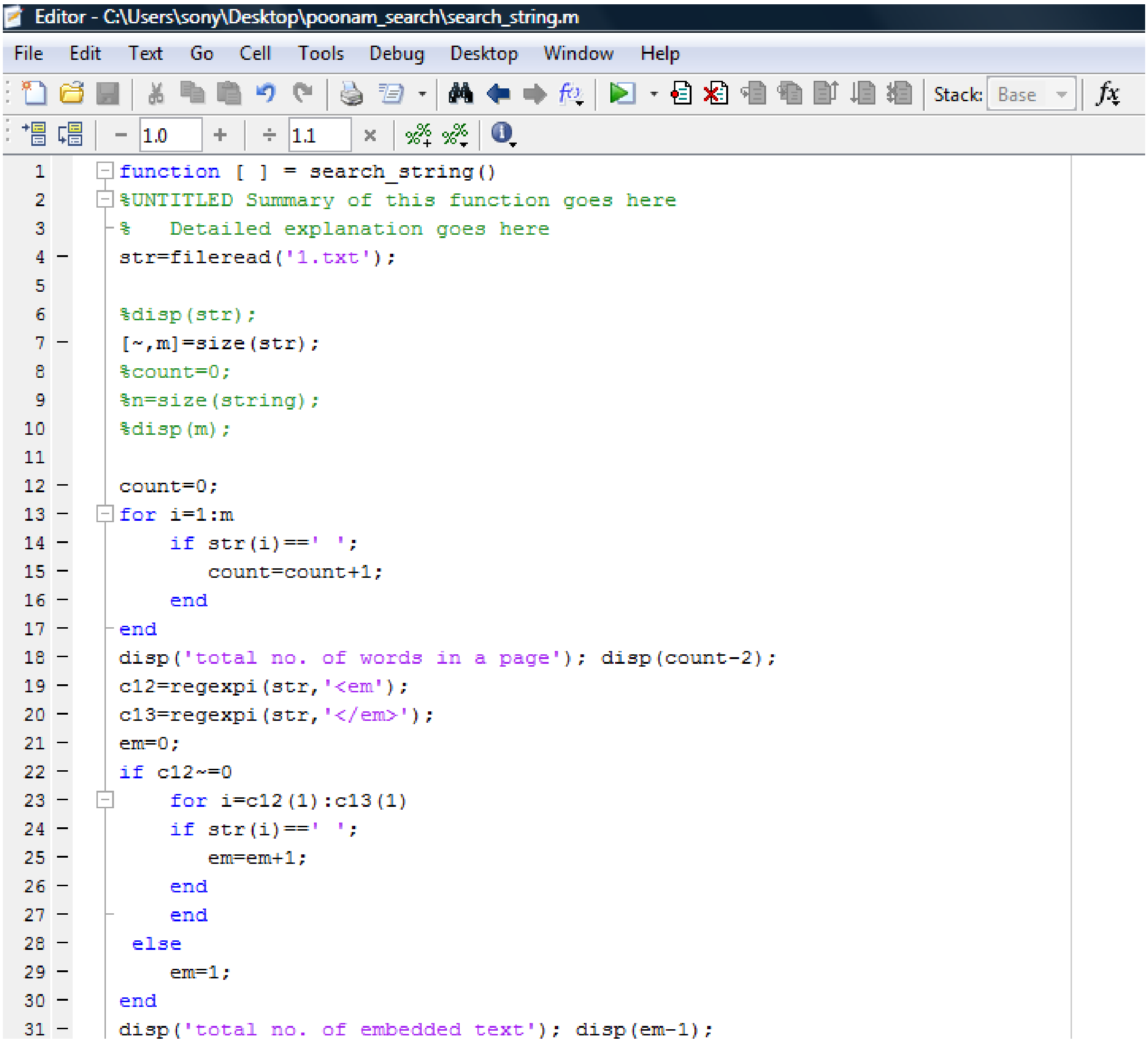
5.3.2 Description of Tools Used

Metrics have been calculated with the help of tool built in MATLAB. This tool is used to produce web design and formatting metric e.g. table, graphics, bogy text length etc. Input of this tool is the source code pages of sites collected from webby award site. Snapshots of the code and output window have been shown below:



```
total no. of internal links=
     0


size of file in KB
   12.6523

>> str=fileread('1.txt');
>> search_string()
total no. of words in a page
   809


total no. of embedded text
     0


total no. of !
    18


total no. of words between body tag
   664
```

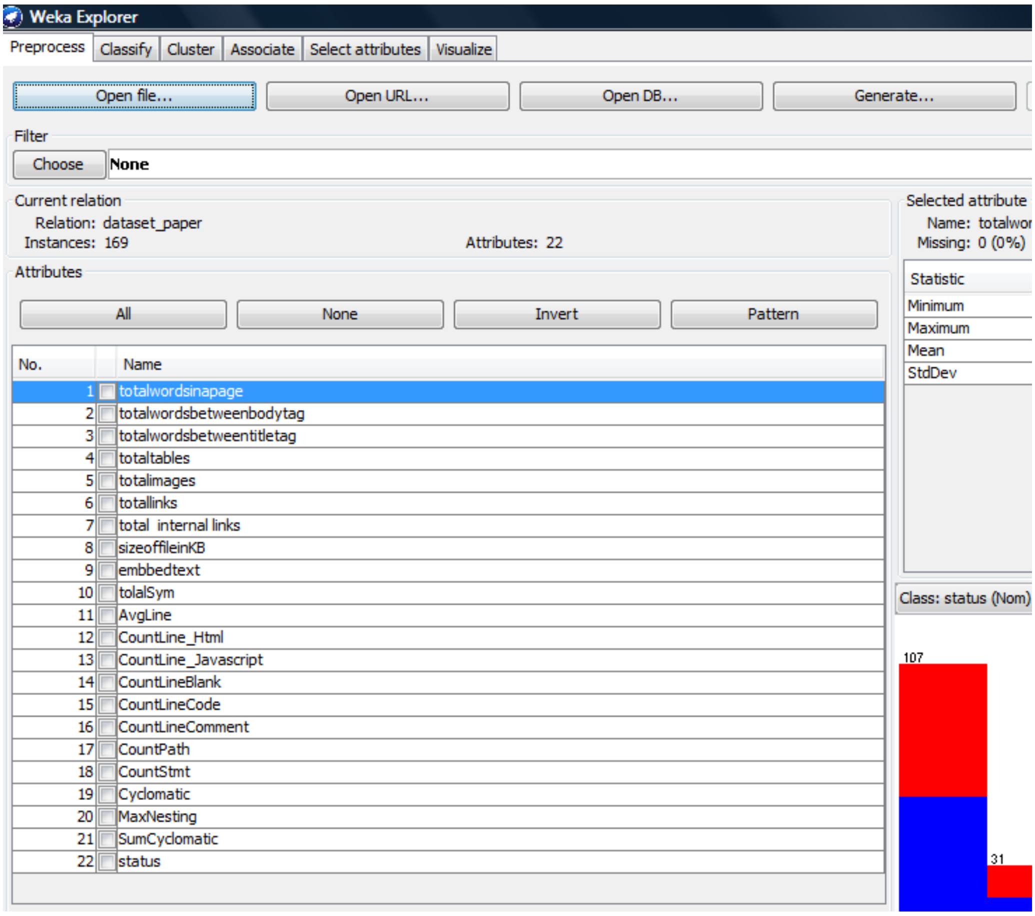
Figure 11 Image of MATLAB Tool for Metric Calculation



```
function [ ] = search_string()
%UNTITLED Summary of this function goes here
%   Detailed explanation goes here
str=fileread('1.txt');

%disp(str);
[~,m]=size(str);
%count=0;
%n=size(string);
%disp(m);

count=0;
for i=1:m
    if str(i)==' ';
       count=count+1;
    end
end
disp('total no. of words in a page'); disp(count-2);
c12=regexpi(str,'<em>');
c13=regexpi(str,'</em>');
em=0;
if c12~=0
    for i=c12(1):c13(1)
    if str(i)==' ';
       em=em+1;
    end
    end
 else
    em=1;
end
disp('total no. of embedded text'); disp(em-1);
```

Figure 12 Image of MATLAB Tool

WEKA Tool



Filter
Choose None

Current relation
Relation: dataset_paper
Instances: 169
Attributes: 22

Selected attribute
Name: totalwor
Missing: 0 (0%)

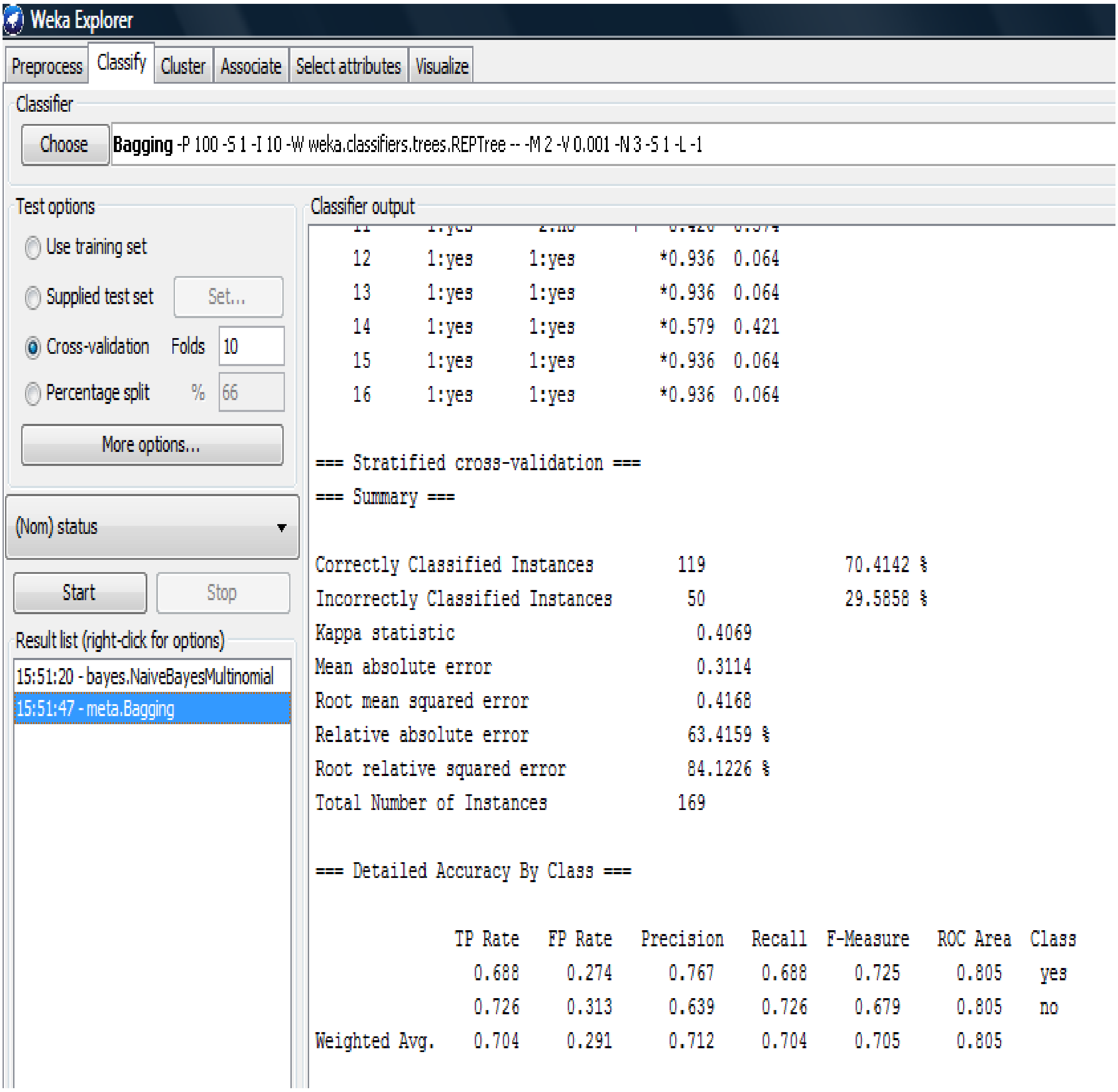
Statistic
Minimum
Maximum
Mean
StdDev

Attributes

No.		Name
1	□	totalwordsinapage
2	□	totalwordsbetweenbodytag
3	□	totalwordsbetweentitletag
4	□	totaltables
5	□	totalimages
6	□	totallinks
7	□	total internal links
8	□	sizeoffileinKB
9	□	embbedtext
10	□	tolalSym
11	□	AvgLine
12	□	CountLine_Html
13	□	CountLine_Javascript
14	□	CountLineBlank
15	□	CountLineCode
16	□	CountLineComment
17	□	CountPath
18	□	CountStmt
19	□	Cyclomatic
20	□	MaxNesting
21	□	SumCyclomatic
22	□	status

Class: status (Nom)

Figure 13 Image of WEKA Tool



Bagging -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) status

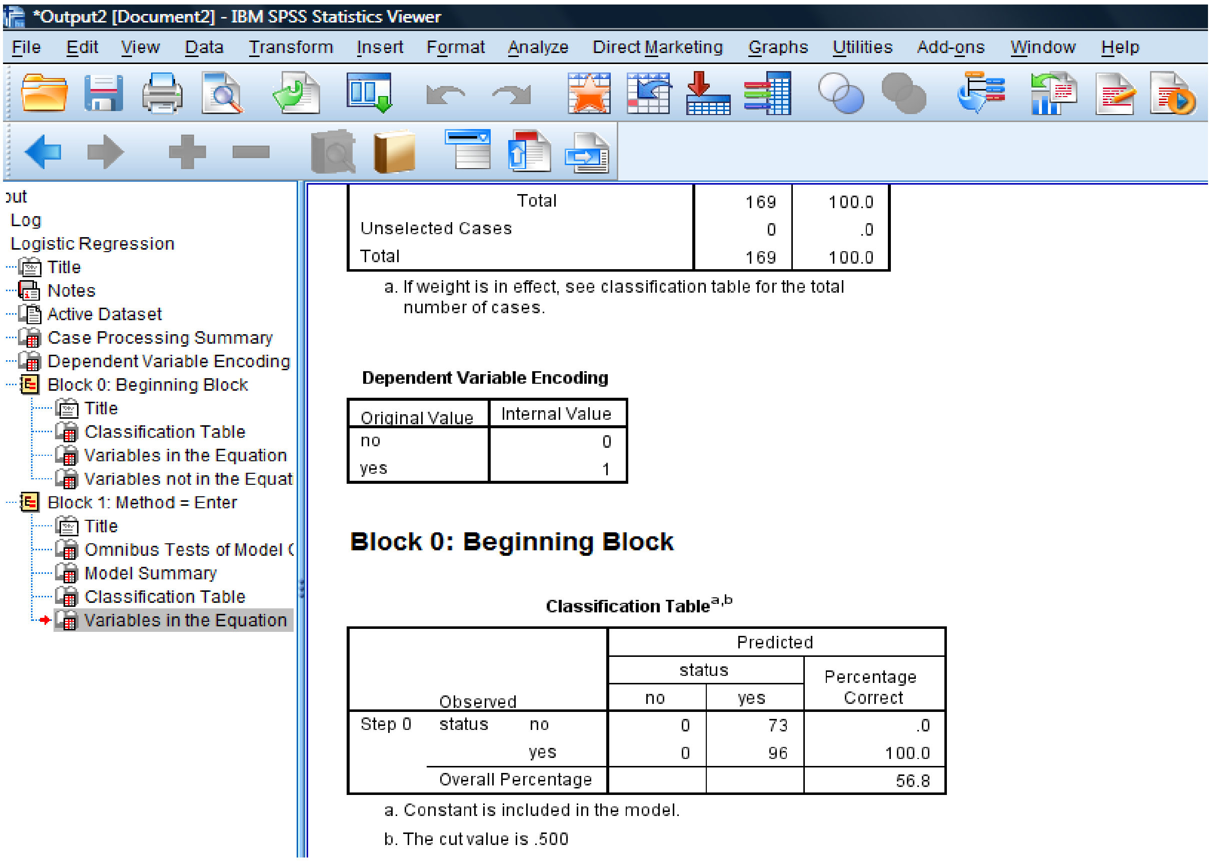
15:51:20 - bayes.NaiveBayesMultinomial
15:51:47 - meta.Bagging

```
    12       1:yes       1:yes          *0.936   0.064
    13       1:yes       1:yes          *0.936   0.064
    14       1:yes       1:yes          *0.579   0.421
    15       1:yes       1:yes          *0.936   0.064
    16       1:yes       1:yes          *0.936   0.064

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         119               70.4142 %
Incorrectly Classified Instances        50               29.5858 %
Kappa statistic                          0.4069
Mean absolute error                      0.3114
Root mean squared error                  0.4168
Relative absolute error                 63.4159 %
Root relative squared error             84.1226 %
Total Number of Instances              169

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.688     0.274      0.767     0.688     0.725      0.805    yes
                 0.726     0.313      0.639     0.726     0.679      0.805    no
Weighted Avg.    0.704     0.291      0.712     0.704     0.705      0.805
```

Figure14 Image of WEKA Tool

SPSS Tool

SPSS is a Windows based program that can be used to perform data entry and analysis and to create tables and graphs. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. SPSS is commonly used in the Social Sciences and in the business world. SPSS is updated often.


*Output2 [Document2] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
›ut
Log
Logistic Regression
Title
Notes
Active Dataset
Case Processing Summary
Dependent Variable Encoding
Block 0: Beginning Block
Title
Classification Table
Variables in the Equation
Variables not in the Equat
Block 1: Method = Enter
Title
Omnibus Tests of Model (
Model Summary
Classification Table
Variables in the Equation

Total	169	100.0
Unselected Cases	0	.0
Total	169	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
no	0
yes	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			status		Percentage Correct
Observed			no	yes	
Step 0	status	no	0	73	.0
		yes	0	96	100.0
	Overall Percentage				56.8

a. Constant is included in the model.

b. The cut value is .500

Figure 15 Image of SPSS Tool

After processing the web pages of the selected websites, determined the various web metrics like body text length, page size, total number of tables, frames etc. After determining the all metrics for both good and bad data points analysis of the relationship between each metrics with the status of the website that is good or bad sites has been done. This is accomplished by using univariate analysis with the help of SPSS tool and predicting the model using seven machine learning algorithm using weka tool [66]. Result of this tool is fed into the SPSS tool to predict the ROC curve which helps in predicting the best method for model prediction.

5.2.2 Descriptive Statistics

Below table presents the mean, median, standard deviation, standard error of mean minimum, maximum for all the metrics that we have considered.

Metrics	Mean	Std. Error of Mean	Median	Std. Deviation
Total words length	8005.272	771.625	4159.0	10031.1
Body text length	6318.467	751.595	3022.0	9770.740
Title text length	3.491	.3705	1.548[a]	4.8170
Table count	1.166	.3989	.282[a]	5.1855
Frame count	.000	.0000	.[a]	.0000
Graphics	28.769	4.9922	12.00	64.8988
List count	.006	.0059	.006[a]	.0769
Links	150.639	13.7017	86.0	178.1218
Internal links	1.042	.2602	.221[a]	3.3624
Size of Page In KB	120.08	12.36	79.20	160.723
Emphasize Text	19.179	13.4214	.142[a]	173.9611
Total No. of !’S	70.892	12.2547	42.000[a]	158.3657
Average Number Of Lines	3.627	.3423	2.750[a]	4.4505
HTML Lines	721.225	135.639	274.000	1763.3180
JavaScript Lines	91.544	12.2447	22.000[a]	159.1810
Blank Lines Of Code	264.089	117.130	28.750[a]	1522.6956
Source Lines Of Code	74.976	10.0594	18.000[a]	130.77
Lines With Comments	35.024	6.1178	9.286[a]	79.5308
Path	6965645.0118	6002929.718	1.213	78038086.335
Statements	79.751	11.5875	29.250[a]	150.6374
Cyclomatic Complexity	3.343	.5212	1.299[a]	6.7762
Nesting	.988	.0916	.770[a]	1.1902

Table 7 Statistical Description of Web Metrics

5.2.3 Machine Learning Model

Along with statistical approach there also used various machine learning algorithm. The thesis has used machine learning method to determine correctness of model using all metrics. There are many machine learning method among which we have used seven methods such as Adaboost[67], Bagging[68], Random Forest[69][70], Multilayer Perceptron[71][72][73], bayesNet[74] , Decision Table[75], Naive Bayes

Multinomial[76][77]. To evaluate quality of websites we have used seven machine learning methods and these methods are being implemented in WEKA TOOL[66] which is freely available open source tool also the research work have used SPSS tool to generate ROC curve for these machine learning methods. This thesis have summarised the description of the observed machine learning methods from various research paper as follows.

Bagging

Bagging method is used to generate multiple version of a predictor and to aggregate this predictor. Bootstrap replicates the learning set that is used to form multiple version of predictor then after used it as new learning set. Bagging is also known as bootstrap aggregating that repeatedly samples from a data set according to uniform probability distribution [68]. On average, a bootstrap sample Di contains approximately 63% of the original training data because each sample has a probability 1- (1- 1/N)N of being selected in each Di. If N is sufficiently large, this probability converges to 1-1/e = 0.632. After training the k classifiers, a test instance is assigned to the class that receives the highest number of votes [78].

Multi Layer Perceptron

A MLP network is one of the most important neural network models and has been used for pattern classification. It is proved to be a powerful technique. This neural network can be trained to form arbitrary surfaces in input space. Mashor showed in [71] that the MLP network was highly nonlinear and ever modelling a linear model using the standard nonlinear network is never be the best solution. The MLP network consists of a set of input layer, one or more hidden layer and an output layer. It has been proven by Cybenko [72] and Funashashi [73] that the MLP network with one hidden layer is always sufficient to approximate any continuous function up to certain accuracy.

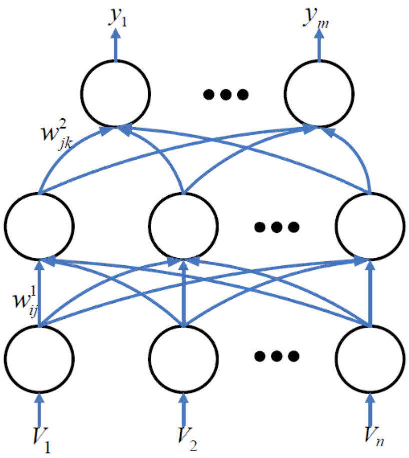


Figure 16 MLP Network with One Hidden Layer [73]

The output of the MLP network with m outputs can be calculated as:

$$Y_k(t) = \sum_{i=1}^{n} w_{jk}^2 F\left[\sum_{i=1}^{n} w_{ij}^1 v_i^o(t) + b_j^1\right] \qquad 1 \le K \le m$$

Where w_{ij}^1 and w_{jk}^2 are weight of the connection between input and hidden layer.

Random Forest

The random forest machine learner, is a meta-learner; meaning consisting of many individual learners (trees). The random forest uses multiple random trees classifications to votes on an overall classification for the given set of inputs. In general in each individual machine learner vote is given equal weight. In Breiman's later work, this algorithm was modified to perform both un-weighted and weighted voting. The forest chooses the individual classification that contains the most votes. Figure 17 is a visual representation of the un-weighted random forest algorithm.

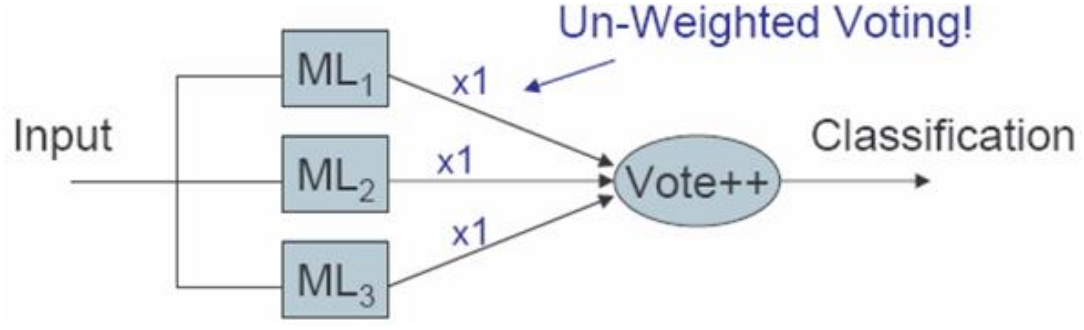


Figure 17 Meta Learners [69]

A random forest is made up of a number of decision trees. Each decision tree is made from a randomly selected subset of the training dataset using replacement. For building a decision tree, a random subset of available variables is used. The final result/outcome is chosen by the majority. Each decision tree in the random forest gives out its own vote for the result and the majority wins. In building a random forest, we can mention the number of decision trees we want in the forest. Each decision tree is built to its maximum size. There are various advantages of a random forest. Very little pre-processing of data is required. A random forest itself takes the most useful variables [70].

Bayes Net

A Bayesian network comprises of a probability distribution and directed acyclic graph, where arcs and nodes in a directed acyclic graph represent random variables and they have direct inter-relation between variables respectively and probability distributions is the set of local distribution for each node. A local distribution is described by a conditional probability table. Bayes Net basically used for the classification problem. In classification learning problem, a learner construct a classifier from a training set which is presented by attribute variable used collectively to predict the value of the class variable. The objective of learning a bayes net is to identify both structure learning and parameter learning. Structure learning corresponds to structure of network and parameter learning corresponds to conditional probability table [74].

Adaboost

Adaboost is one of the popular ensemble methods. Adaboost and its variant have been applied to large domains with great success; have outstanding theoretical structure, efficient prediction and quite simplicity. This algorithm can be described as an abstraction of boosting procedure. Boosting is a popular method for improving the efficiency of a given learning algorithm. It is effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rule of thumb. The basic role of an Adaboost algorithm is to maintain weight distribution of training set. At the beginning all the weights are equal but gradually weights of incorrect classifiers are improved after each round. So the weak learner is focused on strong learning algorithm. In this way weak learning classifiers are improved to strong learning classifiers [67]. Adaboost is short for adaptive boosting. It is a machine learning algorithm that can be used along with many other learning algorithms. This leads to an improvement in efficiency and performance. Adaboost is adaptive as it adapts to the error rates of the individual weak hypothesis. Also, Adaboost is a boosting algorithm as it can efficiently convert a weak learning algorithm into a strong learning algorithm. Adaboost calls a given weak algorithm repeatedly in a series of rounds. The important concept for an Adaboost algorithm is to maintain a distribution of weights over the training set. Initially all the weights are equal but on each round the weights of incorrect classified examples are increased so that a weak learner is forced to focus on the hard examples in the training set. This is how a weak learning algorithm is changed to a strong learning algorithm. Adaboost is less susceptible to an over fitting problem than most learning algorithms [80].

Decision Table

Decision tables are considered as classification models used prediction and are influenced by machine learning algorithms. A decision table consist of a hierarchical table in which every entry is broken down by the pair of additional attributes of higher level table to form another table. The structure of the decision looks similar to dimensional stacking. Visualization method allows a model depending upon many attributes that can be understood even by those unfamiliar with machine learning. Interaction of various form have been used to make this visualization more useful and efficient than other static designs [75].

Multinomial Naive Bayes

Naive bayes is a learning algorithm used to handle text classification. It is very efficient in computation and easy to implement. There are two popular models that have been used: Multivariate Bernoulli Event model and Multinomial Event model. Multinomial event model usually referred to as Multinomial Naive Bayes. It has been found to favourably with more specialised event model. It is a supervised learning technique, in which every new document is classified by assigning one or more class label from a fixed set of predefined classes [76]. In practice many application of MNB uses modification for correcting the strong assumptions [77].

5.2.4 Evaluation Measures of Performance

To predict model have used following evaluation measure:

Sensitivity: It is defined as the percentage of data points correctly measure. It measures the correctness of predicted model.

Specificity: It also measures the correctness of the predicted model.

Receiver Operating Characteristic (ROC) Analysis

ROC analysis is used to evaluate the performance of the output of the predicted models. It is a efficient method for model prediction. The ROC curve is a graph between sensitivity and 1-sensitivity. Where sensitivity corresponds to y coordinate and 1-sensitivity corresponds to x coordinate [81]. This work has been selected many cut-off points while plotting the ROC curves between 0 and 1 and then at each cut-off point determine sensitivity and specificity.ROC curve is used to obtain the optimal cut-off point that maximizes both sensitivity and specificity [81].

The study have used k-cross validation method for our study where value of k is 10.in this method data points is divided into k equal partitions among them one partition is used for testing and remaining k-1 partitions are used for training the model at a time [82

CHAPTER 6

RESULT ANALYSIS

Result Analysis

This section provides the final result of our study. The thesis validated web metrics. Initially this research work analysed the subset of the web metrics that are related to quality of web pages. For this purpose, there has been used statistical method called univariate logistic regression. To predict the model that provides highest accuracy, seven machine learning method has been used. The performance of each model is analysed using sensitivity, specificity, ROC and cut-off points.

6.1 Univariate LR Analysis Result

The study have analysed univariate logistic regression result to find how each independent metrics is associated significantly associated with the status(good or bad) of website which is taken as dependent variable. Table 4 presents the univariate analysis results. It shows the coefficient (B), standard error (S.E.), statistical significance (sig.), and odds ratio (exp (B)) for each metric [83]. The "sig" parameter describes whether each metric significantly predict the dependent variable. The metric is said to be significant if value of "sig" parameter of metric is nearly equals to 0.01 that is significance threshold.

Table 8 represents the significant values in bold. The coefficient (B) tells the strength of independent variable. The higher value of coefficient (B) shows the higher impact of the independent variable. The sign of coefficient represents whether the impact is positive is negative or positive. We can observe that total words in a page, HTML Lines, Paths and Statements are not significant therefore not taken for further analysis. Thus, in this way we can reduce the independent variable and take only significant metrics for analysis of website quality.

Metric name	**B**	**S.E.**	**Sig.**	**Exp(B)**
Total words in a page	0.000	.000	0.167	1.000
Words between body tag	.000	.000	**.016**	1.000
Words between title tag	-0.017	0.32	0.603	.983
Total links	-0.002	.001	0.**030**	1.002
No. of Internal links	0.127	0.075	0.090	1.135
Size of page in KB	0.004	0.002	**0.025**	1.004
Embbed text	0.000	.001	0.0904	1.000
Total no. of!'s	-0.002	0.002	0.294	0.994
Average Number of Lines	-0.145	0.42	**0.001**	0.865
HTML Lines	0.000	0.000	0.883	1.000
JavaScript Lines	-0.003	0.001	**0.030**	0.997
Blank Lines of Code	0.000	0.000	0.883	1.000
Source Lines of Code	-0.003	0.001	**0.030**	0.997
Lines with Comments	-0.010	.004	**0.008**	.990
Paths	0.000	.000	.523	1.000
Statements	-0.002	0.001	0.125	.998
Cyclomatic Complexity	-0.125	0.46	**0.007**	0.883
Nesting	-0.759	0.172	**0.000**	0.468
Graphic	-0.001	0.002	0.728	0.999
List count	-21.49	40192.970	1.000	0.000
Frame count	0.274	0.155	**0.078**	1.315
Table count	0.32	0.38	0.392	1.033

Table 8 Univariate Analysis

6.2 Model Evaluation Using the Roc Curve

The section summarises the result analysis. The study has been used seven machine learning algorithm for prediction. We used 10 cross- validation methods. Table 6 presents the result 10 cross validation of model predicted by Appling machine learning algorithm. Table 9 shows sensitivity, specificity, AUC and the cut-off point for the model prediction. Sensitivity corresponds to y-axis and 1- specificity corresponds the x-axis of the ROC curve. The cut-off point can be called as the point at which (1-specificity) is equal to sensitivity.

S.NO.	**Method used**	**sensitivity**	**specificity**	**AUC**	**Cut-off point**
1.	Adaboost	70.8%	70%	80.4	0.48
2.	Bagging	69.8%	70%	80.5	0.49
3.	Bayes Net	67.7%	69.5%	81.3	0.27
4.	MLP	64.4%	65%	66.8	0.64
5.	Random Forest	72.9%	72.6%	82	0.55
6.	Decision table	66.7%	63%	71.1	0.39
7.	Naive bayes multinomial	69.8%	69.9%	70.8	0.99
8.	Logistic regression	67.7	67.9	69.8	0.55

Table 9 Results of 10-Cross Validation

We can observe Random forest gives the best result as compare to other methods. Random forest has highest value of area under curve i.e. 82 and for Bayes Net AUC value is 81.3.The Adaboost and bagging gives almost similar result i.e. value of AUC are 80.4 and 80.5 respectively. Naive Bayes Multinomial and decision table show average result. AUC values for these methods are 70.8 and 71.1 respectively.

MLP shows poor results having AUC value is 66.8. Thus, this method is not considered as good method for prediction. We can also observe that the logistic regression [84] is poor as compare to machine learning methods. Thus, we can conclude from the discussion that the machine learning methods give better results as compared to the statistical methods. From amongst the machine learning methods under consideration, random forest and Bayes Net are the best predicted models.

Bagging

Bagging method is used to generate multiple version of a predictor and to aggregate this predictor. Bootstrap replicates the learning set that is used to form multiple version of predictor then after used it as new learning set. Bagging is also known as bootstrap aggregating that repeatedly samples from a data set according to uniform probability distribution [68].

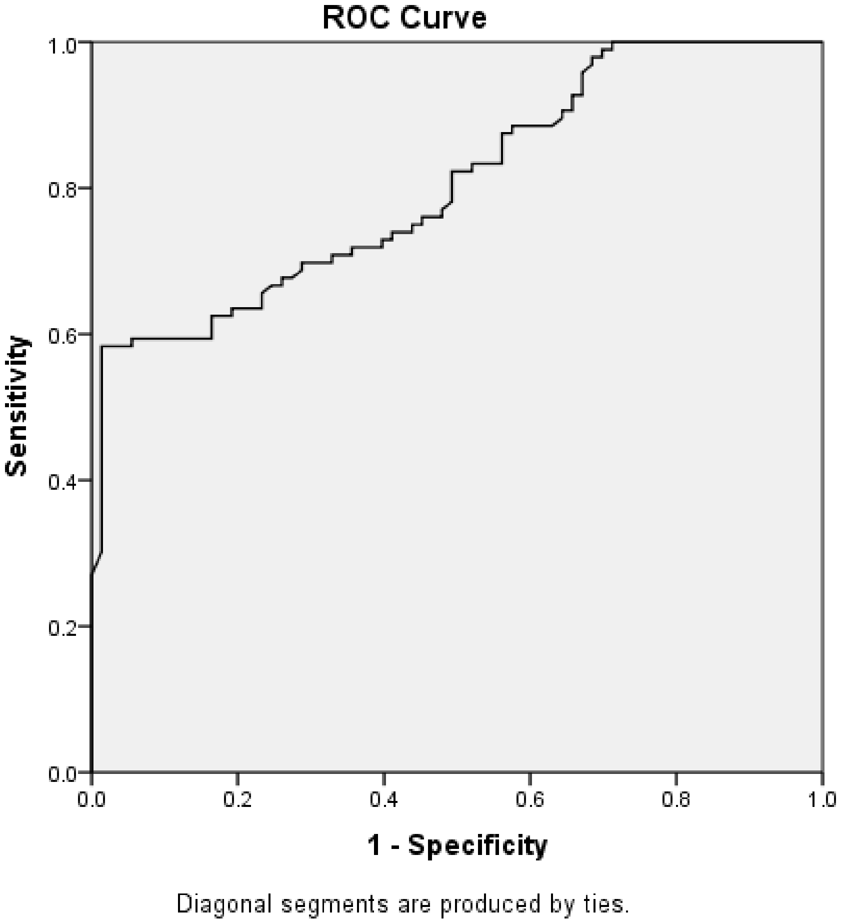


Figure18 ROC Curve for Bagging Algorithm

Multi Layer Perceptron

A MLP network is one of the most important neural network models and has been used for pattern classification. It is proved to be a powerful technique. This neural network can be trained to form arbitrary surfaces in input space. Mashor showed in [71] that the MLP network was highly nonlinear and ever modelling a linear model using the standard nonlinear network is never be the best solution.

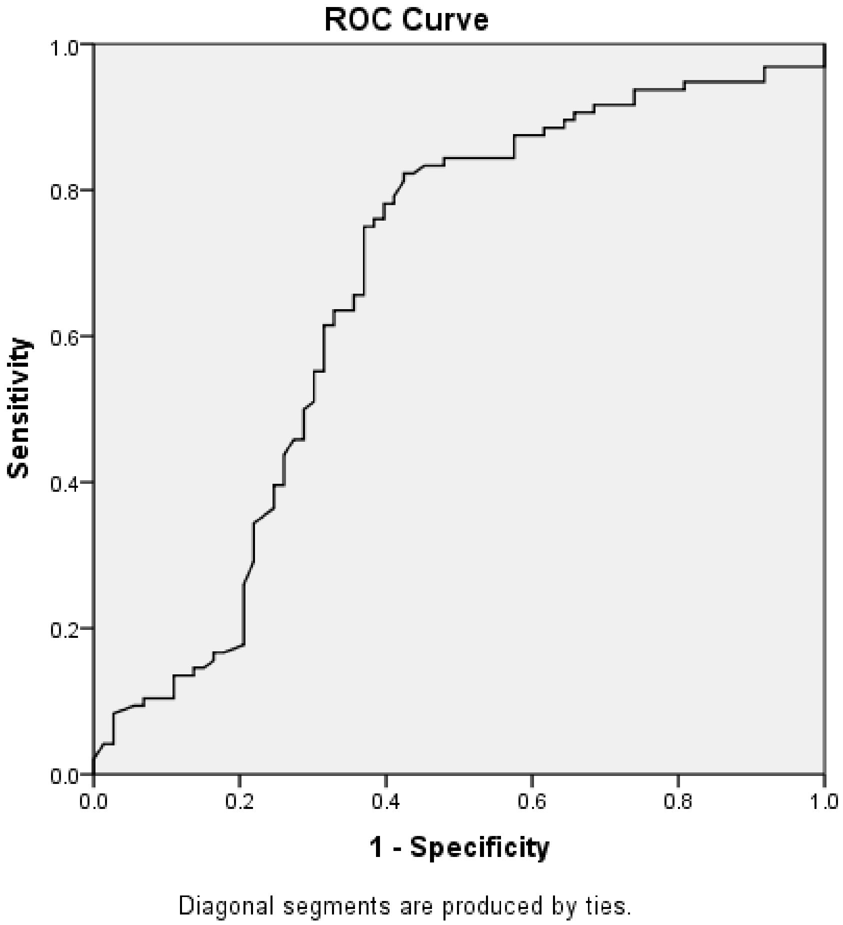


Figure19 ROC Curve for Multi Layer Perceptron Algorithm

Random Forest

The random forest machine learner, is a meta-learner; meaning consisting of many individual learners (trees). The random forest uses multiple random trees classifications to votes on an overall classification for the given set of inputs. In general in each individual machine learner vote is given equal weight. In Breiman’s later work, this algorithm was modified to perform both un-weighted and weighted voting. The forest chooses the individual classification that contains the most votes. Figure1. Below is a visual representation of the un-weighted random forest algorithm [69].

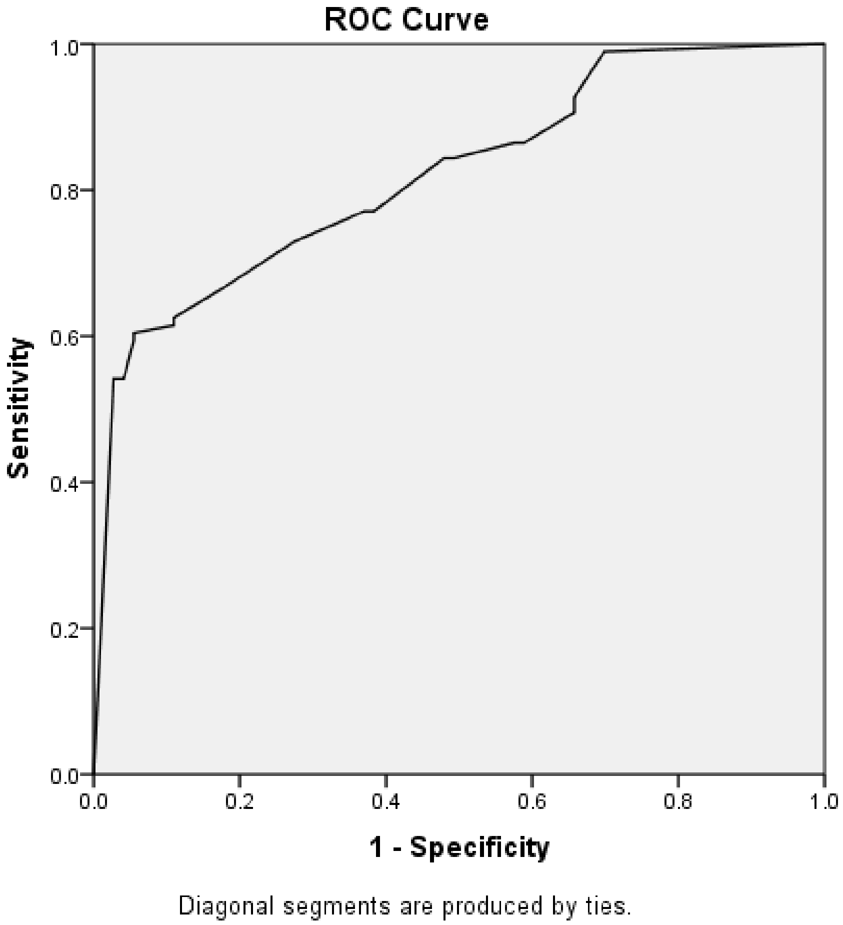


Figure 20 ROC Curve for Random Forest Algorithm

Bayes Net

A Bayesian network comprises of a probability distribution and directed acyclic graph, where arcs and nodes in a directed acyclic graph represent random variables and they have direct inter-relation between variables respectively and probability distributions is the set of local distribution for each node. A local distribution is described by a conditional probability table. Bayes Net basically used for the classification problem.

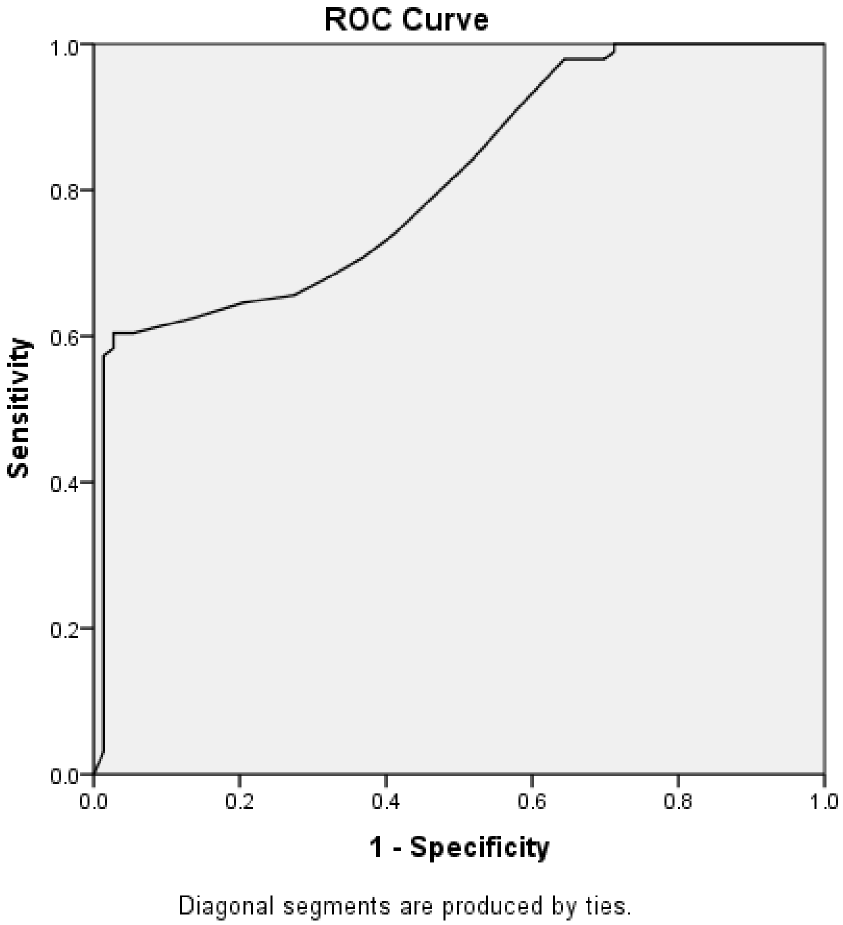


Figure 21 ROC Curve for Bayes Net Algorithm

Adaboost

Adaboost is one of the popular ensemble methods. Adaboost and its variant have been applied to large domains with great success; have outstanding theoretical structure, efficient prediction and quite simplicity. This algorithm can be described as an abstraction of boosting procedure. Boosting is a popular method for improving the efficiency of a given learning algorithm. It is effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rule of thumb.

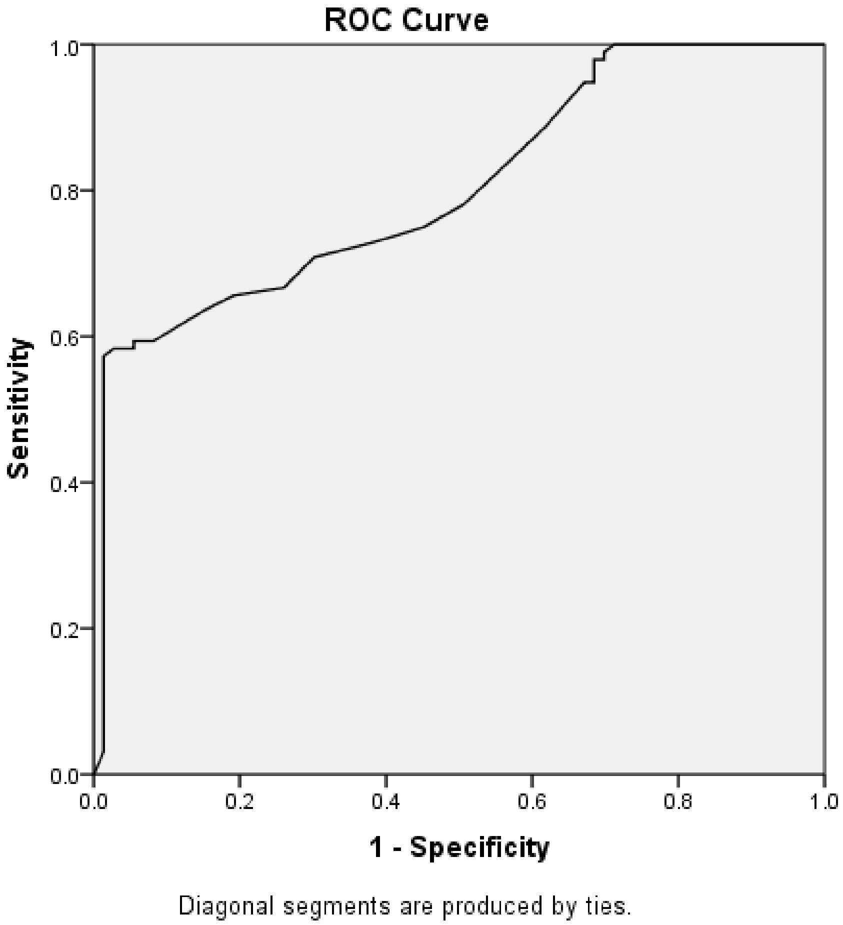


Figure 22 ROC Curve for Adaboost Algorithm

Decision Table

Decision tables are considered as classification models used prediction and are influenced by machine learning algorithms. A decision table consist of a hierarchical table in which every entry is broken down by the pair of additional attributes of higher level table to form another table. The structure of the decision looks similar to dimensional stacking. Visualization method allows a model depending upon many attributes that can be understood even by those unfamiliar with machine learning. Interaction of various form have been used to make this visualization more useful and efficient than other static designs [75].

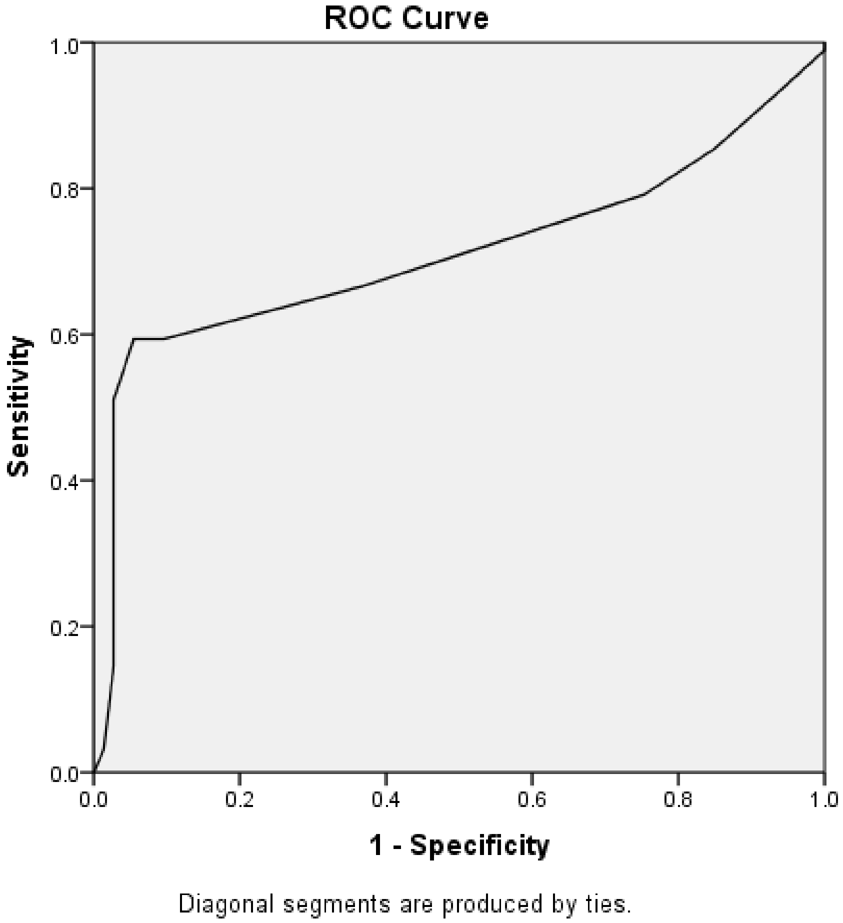


Figure 23 ROC Curve for Decision Table Algorithm

Multinomial Naive Bayes

Naive bayes is a learning algorithm used to handle text classification. It is very efficient in computation and easy to implement. There are two popular models that have been used: Multivariate Bernoulli Event model and Multinomial Event model. Multinomial event model usually referred to as Multinomial Naive Bayes. It has been found to favourably with more specialised event model. It is a supervised learning technique, in which every new document is classified by assigning one or more class label from a fixed set of predefined classes [76].

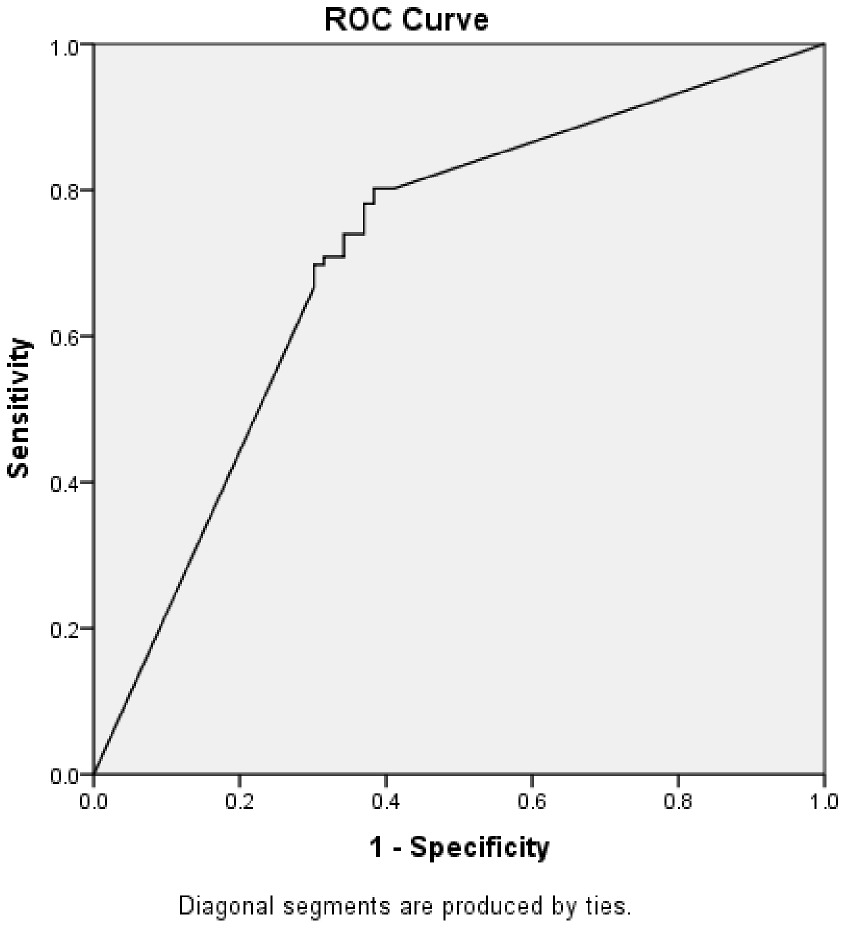


Figure24 ROC Curve for Multinomial Naive Bayes Algorithm

Conclusion

In chapter 4 proposed the V model for web reengineering provides effective and easy way of reconstruction and re-testing web application, it also provides flexibility and reusability that takes into consideration all the user and business standard. The proposed V model of reengineering process is proposed for designing the testing protocol for web application development. By introducing V model into web reengineering process, it increases the maintainability and the effectiveness of the website because V model led to better validation and verification and produce in accordance to web development life cycle. The structure of V model enables testing process starts from unit testing to acceptance testing. It will save the time for reconstructing or refactoring the web due to strong testing and validation.

The STAR paradigm which include situation, tools, application and restructuring that support the comprehension of existing web application during maintenance. The aim of the process is to reconstruct web application illustrating various aspects of web reengineering. STAR paradigm provides effective and easy way of reconstruction web application. By introducing STAR paradigm into web reengineering process, it increases maintainability and effectiveness because it led to better understanding of procedures and practices. It will save the time for reconstructing or restructuring the web application due to strong discerning and contrivance.

The objective of chapter 5 of this thesis is to evaluate the quality of web sites. The E-commerce is an emerging platform and we frequently deal with different web- based interactive applications like shopping and management applications etc. We can expect an increased interest in developing such applications with the wide acceptance of the same. Hence maintenance, quality and security inclusion for websites become very crucial. We validate empirically the relationship between quality and web metrics using statistical ad machine learning methods. This study has collected data from webby award web site. The webby award is a largest human rated website. This organisation consists of six judges who examined many web websites that has submitted for the awards based upon many criteria which are unknown to us. The study has taken all the award winning sites as good data points and other nominated sites bad data points for our research. The study considered all web metrics as independent variable and status (goodness and badness) of a website as a dependant variable. The research work has identified various web metrics for predicting quality of a website. To meet this objective we did univariate logistic regression. It was found that the metrics average number of lines, complexity, nesting, lines with comments, total links is more significant predictor of quality of website.

For model prediction there has been used seven machine learning methods which determine the accuracy in terms of specificity, sensitivity, AUC (area under curve) and cut-off points. Cut-off points maintain the balance between quality data points. Cut-off points are calculated using ROC curve. We observed that the random forest and Bayes Net gave the best result as compare to other model.

Future Work

There is a lot of scope for future research in this area. We can apply STAR paradigm in a practical work by applying these process on some live websites. We can do comparative analysis by modifying any website with different engineering model and analyse the actual time, cost efforts required to rebuild the web sites. The reverse engineering toolset can be considered to be one step in the direction of providing support for natural development for web based applications.

In future, we repeat this process on comparatively larger data set and we will explore the methods and tools. Also we can include more significant metrics that influence on web site quality. We have analysed only level 1 of a web page that is home page we can extend it to level 2 and level 3 in future. Also we can include some more parameter to evaluate the performance of model.

References

[1] J. Conallen, *"Modeling web application architectures with UML"*. *Communications of the Association for Computing Machinery* 1999. 42 (10): 63-70.

[2] S. Tilley, S. Huang "*Evaluating the reverse engineering capabilities of web tools for understanding site content and structure: a case study*". In *Proceedings of 23rd International Conference on Software Engineering* - 2001, IEEE Computer Society Press, Los Alamitos, CA, 2001; 514- 523.

[3] A.Ginige and S.Murugesan, "*Web engineering. An introduction*", IEEE Multimedia, 8(1):14-18, April-June 2001.

[4] L. Baresi, S. Morasca, and P. Paolini. Estimating the design effort of web applications. In Proceedings of the 9th International Software Metrics Symposium, pages 62–72. IEEE Computer Society Press, 2003.

[5] A. K. Mishra ,P. Bhatt, 'Influencing Factors in Outsourced Software Maintenance' , May 2006.

[6] Basili , Mills ' Understanding and Documenting Programs.' IEEE Transactions on Software Engineering SE-8,3(1982), pp. 270-283.

[7] E. Chikofsky and J.H.Cross, "Reverse Engineering and Design Recovery: A Taxonomy", *IEEE Software Engineering journal*, (Jan. 1990), pp 13-17.

[8] IEEE Std 1219-1998, In IEEE Standards Software Engineering, 1999 Edition, Volume Two, Process Standards, IEEE Press.

[9] Di Lucca GA, Fasolino AR, De Carlini U, Pace F, Tramontana P. Comprehending Web applications by a clustering based approach. *Proceedings 10th Workshop on Program Comprehension*. IEEE Computer Society Press: Los Alamitos CA,2002; 261–270.

[10] Anquetil N, Lethbridge TC. Experiments with clustering as a software remodularisation method. *Proceedings 6th Working Conference on Reverse Engineering*. IEEE Computer Society Press: Los Alamitos CA, 1999; 235–255.

[11] Biggerstaff TJ, Mitbander BG, Webster D. Program understanding and the concept assignment problem. *Communications of the ACM* 1993; 37(5):72–83.

[12] Distante, D. “Reengineering Legacy Applications and Web Transactions: An extended version of the UWA Transaction Design Model.” Ph.D. Dissertation, University of Lecce, Italy. June 2004.

[13] UWA (Ubiquitous Web Applications) Project, “Deliverable D3 Requirements Investigation for Bank121 pilot application”, http://www.uwaproject.org, 2001.

[14] UWA (Ubiquitous Web Applications) Project, “Deliverable D6: Requirements Elicitation: Model, Notation and Tool Architecture”, www.uwaproject.org, 2001.

[15] UWA (Ubiquitous Web Applications) Project, “Deliverable D7: Hypermedia and Operation design: model and tool architecture”, www.uwaproject.org, 2001.

[16] UWA (Ubiquitous Web Applications) Project, “Deliverable D8: Transaction design”, www.uwaproject.org, 2001.

[17] UWA (Ubiquitous Web Applications) Project, “Deliverable D9: Customization Design Model, Notation and Tool Architecture”, www.uwaproject.org, 2001.

[18] Xin-Hua Zhang ; Comput. & Inf. Eng. Coll., Hohai Univ., Nanjing, China ; Zhi-jian Wang, e-Business and Information System Security (EBISS), 2010 2nd International Conference on 22-23 may 2010.

[19] M.M. Moore, Representation Issues for Reengineering Interactive Systems, *ACM Computing Surveys Special issue: position statements on strategic directions in computing research*, Vol. 28, No. 4, Dec 1996, article # 199, ACM Press, New York, NY, USA.

[20] M.M. Moore and S. Rugaber, Using Knowledge Representation to Understand Interactive Systems, in *Proc. of the Fifth International Workshop on Program Comprehension IWPC’97* (Dearborn, 28-30 May 1997), IEEE Computer Society Press, Los Alamitos, 1997

[21] G. Mori, F. Paternò, C. Santoro, Tool support for designing nomadic applications, *Proc. of the 2003 international conference on Intelligent user interfaces, Jan 2003*, (Miami, USA), ACM Press, New York, USA, pp141-148

[22]L.Paganelli, F.Paterno, Automatic reconstruction of the underlying interaction design of web applications, in *Proc. Of the 14th international conference on Software engineering and knowledge engineering,* (July 2002, Ischia, Italy), ACM Press, New York, USA, pp 439 – 445.

[23] G. Calvary, J. Coutaz, D. Thevenin, Q. Limbourg, N. Souchon, L. Bouillon, J. Vanderdonckt, Plasticity of User Interfaces: A Revised Reference Framework, in *Proc. of 1st International Workshop on Task Model and Diagrams for user interface design* Tamodia'2002 (Bucharest, 18-19 Jul 2002), INFOREC Publishing House Bucharest, Romania,2002.

[24]. B. Laurent, L. Quentin, V. Jean and M. Benjamin, "Reverse Engineering of Web Pages Based on Derivation and Transformation" , 27 August 2005.

[25]. G. K.Tyagi and D. P. Ballou, "Examinig Data Quality", *communication ACM*, 41 (2), Feb. 1998.

[26] Hammer, Michael (1996), "Beyond Reengineering: How the Process-Centered Organization is Changing Our Work and Our Lives", New York, NY: HarperCollins Publishers, Inc.

[27]. Stanjarzabek, "Strategic reengineering of software: life cycle approach", Department of Information Systems and Computer Science National University of Singapore.

[28]. S. Jarzabek, "Domain Model-Driven Software Reengineering and Maintenance", *Journal of Systems and Software,* January 1993, pp. 37-51.

[29]. M. Giordano, G. Polese, G. Scanniello, and G. *Tortora, "*Visual Modelling of Role-Based Security Policies in Distributed Multimedia Applications". *In Proc. of IEEE6th International Symposium on Multimedia Software Engineering*, Miami, FL, USA, IEEE CS Press, 2004, pp.:138 – 141.

[30]. D. Distante, T. Parveen, and S. Tilley, “Towards a Technique for Reverse Engineering Web Transactions from a User’s Perspective”, In *Proceedings of the 12th IEEE International Workshop on Program Comprehension* (IWPC 2004: June 24-26, 2004; Bari,Italy). Los Alamitos, CA: IEEE CS Press, 2004.

[31] S. Tilley, D. Distante, and S. Huang, “Design Recovery of Web Application Transactions”, Submitted to the 11thIEEE Working Conference on Reverse Engineering (WCRE 2004: Nov. 9-12, Delft, the Netherlands). June 2004.

[32] P. Szekely, P. Luo, and R. Neches, “Beyond Interface Builders: Model-Based Interface Tools”, *Proc. of ACM Conf. on Human Aspects in Computing Systems InterCHI’93*, ACM Press, New York, 1993, pp. 383-390.

[33] J. Vanderdoncktand P. Berquin, “Towards a Very Large Model-based Approach for User Interface Development”,*Proc. of 1st Int. Workshop on User Interfaces to Data Intensive Systems UIDIS’99*, IEEE Computer Society Press, LosAlamitos, 1999, pp. 76-85.

[34] Offutt J, “Quality attributes of Web software applications”, *IEEE Software* 2002; **19**(2):25–32.

[35] M. Moore, “Representation Issues for Reengineering Interactive Systems”, *ACM Computing Surveys Special issue: position statements on strategic directions in computing research*, Vol. 28, No. 4, Dec 1996, article # 199, ACM Press, New York, NY, USA.

[36] M. Moore and S. Rugaber, “Using Knowledge Representation to Understand Interactive Systems”, in *Proc. of the Fifth International Workshop on Program Comprehension IWPC’97* (Dearborn, 28-30 May 1997), IEEE Computer Society Press, Los Alamitos, 1997

[37] G. Mori, F. Paternò, C. Santoro, “Tool support for designing nomadic applications”, *Proc. of the 2003 international conference on Intelligent user interfaces, Jan 2003*, (Miami, USA), ACM Press, New York, USA, pp141-148.

[38] F. Ricca, P. Tonella, I.D. Baxter, “Restructuring Web applications via Transformation Rules”, *Proc. Of IEEE Workshop on Source Code Analysis and Manipulation SCAM.2001* (Florence, 5-9 Nov 2001), IEEE Computer Soc. Press, Los Alamitos, 2001, pp. 150-160.

[39] Vanderdonckt, J., Bouillon, L. and Souchon, N. (2001), "Flexible reverse Engineering of Web Pages with VAQUISTA", Proc 8th Working Conference on *Reverse Engineering, WCRE'01, IEEE*, pp241-248.

[40] L. Paganelli, F. Paterno, "Automatic reconstruction of the underlying interaction design of web applications", in *Pro .Of the 14th international conference on Software engineering and knowledge engineering,* (July 2002, Ischia, Italy), ACM Press, New York, USA, pp 439 – 445.

[41] Di Lucca, G.A., Fasolino, A.R., Pace F., Tramontana, P. and De Carlini, U. (2002a),"*WARE:A Tool for The Reverse Engineering of Web Applications"*, Proc. 6th European Conference onSoftware Maintenance and Rengineering (CSMR'02), pp241-250.

[42] Di Lucca, G.A., Fasolino, A.R. and Tramontana, P. (2004), "*Reverse Engineering Web Applications: The WARE Approach"*, Journal of Software Maintenance and Evolution,Research and Practice, Vol 16, pp71-101.

[43] G. Mori, F. Paternò, C. Santoro, *Tool support for designing nomadic applications*, Proc. of the 2003 international conference on Intelligent user interfaces, Jan 2003, (Miami, USA), ACM Press, New York, USA, pp141-148.

[44]. Draheim, D., Fehr, E. and Weber, G. (2003), "JSPick – A Server Pages Design Recovery", Proc7th IEEE European Conference on Software Maintenance and Reengineering, LNCS, pp230-236.

[45] L.Paganelli, F.Paterno, Automatic reconstruction of the underlying interaction design of web applications, in *Proc. Of the 14th international conference on Software engineering and knowledge engineering,* (July 2002, Ischia, Italy), ACM Press, New York, USA, pp 439 – 445.

[46] M. Abrams, C. Phanouriou, A. Batongbacal and J. Shuster, UIML: *An Appliance-Independent XML User Interface Language,* in Proc. of 8th World Wide Web Conference WWW.8 (Toronto, 11-14 May 1999), Computer Networks, Vol. 31, No. 11-16, pp. 1695-1708.

[47] Jovanovic, N. ; Secure Syst. Lab., Tech. Univ. of Vienna ; Kruegel, C. ; Kirda, E., "a static analysis tool for detecting Web application vulnerabilities", Security and Privacy, 2006 IEEE.

[48] Xin-Hua Zhang ; Comput. & Inf. Eng. Coll., Hohai Univ., Nanjing, China ; Zhi-jian Wang, e-Business and Information System Security (EBISS), 2010 2nd International Conference on 22-23 may 2010.

[49] J. Conallen. *Building Web Applications with UML.* Addison- Wesley Publishing Company, Reading, MA, 1999.

[50] J. Conallen. Modeling web application architectures with uml. *Communications of the Association for Computing Machinery,* 42(10), October 1999.

[51] J. Conallen. Modeling web applications with uml. White paper, Conallen Inc., http://www.conallen.com/whitepapers/ webapps/ModelingWebApplications.htm, March 1999.

[52] N. Anquetil, T.C. Lethbridge, "*Experiments with clustering as a software remodularization method"*. In *Proceedings of 6th Working Conference on Reverse Engineering* - 1999*.* IEEE Computer Society Press: Los Alamitos, CA, 1999; 235-255.

[53] T.A. Wiggerts, "*Using clustering algorithms in legacy systems remodularization*", *4th Working Conference on Reverse Engineering,* IEEE CS Press, Los Alamitos, CA*,* 1997, pp. 33-43.

[54] R.W. Schwanke, "*An intelligent tool for Reengineering Software Modularity*", *Proc. of 13th International Conference on Software Engineering*, IEEE CS Press, Los Alamitos, CA, 1991*,* pp. 83-92.

[55] V. Basili, D. Hutchens, "*System structure analysis: clustering with data bindings*", *IEEE Transactions on Software Engineering*, 11 (8), 1985, pp. 749-757.

[56] H.A. Müller, K. Klashinsky, "*Rigi - A system for programming in the large*". In *Proceedings of International Conference on Software Engineering* - 1988*,* IEEE Computer Society Press, Los Alamitos, CA, 1988; 80-86.

[57] K. Wong, S. Tilley, H.A. Müller, M.A. D. Storey, "*Programmable Reverse Engineering*", *International Journal of Software Engineering and Knowledge Engineering*, 4 (4), Dec. 1994, pp.501-520.

[58] V. Tzerpos, R.C. Holt, "*On the stability of software clustering algorithms*", *8th International Workshop on Program Comprehension,* IEEE CS Press, Los Alamitos, CA, 2000, pp. 211-220.

[59] V. Tzerpos, R.C. Holt, "*ACDC: an algorithm for comprehension-driven clustering*", *7th Working Conference on Reverse Engineering,* IEEE CS Press, Los Alamitos, CA, 2000, pp. 258-267.

[60] N. Anquetil, T.C. Lethbridge, "*Experiments with clustering as a software remodularization method"*. In *Proceedings of 6th Working Conference on Reverse Engineering* - 1999. IEEE Computer Society Press: Los Alamitos, CA, 1999; 235-255.

[61] George W. Furans, "Effective view navigation", *in proceedings of ACM CHI 97 conference on human factors in computing systems*, volume 1 of PAPERS: information structures, pp. 367-374, 1997.

[62] Lincoln D. Stein, "The rating game", http://stein .cshl.org/lstein/rater/,1997.

[63] Kevin Larson and Mary Czerwinski., "Web page design: Implications of memory, structure and scent for information retrieval", *In proceedings of ACM CHI 98 Conference on human Factors in Computing Systems,* volume 1 of Web Page Design , pp. 25-32, 1998.

[64] Yanlong Zhang, Hong Zhu and Sue Greenwood, "Website Complexity Metrics for Measuring Navigability", Proceedings of the fourth conference on quality software (QSIC'04), 0-7695-2207-6/04, IEEE.

[65] http://WWW.webby awards.com/

[66] Weka. Available: http://www.cs.waikato.ac.nz/ml/weka/

[67] Yoav Freund and Robert E. Schapire" *A Short Introduction to Boosting", Journal of Japanese Society for Artificial Intelligence,14(5):771-780, September, 1999.*

[68] L.Breiman, "*Bagging predictors*," *Machine Learning*, Vol.24, 1996, pp.123-140.

[69] Y. Freund and R.E. Schapire, "*A Short Introduction to Boosting*," *Journal of Japanese Society for Artificial Intelligence,* Vol.14, No.5, 1999, pp.771-780.

[70] White, Mark **ECE591Q-Machine Learning – Lecture slides**, Fall 2005.

[71] Mashor M.Y, “Hybrid Multilayer Perceptron Networks”, *International Journal of Systems Sciences*, 2000, Vol 31, No 16, pp. 771-785.

[72] Cybenko G, “Approximations by Superposition of a Sigmoidal Function”, *Mathematics of Control, Signals and Systems 2*, 1989, Vol 2, pp. 303-314.

[73] Funashashi K, “On The Approximation Realization on Continuous Mappings by Neural Networks”, *Neural Network*, 1989, Vol 2, pp. 182-92.

[74] Jiang Su jiang.su@unb.ca,Harry Zhang” Full Bayesian Network Classifier” Faculty of Computer Science, University of New Brunswick, Canada, Appearing in Proceedings of the 23 rd International Con- ference on Machine Learning, Pittsburgh, PA, 2006.

[75] Barry G. Becker, INFOVIS '98 Proceedings of the 1998 IEEE Symposium on Information Visualization Pages 102-105 IEEE Computer Society Washington, DC, USA ©1998.

[76] Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes

for text categorization revisited. In: Proceedings of the 17th Australian joint conference

on Advances in Artificial Intelligence. AI'04, Berlin, Heidelberg, Springer-

Verlag (2004) 488-499.

[77] Schneider, K.M.: Techniques for improving the performance of naive bayes for text classification. In: In Proceedings of CICLing 2005. (2005) 682-693.

[78] R. Malhotra and A.Jain, “*Software Effort Prediction using Statistical and Machine Learning Me thod*,” *International Journal of Advanced Computer Science and Applications* , Vol.2, No.1, 2011.

[79] Guohua Liang, Xingquan Zhu, and Chengqi Zhang” An Empirical Study of Bagging Predictors for Different Learning Algorithms” Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence Sydney, NSW 2007, Australia.

[80] Y. Freund and R.E. Schapire, “*A Short Introduction to Boosting*,” *Journal of Japanese Society for Artificial Intelligence,* Vol.14, No.5, 1999, pp.771-780.

[81] Izzat Alsmadi, Ahmad T. Al-Taani, and Nahed Abu Zaid Web Structural Metrics Evaluation 978-0-7695-4160-0/10 $26.00 © 2010 IEEE DOI 10.1109/DeSE.2010.43

[82] M.Stone, "*Cross-validatory choice and assessment of statistical predictions*," *Journal Royal Stat. Soc.*, Vol.36, 1974, pp.111-147.

[83] Y. Singh, A. Kaur, and R. Malhotra, "Empirical vlidation of object-oriented metrics for predicting fault proneness models," Software Quality Journal, Vol.18,No.1, 2010,pp.3-35.

[84]Goran Mauša, Tihana Galinac Grbac and Bojana Dalbelo Bašić "Multivariate Logistic Regression Prediction of Fault-Proneness in Software Modules" MIPRO 2012, May 21-25,2012, Opatija, Croatia

[85] Dr. Linda H. *"Software Reengineering*" Rosenberg Engineering Section head Software Assurance Technology Center Unisys Federal Systems 301-286-0087

[86]Izzat Alsmadi, Ahmad T. Al-Taani, and Nahed Abu Zaid Web Structural Metrics Evaluation 978-0-7695-4160-0/10 $26.00 © 2010 IEEE DOI 10.1109/DeSE.2010.43