International Journal of Software Engineering & Systems

**ENRICHED
PUBLICATIONS**

International Journal of Software Engineering & Systems

Aims and Scope

Software Engineering has become very important with the ever-increasing demands of the software development to serve the millions of applications across various disciplines. For large software projects, innovative software development approaches are vital importance. In order to gain higher software standards and efficiency, software process adaptation must be derived from social behavior, planning, strategy, intelligent computing, etc., based on various factors. International journals of software engineering address the state of the art of all aspects of software engineering, highlighting the all tools and techniques for the software development process. The journals aims to facilitate and support research related to software engineering technology and the applications. International journals of software engineering welcomes the original research paper, review papers, experimental investigation , surveys and notes in all areas relating to software engineering and its applications. The following list of sample-topics its by no mean to be understood as restricting contributions to the topics mentioned:

Ø Aspect-oriented software development for secure software

Ø Dependable systems

Ø Experience related to secure software system

Ø Global security system

Ø Maintenance and evolution of security properties

Ø Metrics and measurement of security properties

Ø Process of building secure software

International Journal of
Software Engineering & Systems

(Volume No. 12, Issue No. 1, Jan - Apr 2024)

Contents

Comparative Analysis on Scope of Web Mining Techniques and Application

[1]Md. Barique Quamar & [2]Mohd. Arif

1. Al Falah School of Engineering Dhauj Faridabad e-mail: barik011@gmail.com
2. Al Falah School of Engineering Dhauj Faridabad e-mail: mohd.arif.jmi@gmail.com

A B S T R A C T

Data mining is a process of pulling out important information, identifying trends as well as analysing relationships from vast amounts of data. The Paper compares three different techniques, which are widely used in data mining, the task of converting unstructured data into structured and meaningful patterns. The techniques will be applied in four unique datasets that relate to the field of sports, health, law and entertainment respectively. The paper aims at exploring the merits and demerits of various data mining techniques and carry out a comparative analysis to explore which one of them works best in different circumstances. The research seeks to achieve this goal by processing those techniques on unique datasets to gain an understanding of the practical use.

Keywords: ***Data Mining, Comparative Analysis, Web Mining Techniques, Unstructured Data***

INTRODUCTION TO DATA MINING

Data mining is the process of digging in large, often unstructured, data and identifying potentially useful correlations and patterns (Han and Kamber, 2011). Data mining techniques are essentially an extension to the more traditional statistical approaches in the sense that it employs techniques from various disciplines such as machine learning, artificial intelligence, numerical analysis, neural networks, and the like (Rajaraman & Ullman, 2011).

Importance of Data Mining

In the past few decades, data has become the new oil. It is important, therefore, for the organizations to understand the importance of the information lying in their databases and extract useful patterns from it. Data mining is equally important for the analysts and scientists to understand the patterns in the data and gain some insightful analysis to obtain the analytics edge (Pal, 2007). Almost all the organizations use data mining in one way or the other. A large number of those use it in all the stages of their development such as customer acquiring, revenue growth, customer and employee retention, and the like (Han &Kimber, 2011). Data mining helps companies by letting them understand the customer preferences and make the business decisions accordingly. An important term used in this regard in the context of data mining is "profiling". Profiling is the process of determining the traits and characteristics of ideal customers who helped the company achieve a particular level of success. By understanding the characteristics of these customers, the company can focus on those customers who did not bring that particular level of success (Berry &Linoff, 2000). Profiling is also of significant importance when it comes to reducing churn (the act of retaining the inactive customers who are likely to leave).

Applications of Data Mining

Data mining finds applications in a wide array of fields. Contradictory to what the term suggests, data mining is not just the process of exploring data from different sources. Instead, the process extends to making sense of that data and finding useful patterns in it, which may be used for some important purposes. For instance, cellular companies receive the data from various sources (such as customer feedback, call history, social media feedback, and the like) and use that data to explore various meaningful trends and patterns (Pal, 2007). An example is when the companies detect churn based on the engagement level of the customers and take steps accordingly to prevent them from leaving. Data mining experts are hired by companies all over the world to gain a competitive edge. Below are some of the fields in which the use of data mining is most common:

. Data Mining in Education

. Data Mining in Business

. Data Mining in Science

. Data Mining in Fraud Detection

. Data Mining in Health

. Data Mining in Geography

. Data Mining in Sport

. Other Applications of Data Mining

Data Mining and Statistics

Statistics is an integral component of a rather more traditional version of data mining (Rajaraman& Ullman, 2011). The statistical approach towards data mining aims to discover the structure in data and interpret it using a wide variety of visualization techniques (Pal, 2007). This branch of data mining uses descriptive statistical techniques to determine things like mean values, standard deviation, standard error, percentages and correlation values between the variables. The different techniques used are:

Cluster Analysis : This technique organizes the data so that clusters of homogenous members can be formed.

Correlation Analysis : The purpose of correlation analysis is to determine the link between two variables, also known as correlation. The value of correlation lies between -1 and 1. A value of -1 indicates highly negative relationship, 1 indicates highly positive relationship whereas 0 indicates no relationship.

Factor Analysis: This kind of analysis is important when one needs to find out the cause(s) of a particular type of correlation. Another purpose of this analysis is to reduce the number of variables and thus, reduce the amount of data.

Linear Regression: Regression analysis establishes a link between dependent and independent variable so that one can be determined when the other is known. In regression analysis, there is one dependent variable, among one or more independent variables. No matter the statistical relation, linear regression analysis gives an equation of a straight line (best fitting line) which has minimum squared distance from each data point (also referred to as R2).

Logistic Regression: Similar to the linear regression, logistic regression also gives a best fitting line but unlike linear regression, it uses maximum likelihood method to derive the best fitting line.

WHAT IS DATA, INFORMATION & DATA TYPES

Data

According to S.P Bora (2001), "Data are any facts, numbers, or text that can be processed by a computer". Also according to the Business Dictionary, Data is any information in raw or unorganized form, which includes alphabets, numbers or symbols. It could signify conditions, ideas or objects. Another definition from the same source says that; Data are symbols or signals that are stored and processed by a computer device later used as an output for information [businessdictionary.com, 2014].Data however is limitless and present in various forms around the entire world. An example of Data includes; a man is 20 years old, African, speaks Latin, red hair and 6ft tall. All these are 'data'.

Information

Information is a processed, structured or systematized form of data to render it useful for a specific purpose. Data all by itself is useless without being interpreted. The art of making meaning of data for a useful purpose is called information. However, data is more of a computer language, while information is the translation of this language [diffen.com, 2014]. Data can be captured and distributed as information. An example of 'information' include; "the biography of an African man sent to the press". The biography is sent as information, irrespective of where the information goes, the data does not change, the characteristics of the African man remains intact and constant.

Data & Information Types

Data types can vary from simple numerical measurement or text-based documents to more complex information like hypertext documents and spatial data. The following is a non-exclusive list of information collected in digital forms and flat files in databases. [Zaiane, 1999]

- Operational or transactional data: Examples of such are business transactions, Generated purchases, exchanges, sales, cost, inventory, payroll, accounting, stocks and banking. Barcodes can be used to store millions of daily transactions representing terabytes of data. To survive in a highly competitive world, decision making using just-in-time information remains the key to solving critical issues in the businesses struggling to survive.

- Non-operational Data such as forecast data, macro-economic data, and games (sports) etc. [S.P Bora, 2001].
- Meta data: this refers to data about data itself such as logical database design or data dictionary definitions [S.P Bora, 2001].
- Others include; sequence data, images and videos, mixtures of fata, time-series data, etc.

WHAT IS KNOWLEDGE & KNOWLEDGE BASE TYPES

Knowledge

As the name implies, 'Knowledge' is derived from the word to 'Know' or "know how". Knowledge is a more meaningful, richer and deeper form of information. Collins English Dictionary (2014) defines knowledge as the "facts, feelings or experiences known by a person or a group" [InSTEP, 2014].Knowledge involves awareness and understanding of information achieved from identifying consequences, making connections and comparison. Other terms associated to knowledge are "applied information", "information with judgment" or "capacity for effective action" (National Electronic Library for Health: Knowledge Management Glossary). When information is shaped, organized and embedded in a form whereby it conveys meaning that has a purpose and leads to an understanding of something about the world, it becomes 'Information' (Stoll, Fink, and Earl, 2003, page 6).

Knowledge Base Types

"Knowledge base" can be explained as a shared body of knowledge from which members of a similar profession can draw information or learn from. This can be an individual as well as a shared professional knowledge. According to SearchCRM (2014), we learnt that a knowledge base is not necessarily a static collection of information but a dynamic resource with capacity to learn as an expert system (AI). There are various types of knowledge based system types such as [Kendal & Green, 2007]

Expert Systems: They simulate decision-making abilities of the human brain in experts. They contain algorithms, which do not change but also has the ability to learn from experience.

Neural Networks: They model the brain at a biological level by adapting to learn through experience, which can be used to predict trends in the future.

Case-based reasoning: They model the human ability to reason through analogy. Mostly used in legal cases.

Genetic algorithms: It is a method of generating solutions to complex problems

Intelligent agents: This is a software program where goals are specified and the software gets to make decisions on its own. It is built to learn, adapt and make complex decisions on behalf of users

Data mining: Another word is *data archaeology* or *data dredging*. It is used to describe knowledge discovery by recognizing past relationships in data.

Intelligent monitoring systems: Similar systems like educational learning and tutoring systems are able to continually teach students with various levels of explanations/examples/domain materials with respect to various students learning.

[Kendal & Green, 2007]

Knowledge Discovery From Data (KDD)

KDD or Knowledge discovery from data is another synonym associated with the term 'data mining'. Data Mining is by all means a part of the knowledge discovery process. The process to knowledge discovery is an iterative systematic process, which can be described as follows [Zaiane, 1999]

1. **Data Cleaning**: This is also called data cleansing. Noise data and irrelevant data are removed from the pile up of data in this phase

2. **Data Integration**: In this phase, multiple sources of data are combined in a common source

3. **Data Selection**: Here, only data related to the analysis being made are retrieved from the pool of data

4. **Data Transformation**: This is also called data consolidation. In this phase, the selected data is consolidated into groups appropriate for mining by performing aggregation operations

5. **Data Mining**: A phase whereby clever or intelligent methods are used to extract trends or patterns of data that are potentially useful

6. **Pattern Evaluation**: Here, data mined according to patterns representing knowledge are identified based on given measures

7. **Knowledge Representation**: In this final stage, knowledge representation and visualization techniques are used to present mined knowledge to users. These techniques are useful to interpret the data mining results.

[Osmar R. Zaiane, 1999]

The diagram as shown in fig 4, depicts the waterfall or iterative process in knowledge discovery of data.

Pre-Modeling Data Mining Processes

To extract useful information from the data set, the data is sent to the modeller; hence, it is important to undergo few processes, which are listed below:

Business Understanding

The first and foremost step in the data mining process as described by Liu (2006) is to understand the project objectives and requirements. This preliminary knowledge helps in formulating the actual data-mining problem and a plan that can be followed to achieve the project objectives. The data analyst needs to precisely understand the requirements of the client as to what he actually wants to achieve with the data. The goal of the analyst is to reveal different crucial factors that can possibly affect the outcome (Hornick, Marcade and Venkayala, 2010). Ignoring the importance of the pre-analysis can result in the right results for a wrong problem (Liu, 2006). The next step involved in the business-understanding phase is to assess the situation (Windt, 2012). It involves collecting all the facts regarding the available resources, constraints and opportunities, assumptions and other important factors that need to be considered for defining the data-mining goal.

Data Understanding

After an overall picture of the business has been developed and the basics of the business are well established, the next step as discussed by Liu (2006) is to gain an understanding of the data that will be used in the data mining process. This involves gaining an understanding of all the variables in the data and extracting their meaning in the context of the business (Haynes, 2006). The data that needs to be collected is generally listed in the project resources. Data loading can be done if required for gaining the necessary understanding of the data. Data loading means applying the data to a physical instrument or device just for the sake of developing its understanding. Once the data is collected, it is then examined for its properties and possible results in this regard are reported (Olson and Delen, 2008).

Data Preparation

After understanding the business needs and the data, the last step is to actually go on and prepare the data that would be fed into the data-modelling tool (Liu, 2006). The data preparation phase of data mining essentially covers all the activities that are needed to prepare the final set of data for modelling. In order to increase the reliability level, data preparation is repeated multiple times in a random manner. Weiss (1998) argues that a set of actions should be taken before preparing the data. These include filtering the redundant data, decreasing the excessive data volume to make the data mining more manageable, adding usable data types and considering the quality and technical constraints (Yan, 2009).

Summary

Before the data is analyzed and decisions are reach, it is critically important to undergo a few pre-modeling processes. These include, 1) developing an understanding of the business and project requirement, 2) developing understanding of the data under consideration, and 3) preparing the data for analysis by removing missing values and biases, if any. The modeling techniques in this chapter are linear regression, logistic regression, classification and regression trees (CART) and clustering. The final portion of this chapter aimed at discussing the recent literature on the comparative analysis of these modelling techniques.

DATA ANALYSIS

Comparative Analysis of Data Mining Techniques

Linear regression provides effective solution to the problems whose outcomes are categorical in nature. Linear regression finds extensive applications in sports analytics where the goal often is to find solid figures to either help the team win a competition or determine the ideal attributes in players that can help the team win matches. However, in cases where the outcome is of binary nature, linear regression does not prove to be the most effective technique.

LINEAR REGRESSION	LOGISTIC REGRESSION
Linear regression is used to establish a relationship between dependent and independent variables and determine the impact on the dependent variable (also categorized as outcome) in case one of more of the dependent variables change.	Logistic regression is used to determine probability of an event which is given in a binary format, 0 and 1, albeit also used to build a relationship between dependent and independent variables.
Linear regression gives a continuous value of dependent variable (or the outcome)	Logistic regression gives the probability of a categorical dependent variable (or the outcome)
Linear regression is generally solved by minimizing the value of least R-squared (or by finding the best fit line. Thus, there is a greater penalty for large errors.	The logistic regression function does not have a large penalty for larger value of errors. Instead, the logistic response function causes the large errors to be penalized to an asymptotically constant.
When plotted on a graph, linear regression returns a straight line.	The logistic response is visualized graphically by plotting true positive rate versus the false positive rate. This returns a curved line.
In linear regression, the measure of association between the desired outcome (dependent variable) and the predictor variable (independent variables) is given by a multiplicative factor.	In logistic regression, the measure of association between the desired outcome (dependent variable) and the predictor variable (independent variables) is given Odds Ratio.

IC RGER

In such cases where outcome is either 0 or 1 (such as whether or not the decision is reversed, or whether or not patient develops a certain disease) logistic regression and CART (Classification and regression trees) are the two most relevant techniques. Logistic regression determines the odds of a particular event happening and calculates the probability if the outcome will be 0 or 1. Often this is what is desired as in health care, where the goal is to predict beforehand if a patient is at risk of getting a disease. In such situations, logistic regression and CART can both prove helpful. The researcher analyzed that although logistic regression is somewhat more accurate than CART, it is less interpretable and often it proves very hard to reach quick decisions. In such scenarios, CART turns out to be a very useful technique that makes decisions by following the splits of the tree. An added advantage that CART provides over logistic regression is that it can be used not only in situations where the response variable consists of only two variables but also when the response variable is of continuous nature. The concept is illustrated in the figure below:
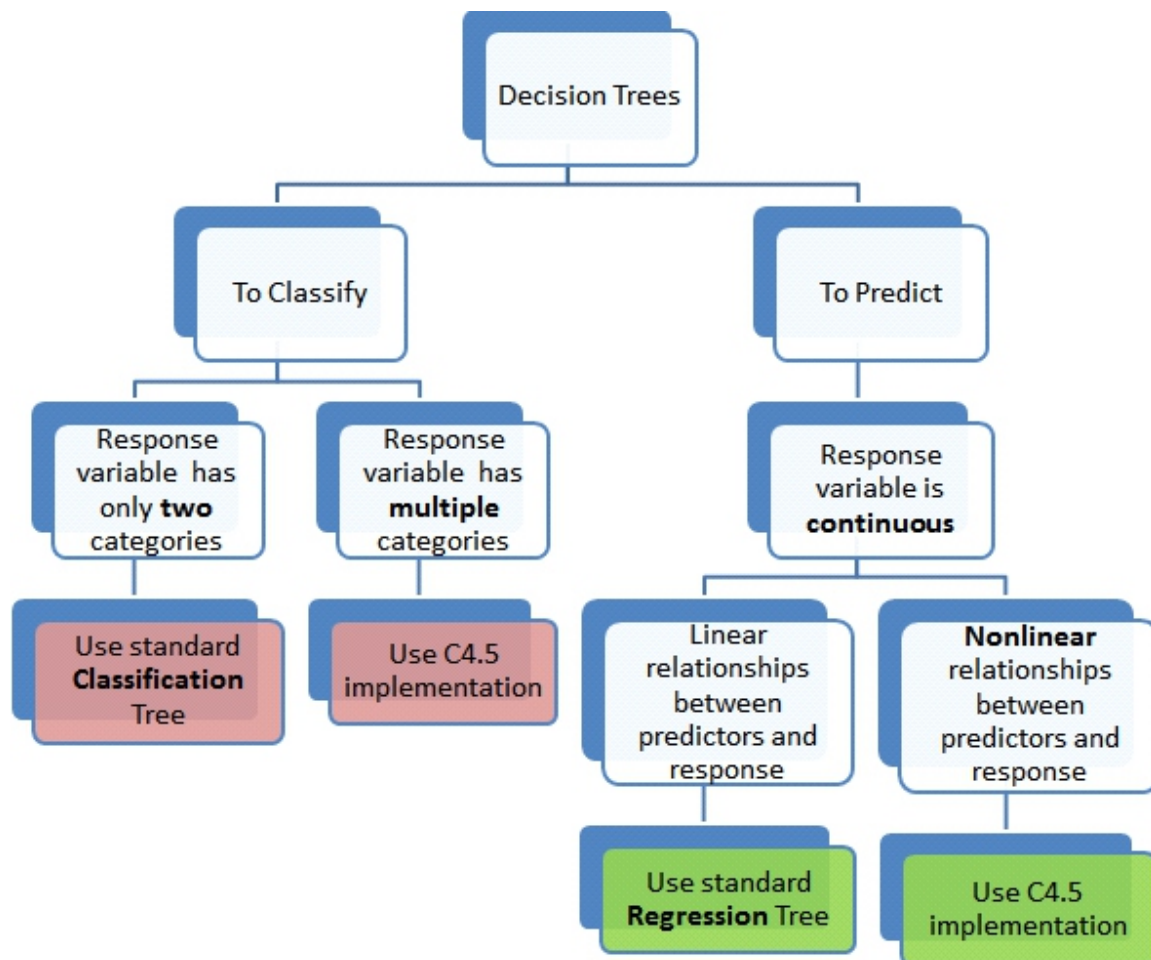
Figure 1: Classification of Decision Trees into Regression Techniques and CART

The figure above sums up the comparative analysis of three data mining techniques discussed above. When the decision trees are used for classification purposes and there are only two categories of response variable, CART is used. CART is also used in classification cases where the response variable has multiple categories and in prediction cases where there is a non-linear relationship between the predictor and response variables. Finally, in cases where there is linear relationship between predictors and response function, the standard regression techniques (linear and logistic regression) are used.

Based on the analysis of this research, Table draws a comparison between logistic regression and CART models :

LOGISTIC REGRESSION	CART (Classification & Regression Trees)
Logistic regression is used to determine probability of an event, which is given in a binary format, 0 and 1, albeit also used to build a relationship between dependent and independent variables. In logistic models, the idea is to fit a regression model using each of the dependent variables and predict the outcome of the problem using the specified conditions. A Logistic regression model makes a decision (0 or 1) by calculating the probability of an event, or what is called Logit. Logistic regression model is less interpretable. In other words, it is not a straightforward task to reach quick decisions. Logistic regression model works best in situations where there is a single decision boundary that is not necessarily parallel to the axes.	CART can be applied not only to cases where the outcome variable (the dependent variable) is of binary nature but also when the outcome is continuous, as in the cases of linear regression. However, an added advantage that CART provides over linear regression is that it is also applicable when the dependent and independent variables are non-linearly related. In a CART model, the idea is to split the data into various regions based on the homogeneity of the data and predict the outcome depending on which region a particular data point lies in. This make CART model more interpretable than logistic regression. A CART model makes a decision by following the splits of the trees based on the specified conditions.

Table 1: Comparison between Logistic Regression and CART

PREDICTIVE DATA MINING	CLUSTERING TECHNIQUE
The goal of predictive data mining techniques is to develop a statistical model that relates the dependent and independent variables and predicts the response of interest.	The goal of clustering techniques is to aggregate the information in large datasets into manageable clusters, which exhibit similar properties.
The predictive data mining techniques such as linear and logistic regression are developed through supervised learning whereby the inputs are fed to the model in the beginning stages and outputs are at the end of chain.	Clustering is an unsupervised learning process whereby the observations or input come into play at the end stages of the chain. Unsupervised learning, as in clustering, provides the ability to learn large and significantly complex models.

Table 2: Comparison between Predictive Data Mining Techniques and Clustering

Finally, there are some situations when the demand of the problem is to segment data instead of making direct predictions. In such cases, clustering can readily be used by building a predictive model by putting the data points with similar attributes into clusters. Table 2 gives a comparison of clustering with the other three data mining technique studied in this research.

CONCLUSION AND RECOMMENDATIONS

Conclusion:

In general, regression techniques work for a large variety of datasets. Linear regression assumes a linear model and works best when not only the data is of linear nature but also the outcomes are categorical. This helps in various events such as predicting the runs scored and number of wins registered by a team. Logistic regression, on the other hand, is useful when the probability of an event is determined such as in health care, when the goal is to differentiate between high risk and low risk patients.

Classification and regression trees (CART) technique, however, has certain advantages over the regression techniques. First, CART is able to predict categorical outcomes similar to linear regression but the advantage of it is that it also works when the data is non-linear in nature. This is one of the significant advantages of CART that makes it extremely useful. When compared with logistic regression, CART gives similar results but with a little less accuracy. However, CART has a more interpretable nature and helps reach quick decisions, whereas logistic regression can prove to be cumbersome at times.

Finally, clustering is a useful data mining technique when the goal of the problem is to segment data instead of predicting. Eventually, however, clustering helps develop a predictive model which is of great use in applications such as recommendation systems, where users are recommended some similar stuff based on their viewing history. All the data mining techniques studied in this report have their own merits and demerits. It all comes down to the nature of the problem and circumstances dictate which technique should be preferred over the other. A good deal of knowledge about the problem and data mining goals is, therefore, essential whenever dealing with the data mining techniques.

Limitations and Future Recommendations

The researcher has used only a handful of techniques due to the limited scope of this project. In practice, there are a lot of alternatives to the techniques discussed in this project and can all be used at the same time. However, each technique offers its own advantages and shortcomings. For instance, as it was analysed in this chapter, logistic regression provides a greater accuracy while CART has a more interpretable nature.

Another limitation of this research project was regarding the specificity of each technique. Due to the time constraint, the researcher did not get the chance to fully dive into each technique and explore in depth the underlying assumptions and concepts of each mode. The scope of the current research was limited to understanding the differences in various techniques in terms of their applications. The future research could, however, be aimed at carrying out a rigorous analysis of these techniques in terms of their underlying concepts and statistical models on which they are built.

An important limitation of this research project that could not be ignored was the size of the data. Due to educational and research nature of the project, the researcher could only access information which was publically available on the websites of different government agencies and private institutions. In practice, the companies have access to databases that contain a vast amount of information collected from multiple sources. In future projects, this limitation could be addressed by contacting government agencies and other organization to reveal information specifically for research use which is otherwise not publically available. However, the research can take comfort from the fact that big data is not necessarily better than good data and so the challenge can also be addressed by ensuring that data is selected without inducing any bias in it.

Finally, the scope of this project was limited to understanding which data mining technique is the most potent in terms of their application. As it turns out, each technique is effective with its own merits and demerits and their use depends very much on the nature of the problems. Due to the time and resources constraint, the researcher designed problems by herself by looking for what were the most common problems in the industry. In the future, contacting companies and enquiring about the specific data mining problems could better this. This would not only increase the validity of the analysis but also help in developing an understanding as to which technique is genuinely better than the others.

The direction of the future studies can be aimed at putting a greater emphasis on understanding the requirements of the project and selecting the relevant data mining techniques accordingly. One of the main takeaways for the researcher from this project is that understanding the requirements of the data mining problem is of essence and without a complete and thorough understanding of the project, no data mining technique can be correctly applied. In researcher's opinion, the future studies should be focused at developing a framework that helps select a relevant data mining technique based on the requirements and goals of the data mining problem.

REFERENCES:

- *Abonyi, J., Feil, B. (2007). "Cluster Analysis for Data Mining and System Identification". Springer Science & Business Media: Germany*
- *Alhajj, R. (2007). "Advanced Data Mining and Applications: Third International Conference, ADMA 2007, Harbin, China, August 6-8, 2007 Proceedings". Springer Science & Business Media: Germany*
- *Berry, M. J., Linoff, G. S. (2000). "Mastering Data Mining: The Art and Science of Customer Relationship Management". Wiley Computer Publishing: New York. Bishop, C. M. (2006). "Pattern Recognition and Machine Learning". Springer: USA*
- *Chang, Y. (2008). "Robustifying Regression and Classification Trees in the Presence of Irrelevant Variables". ProQuest: USA*
- *Chatterjee, S., Hadi, A. S. (2013). "Regression Analysis by Example". John Wiley & Sons: USA*
- *Christensen, R. (2013). "Log-Linear Models and Logistic Regression". Springer New York: USA Collins English Dictionary (2014)*
- *Gelman, A., Hill, J. (2006). "Data Analysis Using Regression and Multilevel/Hierarchical*
- *Models". Cambridge University Press: USA*
- *Ghmueli, G., Patel, N. R. and Bruce, P. C. (2006). "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner". John Wiley and Sons: US*
- *Gorunescu, F. (2011). "Data Mining: Concepts, Models and Techniques". Springer Science & Business Media: USA*
- *Han, J., Kamber, M. (2011). "Data Mining: Concepts and Techniques", Morgan-Kaufmann Academic Press: San Francisco*

- *Jiawei Han, Michelin Kamber & Jian Pei (2012), "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems, Third Edition, published by Elsevier Inc. 2012*

- *Haynes, K. L. (2006). "Object Recognition Using Rapid Classification Trees". Florida State University: USA*

- *Hilbe, J. (2009). "Logistic Regression Models". CRC Press INC: USA*

- *Hornick, M. F., Marcade, E. and Venkayala, S. (2010). "Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for architecture, design, and implementation". Morgan Kaufmann: US*

- *Liu, Y. (2006). "Interactive Visual Data Mining Modelling to Enhance Understanding and Effectiveness of the Process." ProQuest: US*

- *Manel, S., Dias, J. M., & Ormerod, S. J. "Comparing Discriminant Analysis, Neural Networks and Logistic Regression for Predicting Species Distribution: A Case Study with a Himalayian River Bird," Ecol. Model. 120 (1999), pp. 337-347.*

- *Marsland, S. (2011). "Machine Learning: An Algorithmic Perspective". CRC Press: USA Menard, S. (2010). "Logistic Regression: From Introductory to Advanced Concepts and Applications". SAGE: USA*

- *M. C. Okur (2014), "On Ethical and Legal Aspects of Data Mining", Journal of Yasar University, Izmir, Turkey*

- *Micheal Koploy, (2014), "Compare Data Mining Software: Top 10 Most Recommended Data Mining Software Systems", accessed from http://www.softwareadvice.com/bi/data-mining- comparison/ on 7th April, 2014.*

- *Mirkin, B. (2005). "Clustering for Data Mining: A Data Recovery Approach". CRC Press: USA*

- *Montgomery, D. (2007). "Introduction to Linear Regression Analysis, Fourth Edition Solutions Set". Wiley: USA*

- *Munoz, J., & Felicisimo, A. M. "Comparison of Statistical Methods Commonly used in Predictive Modelling," Journal for Vegetation's Science 15 (2004), pp. 285-292.*

- *National Electronic Library for Health: Knowledge Management Glossary*

- *Olson, D. L. and Delen, D. (2008). "Advanced Data Mining Techniques". Springer Science & Business Media: USA*

- *[Online]: SQL Data Mining (2012),"History of Data Mining", accessed from http://www.sqldatamining.com/index.php/data-mining-basics/history-of-data-mining*

- *[Online]: Diffen, "Data Vs Information" (2014), accessed from http://www.diffen.com/difference/Data_vs_Information*

- *[Online]: Business Dictionary, (2014) "Data", accessed from http://www.businessdictionary.com/definition/data.html*

- *[Online]: Search CRM, (2014) "Knowledge Base",accessed from http://searchcrm.techtarget.com/definition/knowledge-base*

- *[Online]: In STEP, (2014)"What is Knowledge?", accessed from http://instep.net.nz/Knowledge-and-theory/What-is-knowledge*

- *Osmar R. Zaiane, (1999), "Introduction to Data Mining", Principles of Knowledge Discovery in Databases, CMPUT690, 1999.*

- *Pal, N. (2007). "Advanced Techniques in Knowledge Discovery and Data Mining". Springer Science & Business Media: USA*

- *Pampel, F. C. (2000). "Logistic regression: a primer". SAGE Publications: USA*

- *Pan, T. (2008). "Using the Multivariate Multilevel Logistic Regression Model to Detect DIF: A Comparison with HGLM and Logistic Regression DIF Detection Methods". Michigan State University: USA*

- *Pittman, K. (2008). "Comparison of Data Mining Techniques Used to Predict Student Retention". ProQuest: USA*

- *Pyle, D. (2003). "Business Modelling and Data Mining". Morgan Kaufmann: USA*

- *Rajaraman, A., Ullman, J. D. (2011). "Mining of Massive Datasets". Cambridge University Press: United Kingdom*

- *Rajagopalan, B., Krovi, R. (2002). "Benchmarking Data Mining Algorithms", Journal of Database Management, Jan-Mar, 13, 25-36*

- *Richard J. Roiger & Micheal Geatz, (2002), "Data Mining: A Tutorial Based Primer", Published by Addison-Wesley Publishing, September 26, 2002. Reviewed by Prof. Keith Rennolls.*

- *Rokach, L. (2008). "Data Mining with Decision Trees: Theory and Applications". World Scientific: Singapore*
- *Shital P. Bora, (2001), "Data Mining and Warehousing", Department of Computer Science and Application, Sardar Patel College, Chandrapur, India. Publisehd by IEEE, 2011.*
- *Stoll, L., Fink, D., and Earl, L. (2003). It's about Learning (and it's about Time): What's in it for Schools? London: Routledge Falmer.*
- *S.L. Kendal and M. Creen, (2007), "An Introduction to Knowledge Engineering", School of Computing & Technology, University of Sunderland, Tyne and Wear, UK. Published by Springer Link in 2007*
- *Tamim Jabban,(2010),"Data Mining", http://www.contrib.andrew.cmu.edu/~tjabban/datamining.html*
- *Two Crows Corporation (2005), "Introduction to Data Mining and Knowledge Discovery", Technology Report, Two Crows Corporation, Third Edition.*
- *Weisberg, S. (2005). "Applied Linear Regression". John Wiley & Sons: USA*
- *Weiss, S. M. (1998). "Predictive Data Mining: A Practical Guide". Morgan Kaufmann: US Wel, L. Van and Royakkers, L. (2004), "Ethical Issues in Web Data Mining. Ethics and Information Technology", 6; 129-140.*
- *Windt, K. (2012). "Robust Manufacturing Control: Proceedings of the CIRP Sponsored*
- *Conference RoMaC 2012, Bremen, Germany". Springer Science & Business Media: USA*
- *Wing Kee Ho and Xiaohua Luan (2012), "Data Mining: An Introduction", A power point presentation accessed from http://www.docstoc.com/docs/108233884/Data-Mining-An- Introduction on the 2nd April 2014.*
- *Yan, X. (2009). "Linear Regression Analysis: Theory and Computing". World Scientific: Singapore*
- *Yohannes, Y., Webb, P. (1998). "Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity". Intl Food Policy Res Inst: USA*

CONTEMPORARY TESTING BENEFITS OVER TRADITIONAL TESTING

Mohd Shafaat Hussain

M. Tech. Scholar, Department of Computer Science and Engineering,
Al-Falah School of Engineering and Technology, Dhauj, Faridabad (Haryana),India
ershafaat@gmail.com

A B S T R A C T

This research papers based on modern testing advancement over the traditional testing. Normally traditional testing follow a hierarchy (1) Requirement Phase (2) Design phase (3) Code and Build (4) Testing (5) Maintenance .In this hierarchy testing follow a static strategy where it lies at a certain position, this is the reason why some of the early softwares were failed because they follow this rigid hierarchy. The best thing is that as early as we find and eliminate the bug is more cost effective than to eliminate and fix at later stages. For example, fixing an error in maintenance is ten times more expensive than fixing it during execution, but many organizations think that some testing should take place at every stage. The cost is very much important factor for both the client and the developing team. If due to a bug the overall cost will increase or if the bug will reach to the customer end it will become failure. A software bug in a computer program or system that causes it to produce an incorrect or unexpected result, or to behave in unintended ways. Due to unidentified bugs in the Software there are many accidents occurred in the history .Earlier when the software development follow the basic models like Waterfall the testing was bound to a certain stage and it came into the picture when the early stages completed due to this many early softwares failed and the necessity of the early testing came into the picture which results in the advancement of the Software models too and other models like Build and Fix model, Iterative Model ,Spiral model and the V- model which emphasized the necessity of the early testing ,it includes some amount of testing with each phase and the Software testers involve in the development process from the requirement phase.

Index Terms— Modern testing, Traditional approach, Development Models, V-Model.

I. INTRODUCTION

The Testing is the important phase of the software development, Software testing [1] is something which is much more than error detection and plays an important role to produce a quality and bug free software. Due to lack of proper software testing many accidents happened that results the loss of cost of millions and many lives .One of the subcontractors NASA used when building its Mars climate orbiter had used English units instead of the intended metric system, which caused the orbiter's thrusters to work incorrectly. Due to this bug the orbiter crashed almost immediately when it arrived at Mars in 1999. The cost of the project was $327 million, not to mention the loss of time. In 1994 in Scotland, A chinook helicopter crashed and killed all 29 passengers. While initially the pilot was blamed for the crash, that decision was later overturned since there was evidence that a systems error had been the actual cause. These are few example but there are many such examples which prove this how much testing is important in the software development software testing involves the following strategies [2] that are (1) Verification of the Software (2) Error detection (3)Validation of the Software, In verification [3] we

check the consistency of the software by evaluating the consequences against the pre-defined conditions , Error detection is the process of checking a condition because of which software is not working as specified. In validation we will check what has been specified and what the user actually wanted. The software development generally entails 40-50% cost for the software development. The traditional model like waterfall model [4] follow the approach where testing lies below the Coding phase because of that we have bugs come not only from coding phase but also from requirement and design phases . So instead of choosing traditional testing we generally choose the way of testing which is bound to provide a bug free product. In earlier models of software development models every stage was bound at certain position and therefore they are much vulnerable for the bugs. Bug is something which is present in the software and it was not expected. Bugs can be divided into three parts [5] and they are as follows (1) Extra (2) Wrong (3) Missing, suppose a feature in the software which was not demanded then that particular bug come into the category of Extra. Suppose there was requirement and it is wrongly implemented in the software then It comes into the category of the Wrong. The last bug comes into the category "Missing ", when a specified requirement is absent from the software which was specified in the software. Because of the bounded stages in the early models of the software development , all categories of bugs were not covered .The organizations keep this fact into their considerations and the new models came into existence which emphasize the need of early testing [6] and the new models like Code and Fix , Incremental Model , Iterative model and V Model .

II. ROLE OF TESTING IN VARIOUS DEVELOPMENT MODELS

For every organization which deals with the software development work for them software testing is precisely an important and precious stage in the Software Development Life Cycle (SDLC) [7] . Software testing has turn into a part of programming and it is good to begin testing from the initial stage, to avoid the afflictions by repairing the errors at the last stage. Also in software development life cycle (SDLC) testing entails pick up consistency, interpretation and other major elements, which may state under software requirement specification. Prior to see the role of testing in the different Software development models we will see what is the Software Development Life Cycle (SDLC)? The first phase of the SDLC is the Requirement Analysis, In this phase the requirements gather by the Business Analyst, for that interviews and various interactions with client will be made, sometimes the client is technically sound and sometimes he/ she is not as much technically sound, In this regard proper interviews conduct and the BA [8] collect the necessary requirements from the client. In the requirement phase Business manager check whether the requirements are technically feasible or not, whether it is feasible in terms of time and cost .In the second phase which is design phase with the help of the requirements gathered from the Requirement Analysis phase, the system design will make. In the next phase implementation of project is done. With Respect to the system design, correct development is made to expand that design. As per the necessity we will choose the correct programming language for the development, After the coding phase the testing phase comes into the picture, testing is done to make sure that what we developed meet our desired requirements .The last phase is Maintenance phase It is the stage of SDLC, where the application which is implemented is handled to client who is responsible for conserving and using it for appropriate actions. The implemented application should be available for any adjustment to do in coding. These are the overall phases of the SDLC and it is a static hierarchy and earlier models like waterfall model use this hierarchy as it is and the testing phase lies between the implementation and maintenance but because of this many early projects which are based on this rigid hierarchy were failed. Sometimes the world suffered from serious hazards from the software bugs like for example In 1980,

NORAD reported that the US was under missile attack. The problem was caused by a faulty circuit, a possibility the reporting software hadn't taken into account. In 1994 in Scotland, a Chinook helicopter crashed and killed all 29 passengers. While initially the pilot was blamed for the crash that decision was later overturned since there was evidence that a systems error had been the actual cause. There are many such examples which have proved what happened due to unhanding of these bugs [9]. The organizations and the it professionals thought that there is immense need of early testing in the projects for their success , for this they worked on exploring the models in which they can implement the early testing and their search came to an end on the models like incremental Model, iterative model , Spiral and V-Model. Now if we see these models individually then we will find that how they involve early testing and the eventually how the positive consequences came.

If we see the incremental model [10] then we will find that every next build will generate on the basis of what the previous model lack for. For the incremental model a proper testing apply to make a better build till the customer satisfaction.
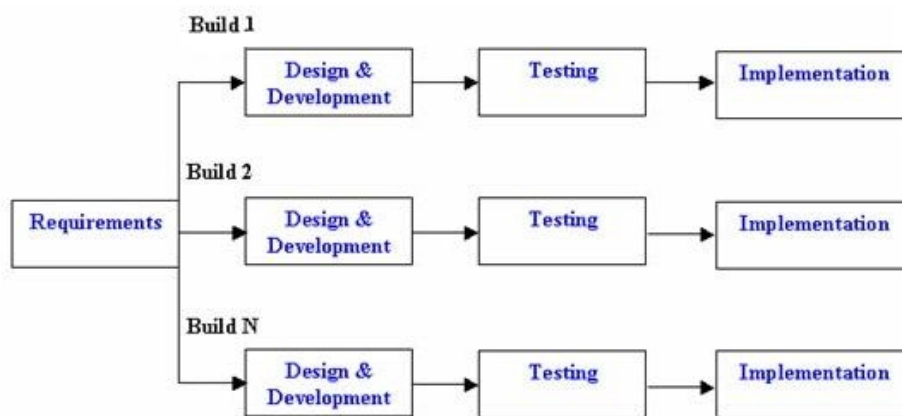
Incremental Life Cycle Model

An iterative life cycle model [11] does not attempt to start with a full specification of requirements. Instead, development begins by specifying and implementing just part of the software, which can then be reviewed in order to identify further requirements. This process is then repeated, producing a new version of the software for each cycle of the model. In Iterative model earlier we do not focus on the whole software development but on just part of it and simultaneously do testing with each part.
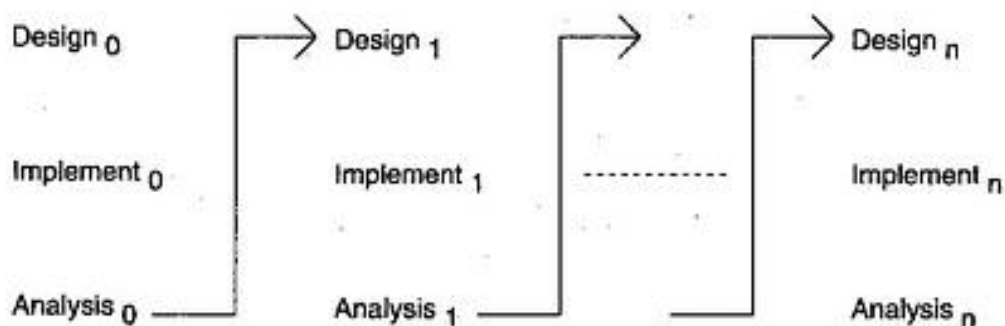

Fig2. Iterative Life Cycle Model

If we talk about the spiral model [12], it encompass a most common issue that is Risk [13]. Risk is something that can cause failure, so to avoid risk is its most common feature which this model dealt with. In spiral model the early testing plays an important role, it has four phases and they are as follows 1. Planning Phase 2. Risk Analysis 3. Engineering Phase 4.Evaluation phase, The spiral model encompass an important thing that is risk. Risk is something that can cause failure. In this model testing also involve in the Risk analysis phase to determine the risk and remove them.
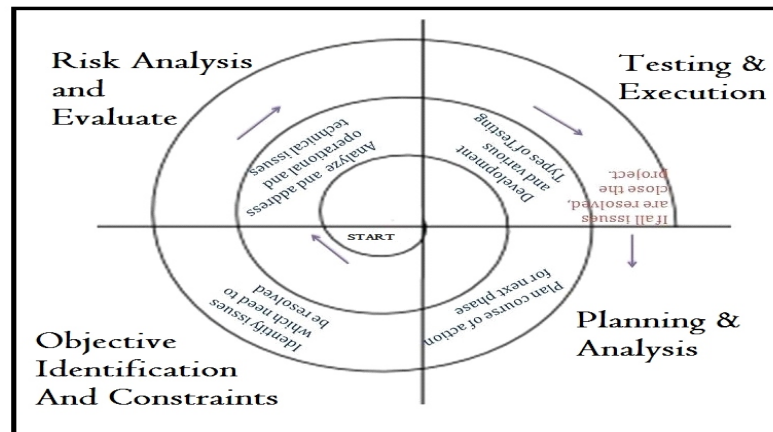

Fig3. Spiral Model

III. IMPORTANCE OF EARLY TESTING IN V MODEL OVER OTHER MODELS:

V model precisely modeled to emphasize the need of the early testing [14], it modeled in such a way such that some part of the testing involve in each phase. It is a model which is based on the fact that "as early bug will remove the more cost effective it is ".
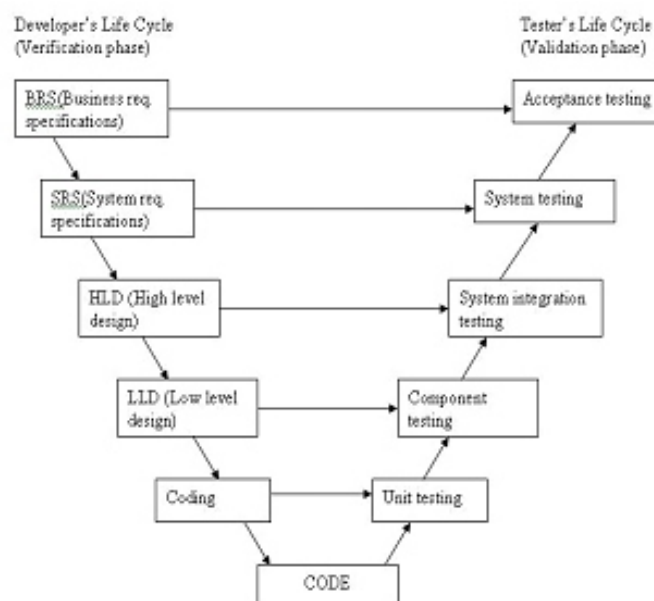

Fig4 . V Model

Like other models start with the documents like SRS and BRS this model generate a System test plan .This test plan focuses on meeting the functionality specified in the requirements gathering. HLD (High level design) phase focuses on system architecture and design. It provide overview of solution, platform, system, product and service/process. An integration test plan is created in this phase as well in order to test the pieces of the software systems ability to work together. LLD (Low Level Design) phase is where the actual software components are designed. It defines the actual logic for each and every component of the system. Class diagram with all the methods and relation between classes comes under LLD. Component tests are created in this phase as well. The Implementation phase is, again, where all coding takes place. Once coding is complete, the path of execution continues up the right side of the V where the test plans developed earlier are now put to use. Coding is at the bottom of the V-Shape model [15]. Module design is converted into code by developers. Unlike the others model V model emphasize the need of early testing the application we will get eventually is less vulnerable to the bugs.

IV. ENHANCEMENT OF THE V MODEL ACCORDING TO THE CURRENT SCENARIO

V model is also a software development model which is the extension of the waterfall model. V model signifies the relationship of each phase of the development and its associated plans of testing, instead of moving downwards it bent upwards after the coding phase and form a structure which is alike to the English alphabet V. The V-model organizes a well-structured technique in which each stage can be implemented by the comprehensive documentation of the previous phase. Early testing start in V model before coding saves a lot of time of the development. Activities like test designing start before coding really assist to reduce the overall time for the development. The overall aim of the V model is the amendment in the efficiency and effectiveness of the overall development process, it also relates various development activities and the testing activities. This model demonstrates how each following stage should verify and validate work done in the previous phase, and how work done during development is used to guide the separate testing stages.
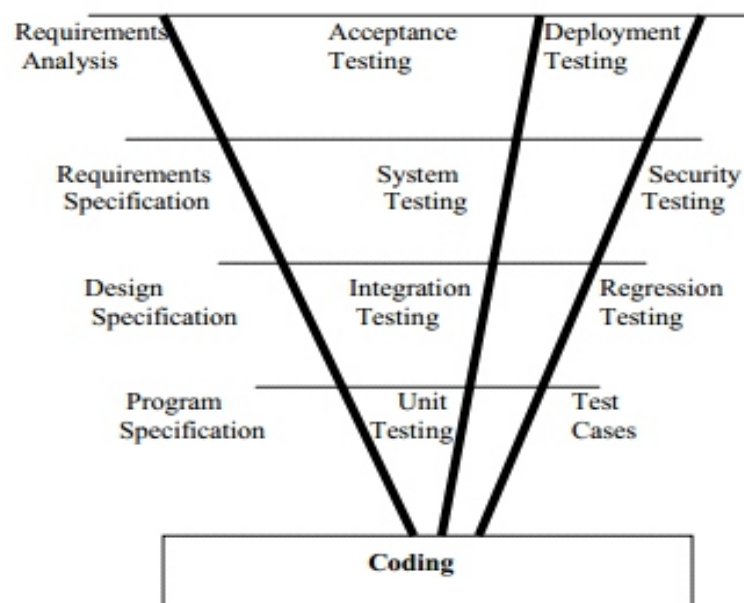

Fig5. Enhanced V Model

In V model. Like any other model V model have also some limitations for an examples if all the development and testing activities have been completed but if they are not maintained properly it will be a serious issue , so in case of the enhancement of the V model we will include the maintenance [16] activities too which increase the percentage of efficiency and effectively of overall development, Once the processes of the development phase starts simultaneously the activities of the testing phase and maintenance phase also begin. For example in Unit testing the modules which are formed tested properly and the test cases designed Secondly if we come on to the Integration testing [17] where the different modules are integrated and the testing for integration will check. But integration will be more vulnerable if the regression testing [18] will not perform, here regression testing will act for maintenance activity, , Now when the integration part will complete we have the system as whole and system testing will come into the picture where the whole system will be tested , Along with system testing , we need to do the security testing just to ensure that the various components are safe from various security threats accordingly we will apply certain security measures . Eventually when the client will satisfied by the alpha and beta acceptance tests [19] , the product will be deploy to the client , Thus we see how the continuous interaction of the three activities development , testing and maintenance assist the overall development assure the amendment in the efficiency and the effectiveness of the software and also increase the reliability. We do not do anything with the V model but just entail a more important activity which is the maintenance activity which bring not only customer satisfaction but also reliability on the development organization

V. CONCLUSION

After investigation of all models through the various attributes, it has been found that the original waterfall model is utilized by different big companies for their own projects .Since the development team is acquainted to the environment and it is practicable to specify all requirements of working environment, It has been found that the waterfall model suited for the small projects in which a bug will entail not a huge loss. Iterative waterfall model overcome the disadvantage of actual waterfall model .It permit feedback to next development stage. Since importantly reduce rework and lead to the creation of working model in lower capital cost. Spiral model is used for development of big, complex and pricy projects like scientific Projects .Since spiral model enables the project term to signify the biggest risk at the smallest total cost. In case of V model as the early testing is the major concern, and different test plans made at different stages of the development cycle to take out the bug free application eventually. We also propose how the Enhanced V model describing that for capable software testing management along with the development and testing process, the maintenance practice is also vital. Thus we have integrated these processes for efficient software testing management.

VI. ACKNOWLEDGMENT

V1I. REFERENCE

1. Bertolino , A : A knowledge Area Description of the   Software Testing . Guide to the SWE- BOK.

2. Boris Beizer, "Software testing techniques", Second edition, (1990).

3. B. Beizer, "Software Testing Techniques," Second Edition, Van No strand Reinhold Company Limited, 1990

4. N. Munassar and A. Govardhan, "A Comparison between Five Models of Software Engineering",

5. B.B Agarwal, S.P Tayal, ", Software Engineering and Testing ".

6. Rook, Paul, E. Rook, "Controlling software projects".

7. Roger S. Pressman, Software Engineering: A Practitioner's Approach

8. http://www.zsassociates.com/pdfs/ZS%20FocusSheet%20-%20Business%20Analyst.pdf

9. http://royal.pingdom.com/2009/03/19/10-historical-software-bugs-with-extreme-consequences/ "10 historical software bugs with extreme consequences"

10. http://en.wikipedia.org/wiki/Incremental_build_model "Incremental Build Model "

11. Software Methodologies Advantages various SDLC models.mht & disadvantages of

12. Boehm B, "A Spiral Model of Software Development and Enhancement", ACM SIGSOFT Software Engineering Notes", "ACM", 11(4):14-24, August 1986

13. LIONEL GALWAY, "Quantitative Risk Analysis for Project Management A Critical Review ","Working Paper" , February 2004

14. http://en.wikipedia.org/wiki/V-Model_%28software_development%29"V-Model(software development).

15. Rook, Paul, E. Rook, "Controlling software projects", IEEE Software Engineering Journal, 1(1), 1986, pp. 7-16.

16. IEEE. 1993. IEEE Standard for Software Maintenance. IEEE Std 1219-1993. Institute of Electrical and Electronics Engineers, inc., New York, NY.

17. http://en.wikipedia.org/wiki/Integration_testing "Integration testing "

18. G. Rothermel and M. J. Harrold, Analyzing Regression Test Selection Techniques, IEEE Transactions on Software Engineering, V.22, no. 8, August 1996, pages 529-551.

19. http://en.wikipedia.org/wiki/Acceptance_testing "Acceptance Testing"

Contribution of e-Governance in Present Time

Alima Naqvi

Research Scholar, Department of Western History, Lucknow University

A B S T R A C T

Electronic governance or e-governance is the application of information and communication technology (ICT) for delivering government services, exchange of information communication transactions, integration of various stand-alone systems and services between government-to-customer, government-to-business.

The National e-Governance Plan of Indian Government seeks to lay the foundation and provide the impetus for long-term growth of e-Governance within the country. This section provides information on creation of the right governance and institutional mechanisms, setting up the core infrastructure and policies and implementation of a number of Mission Mode Projects at the Centre, State and integrated service levels.

*Keywords – **e-governance ,government policy ,impact on business , affect on people, urban and rural***

The "e" in e-Governance stands for 'electronic'. Thus, e-Governance is basically associated with carrying out the functions and achieving the results of governance through the utilization of ICT (Information and Communications Technology).

While Governance relates to safeguarding the legal rights of all citizens, an equally important aspect is concerned with ensuring equitable access to public services and the benefits of economic growth to all. It also ensures government to be transparent in its dealings, accountable for its activities and faster in its responses as part of good governance.

However, this would require the government to change itself – its processes, its outlook, laws, rules and regulations and also its way of interacting with the citizens. It would also require capacity building within the government and creation of general awareness about e-Governance among the citizens.

Information and Communication Technology

Provides efficient storing and retrieval of data, instantaneous transmission of information, processing information and data faster than the earlier manual systems, speeding up governmental processes, taking decisions expeditiously and judiciously, increasing transparency and enforcing accountability. It also helps in increasing the reach of government – both geographically and demographically.

In India, the main thrust for e-Governance was provided by the launching of NICNET National Information Centre Network (NICNET) in 1987 – the national satellite-based computer network. This was followed by the launch of the District Information System of the National Informatics Centre (DISNIC) programme to computerize all district offices in the country for which free hardware and

software was offered to the State Governments. NICNET was extended via the State capitals to all district headquarters by 1990. In the ensuing years, with ongoing computerization, tele-connectivity and internet connectivity established a large number of e-Governance initiatives, both at the Union and State levels.

Some Initiatives by the Government

Government to Citizen Initiatives:

Computerization of Land Records:

In collaboration with NIC. Ensuring that landowners get computerized copies of ownership, crop and tenancy and updated copies of Records of Rights on demand.

Bhoomi Project:

Online delivery of Land Records. Self-sustainable e-Governance project for the computerized delivery of 20 million rural land records to 6.7 million farmers through 177 Government-owned kiosks in the State of Karnataka.

Gyandoot:

It is an Intranet-based Government to Citizen service delivery initiative. It was initiated in the Dhar district of Madhya Pradesh in January 2000 with the twin objective of providing relevant information to the rural population and acting as an interface between the district administration and the people.

Lokvani Project in Uttar Pradesh:

Lokvani is a public-private partnership project at Sitapur District in Uttar Pradesh which was initiated in November, 2004. Its objective is to provide a single window, self-sustainable e-Governance solution with regard to handling of grievances, land record maintenance and providing a mixture of essential services.

Project Friends in Kerala:

FRIENDS (Fast, Reliable, Instant, Efficient Network for the Disbursement of Services) is a Single Window Facility providing citizens the means to pay taxes and other financial dues to the State Government. The services are provided through FRIENDS Janasevana Kendrams located in the district headquarters.

e-Mitra Project in Rajasthan:

e-Mitra is an integrated project to facilitate the urban and the rural masses with maximum possible services related to different state government departments through Lokmitra-Janmitra Centres/Kiosks.

e-Seva (Andhra Pradesh):

This project is designed to provide 'Government to Citizen' and 'e-Business to Citizen' services. The highlight of the eSeva project is that all the services are delivered online to consumers /citizens by connecting them to the respective government departments and providing online information at the point of service delivery.

Admission to Professional Colleges – Common Entrance Test (CET):

With the rapid growth in the demand as well as supply of professional education, the process of admission to these institutions became a major challenge in the early 1990s. Recourse was then taken to ICT to make the process of admission transparent and objective. One of the pioneering efforts was made by Karnataka. The State Government decided to conduct a common entrance test based on which admission to different colleges and disciplines was made.

Government to Business Initiatives:

e-Procurement Project in Andhra Pradesh and Gujarat: To reduce the time and cost of doing business for both vendors and government.

Ministry of Corporate Affairs

The project aims at providing easy and secure online access to all registry related services provided by the Union Ministry of Corporate Affairs to corporates and other stakeholders at any time and in a manner that best suits them.

Government to Government Initiatives:

Khajane Project in Karnataka:

It is a comprehensive online treasury computerization project of the Government of Karnataka. The project has resulted in the computerization of the entire treasury related activities of the State Government and the system has the ability to track every activity right from the approval of the State Budget to the point of rendering accounts to the government.

Smart Government (Andhra Pradesh):

Smart Government has been developed to streamline operations, enhance efficiency through workflow automation and knowledge management for implementation in the Andhra Pradesh Secretariat.

National E-governance Plan :

The National e-Governance Plan (NeGP) has been formulated by the Department of Electronics and Information Technology (DEITY) and Department of Administrative Reforms and Public Grievances (DARPG) in 2006.

The NeGP aims at improving delivery of Government services to citizens and businesses with the following vision: "Make all Government services accessible to the common man in his locality, through common service delivery outlets and ensure efficiency, transparency & reliability of such services at affordable costs to realise the basic needs of the common man."

Central government initiatives as mission mode projects (MMP)

e-office

The Government of India has recognized the need to modernize the Central Government offices through the introduction of Information and Communications Technology. e-Office is aimed at increasing the usage of work flow and rule based file routing, quick search and retrieval of files and office orders, digital signatures for authentication, forms and reporting components.

Unique Identification

The unique identification project was conceived as an initiative that would provide identification for each resident across the country and would be used primarily as the basis for efficient delivery of welfare services. It would also act as a tool for effective monitoring of various programs and schemes of the government.

Pensions

The pensions Mission Mode Project is primarily aimed at making the pension/ retirement related information, services and grievances handling mechanism accessible online to the needy pensioners, through a combination of interactive and non-interactive components, and thus, help bridge the gap between the pensioners and the government.

Banking

The Banking Mission Mode Project is yet another step towards improving operational efficiency and reducing the delays and efforts involved in handling and settling transactions. The MMP which is being implemented by the banking industry aims at streamlining various e-services initiatives undertaken by individual banks. Implementation is being done by the banks concerned, with the banking Department providing a broad framework and guidance.

Posts

Modernization of Postal Services has been undertaken by the Department of Posts through computerization and networking of all post offices using a central server-based system, and setting up of computerized registration centres (CRCs).

State Mission Mode projects

e-Governance in Municipalities

It is a unique initiative of the Government of India conceptualized under the umbrella of the overall National e-Governance Plan (NeGP) and the Jawaharlal Nehru National Urban Renewal Mission (Jnnurm) aimed at improving operational efficiencies within Urban Local Bodies (ULBs).

Public Distribution System

Computerization of the Public Distribution System is envisaged as an end-to-end project covering key functional areas such as supply chain management including allocation and utilization reporting, storage and movement of food grains, and transparency portal, digitization of beneficiary database, Fair Price Shop automation, etc.

Health

Information and Communication Technology for programme management has been undertaken by the Ministry of Health & Family Welfare in the Mother and Child Tracking System (MCTS) programme and the Ministry envisages a more comprehensive use of ICT including for Hospital Information Systems, supply chain management for drugs and vaccines, providing ICT tools to Accredited Social Health Activist (ASHA) and Auxillary Nurse Midwife (ANM) workers, programme management of National Rural Health Mission (NRHM), etc through this MMP.

e-panchayat

The Panchayati Raj Institutions (PRIs) are saddled with the problems of inadequate physical and financial resources, technical capabilities and extremely limited computerization. As a result, the potential of PRIs as the preferred delivery channel for the schemes of State and Centre as well as for citizen services has not been fully realized. While some computerization efforts for PRIs have been made by NIC over the years, the e-Governance revolution sweeping the country has not touched the PRIs yet in significant measure. The Ministry of Panchayati Raj, Government of India has therefore decided to take up the computerization of PRIs on a mission mode basis

e-District

e-District is one of the 31 Mission Mode Projects under National e Governance Plan (NeGP) with the DIT, GoI being the nodal ministry. This project aims at providing support to the basic administrative unit i.e. District Administration by undertaking backend computerization to enable electronic delivery of high volume citizen centric government services which would optimally leverage and utilize the three infrastructure pillars of State Wide Area Networks (SWAN), State Data Centers (SDC) and Common Service Centers (CSCs) to deliver services to the citizen at his doorsteps.

National Land Records Modernization Programme (NLRMP)

A Project for Computerization of Land Records (CLR) was launched in 1988-89 with the intention to remove the inherent flaws in the manual system of maintenance and updation of Land Records. In 1997-98, the scheme was extended to tehsils to start distribution of Records of Rights to landowners on demand. The focus of the entire operation has always been to employ state of the art information technology (IT) to galvanize and transform the existing land records system of the country.

Integrated Mission Mode Projects

e-procurement

Ministry of Commerce & Industry (Department of Commerce) has been nominated as the Nodal Ministry for implementation of e-Government Procurement (e-GP) Mission Mode Projects (MMP). The vision of the e-Procurement MMP is "To create a national initiative to implement procurement reforms, through the use of electronic Government procurement, so as to make public procurement in all sectors more transparent and efficient".

e-Courts

The e-Court Mission Mode Project (MMP) was conceptualized with a vision to transform the Indian judiciary by making use of technology. The project had been developed, following the report submitted by the e-Committee under Supreme Court on national policy & action plan on implementation of information communication tools in Indian judiciary.

A clear objective – to re-engineer processes and enhance judicial productivity both qualitatively and quantitatively to make the justice delivery system affordable, accessible, cost effective, transparent and accountable.

e-Biz

The e-Biz Mission Mode Project, being executed by Department of Industrial Policy and Promotion (DIPP), Ministry of Commerce and Industry, Government of India, was conceptualized with the vision. Its vision is "To transform the business environment in the country by providing efficient, convenient, transparent and integrated electronic services to investors, industries and business throughout the business life cycle".

Common Services Centres

The CSCs would provide high quality and cost-effective video, voice and data content and services, in the areas of e-governance, education, health, telemedicine, entertainment as well as other private services. A highlight of the CSCs is that it will offer web-enabled e-governance services in rural areas, including application forms, certificates, and utility payments such as electricity, telephone and water bills.

We have seen how the concept of e-governance has evolved in Indian scenario and how much it is required for transparency and accountability on the part of government and at the same time it is also a toll to increase the participation of people in policy making by empowering them with the right information at right time. The penetration of internet, telecommunication services in India has increased in the last decade and this gives a ray of hope to the citizens of India to fight with the long persisting problems of poverty, corruption, regional disparity and unemployment. But at the same time, due to slow pace of project completion, red-tape and resistance from the side of government employees and citizens too has not given the desired result.

References

NeGP website. "Approval Details of NeGP". NeGP Website. Retrieved 17 July 2014. "Second Administrative Reforms Commission Report". Arc.gov.in. Retrieved 2011-03-01. "Second Administrative Reforms Commission Preface to the Report" (PDF). Retrieved 2011-03-01.

Indg. "about NeGP". Indg.in. Retrieved 17 July 2014.

^ Revisiting NeGP: eBharat2020:The proposed future NeGP 2.0 by CSR Prabhu in CSI Communications Oct 2011 (access date 5 Mar 2012)

Cost Estimation Based On Functional & Non Functional Requirements

Md. Shahzad Hannan Mansoor

M. Tech Scholar, Department of Computer Science & Engineering Al-Falah School of Engineering and Technology Dhauj, Faridabad, Haryana, mail:khan.shahzad12@gmail.com

A B S T R A C T

Function Point methodology with an enhanced approach, focus on type of the domain of the Software Projects. The assessment of a Project based on Functional and Non Functional Requirements should be the first step for a Project manager before going to find the Cost or effort of the project.

Functional Requirements (FR):

Functional requirements describe what the system should do.

Non Functional Requirements (NFR):

Non-functional requirements describe how the system works.

The **NFR consist of** Performance, Response Time, Throughput, Utilization, Static Volumetric, Scalability, Capacity, Availability, Reliability, Recoverability, Maintainability, Serviceability, Security, Regulatory, Manageability, Environmental, Data Integrity, Usability, Interoperability

1. Requirements directly associated with the Non Functional Requirements (NFR) along with Functional Requirements (FR).

2. Requirements associated with Functional Requirements (FR) and minor impact of the Non Functional Requirements (NFR) and Value Adjustment Factor (VAF).

Mathematical algorithms implemented to compute cost as a function of a number of variables for above 2 use cases.

Based on these approaches we have categories the do Requirements and collected the 30 software developed projects from different streams, we evaluated the Function Point and match with the actual effort based on Project Kickoff meeting started and the Project Closure submitted.

The data showed that the estimation of the mathematical algorithms implemented values is on a par in accuracy, while there is still significant room for improvement in order to better address the prediction challenges faced in practice.

INTRODUCTION

This paper presents an empirical evaluation of Effort and Cost Estimation based on Functional and Non Functional Requirements and focus on type of the domain of the Software Projects.

Responsibility of the project manager is to have accurate estimates of effort, cost, and schedule involved in software development.

There is major relation and impact based on accurate estimates of effort, cost, and schedule or timeline or delivery get failed, exceed budget and go overscheduled or bottleneck on work environment.

Software Effort and Cost estimation is the set of techniques and procedures that organize an estimate for proposal bidding, project planning and probability estimates.

Accurate Effort and Cost estimation means better planning and efficient use of project resources such as cost, duration and effort requirements for software projects.

Function Point methodology with an enhanced approach, focus on type of the domain of the Software Projects. The assessment of a Project based on Functional and Non Functional Requirements should be the first step for a Project manager before going to find the Cost or effort of the project.

Functional Requirements (FR):

Functional requirements describe what the system should do.

Non Functional Requirements (NFR):

Non-functional requirements describe how the system works.

The **NFR consist of** Performance, Response Time, Throughput, Utilization, Static Volumetric, Scalability, Capacity, Availability, Reliability, Recoverability, Maintainability, Serviceability, Security, Regulatory, Manageability, Environmental, Data Integrity, Usability, Interoperability

Domains are: Banking and Financial Services, Defense Projects, Online Examination System, Confidential, Restricted, and Regulatory (Company Policy or Compliance) like SOX (Sarbanes–Oxley Act), Gramm-Leach-Bliley Act (GLBA).

Computer operating environment, hardware components and peripherals, and infrastructure services, enterprise resource management, customer relations management, Supply chain and manufacturing tasks, application development, information management and access, and tasks performed by both business and technical equipment.

Educational software such as training, Education materials based on E-learning, office productivity suites, file viewers, multimedia players, file viewers, and Web browsers.

Based on these approaches we have categories the domains and collected the 30 software developed projects from different streams, we evaluated the Function Point and match with the actual effort based on Project Kickoff meeting started and the Project Closure submitted.

The data showed that the estimation of the mathematical algorithms implemented values is on a par in

accuracy, while there is still significant room for improvement in order to better address the prediction challenges faced in practice.

PROBLEM STATEMENT OF THE EXISTING MODELS

From the existing process of the cost estimation models, surveys and studies have attempted to evaluate the cost estimation models.

Unfortunately, the results are not encouraging and having a lot of variation from developed cost or effort as compare to estimated cost calculated at the time of Requirement engineering, as many of them were found to be not very accurate. Based on Kemerer empirical validation of four algorithmic models (SLIM, COCOMO, Estimacs, and FPA) [10]. Most models showed a strong over estimation bias and large estimation errors, ranging from a Mean Absolute Relative Error (MARE) of 57% to 800%.

Vicinanza, Mukhopadhyay and Prietula used experts process to estimate the project cost using Kemerer's data set without formal algorithmic techniques and found the results outperformed the models in the original study [11]. However, the Mean Absolute Relative Error (MARE) ranges from 32 to 1107%.

Ferens and Gurner evaluated three models (SPANS, Checkpoint, and COSTAR) using 22 projects from Albrecht's database and 14 projects from Kemerer's data set [12]. The estimation error is also large, with Mean Absolute Relative Error (MARE) ranging from 46% for the Checkpoint model to 105% for the COSTAR model.

Jeffery and Low investigated the need for model calibration at both the industry and organization levels [16]. the estimation error was large, with Mean Absolute Relative Error (MARE) ranging from 43% to 105%.

Jeffery, Low and Barnes later compared the SPQR/20 model to FPA using data from 64 projects from a single organization [18].

The models were recalibrated to the local environment to remove estimation biases. Improvement in the estimate was observed with a Mean Absolute Relative Error (MARE) of 12%, reflecting the benefits of model calibration.

Shepperd and Schofield found that estimating by analogy outperformed estimation based on statistically derived algorithms [19].

Heemstra surveyed 364 organizations and found that only 51 used models to estimate effort and that the model users made no better estimate than the non-model users [15]. Also, use of estimation models was no better than expert judgment.

A Report of software development within JPL found that only 7% of estimators use algorithmic models as a primary approach of estimation [17].

Researchers have attempted different approaches for Cost estimation. Models based on artificial intelligence have been developed.

For example, Finnie and Wittig applied artificial neural networks (ANN) and case-based reasoning (CBR) to estimation of effort [13].

Using a data set from the Australian Software Metrics Association, ANN was able to estimate development effort within 25% of the actual effort in more than 75% of the projects, and with a Mean Absolute Relative Error (MARE) of less than 25%.

However, the results from CBR were less encouraging. In 73% of the cases, the estimates were within 50% of the actual effort, and for 53% of the cases, the estimates were within 25% of the actual.

Mukhopadhyay, Vicinanza and Prietula found that an expert system based on analogical reasoning outperformed other methods [21].

Srinivasan and Fisher used machine learning approaches based on regression trees and neural networks to estimate costs [22]. The learning approaches were found to be competitive with SLIM, COCOMO, and function points, compared to the previous study by Kemerer [10].

Briand, El Eman, and Bomarius proposed a hybrid cost modeling method, COBRA: Cost estimation, Benchmarking and Risk Analysis [20].

This method was based on expert knowledge and quantitative project data from a small number of projects. Encouraging results were reported on a small data set.

PROCESS MODEL OF THE ESTIMATION

The assertion of this work is that estimation of the amount of effort required for the development of a project. Software cost estimation process is a set of techniques and procedures that is used to derive the software cost estimate. A set of inputs to the process and then the process uses these inputs to generate or calculate a set of outputs.

Software projects are notorious for going past their deadline, going over budget, or both. Studies within the last few years have shown that a great deal more money is often spent on projects than is initially anticipated. IBM's Consulting Group did a survey of 24 leading companies in 1994 and found that 55% of the software developed cost more than the initial cost estimates. The Standish Group also did a study in 1994 of 8,380 projects in the United States and found that 53% of the software projects that were completed cost 189% of the original estimates. Although these numbers are from a few years ago, they likely have not changed much recently.

The problem of making accurate cost estimates can be attributed to a number of reasons:

1. The wrong cost estimation processes may be used,

2. No base line processes may be used

3. The nature of the problem may not allow for accurate cost estimation.

4. Cost estimation is an often overlooked project management practice.

The cost estimation is usually dependent upon the size estimate of the project, which may use lines of code or function points as metrics.

There are several different techniques for performing software cost estimation, including expert judgment and algorithmic models.

Based on my 10 years of corporate knowledge as a consultant in the GE, we have analyze all the process and come to know that the System will use the Function Point for cost estimation but have to follow the detail analysis and the process as per expert judgment and heuristic knowledge.

Estimation by expert judgment is a common way of estimating the effort required for a project. Unfortunately, this method of estimation does not emphasize re-estimation during the project life cycle, which is an important part of project tracking, because it allows the estimates to be improved during the project life cycle. The quality of a cost estimation model is not so much attributed to the initial estimate, but rather the speed at which the estimates converges to the actual cost of the project. COCOMO is a popular algorithmic model for cost estimation whose cost factors can be tailored to the individual development environment, which is important for the accuracy of the cost estimates. More than one method of cost estimation should be done so that there is some comparison available for the estimates. This is especially important for unique projects. Cost estimation must be done more diligently throughout the project life cycle so that in the future there are fewer surprises and unforeseen delays in the release of a product.

Project estimation process map:
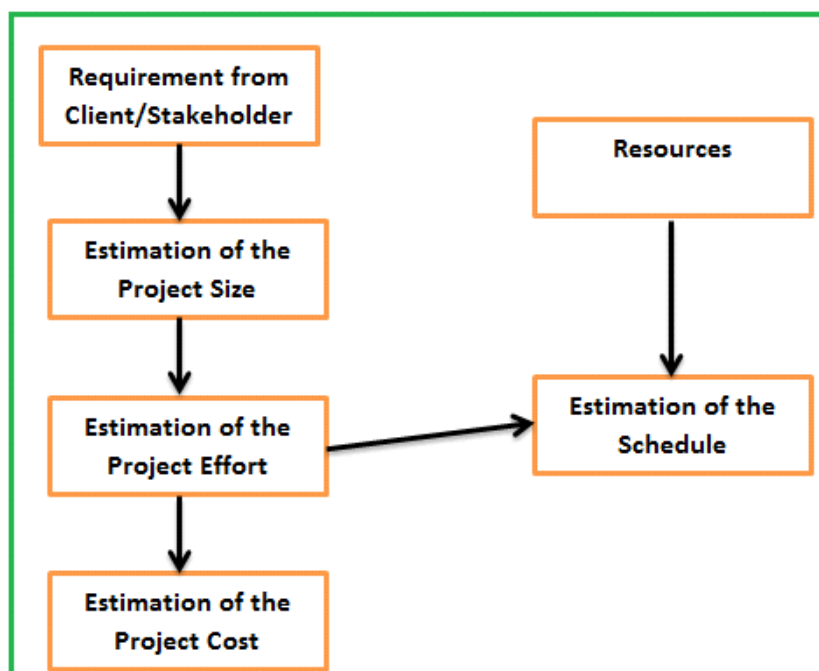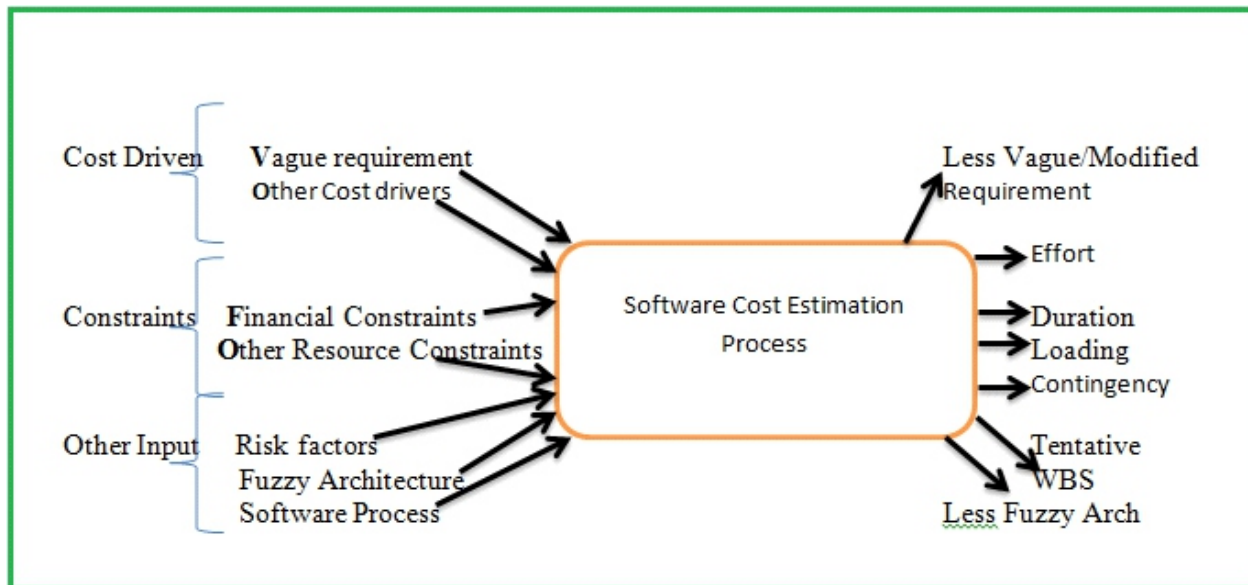
Estimation is an integral part of the software development process and should not be taken lightly. A well planned and well estimated project is likely to be completed in time. Incomplete and inaccurate documentation may pose serious hurdles to the success of a software project during development and implementation. Software cost estimation is an important part of the software development process. Metrics are important tools to measure software product and process. Metrics are to be selected carefully so that they provide a measure for the intended process/product. Models are used to represent the relationship between effort and a primary cost factor such as software product size. Cost drivers are used to adjust the preliminary estimate provided by the primary cost factor. Models have been developed to predict software cost based on empirical data available, but many suffer from some common problems. The structure of most models is based on empirical results rather than theory. Models are often complex and rely heavily on size estimation. Despite problems, models are still important to the software development process. A model can be used most effectively to supplement and corroborate other methods of estimation.

1.1. Software Project Estimation

The four basic steps in software project estimation are:

1) Estimate the size of the development product. This generally ends up in either Lines of Code (LOC) or Function Points (FP), but there are other possible units of measure. A discussion of the pros & cons of each is discussed in some of the material referenced at the end of this report.

2) Estimate the effort in person-months or person-hours.

3) Estimate the schedule in calendar months.

4) Estimate the project cost in dollars (or local currency)

1.2. Estimating size

An accurate estimate of the size of the software to be built is the first step to an effective estimate. Your source(s) of information regarding the scope of the project should, wherever possible, start with formal descriptions of the requirements - for example, a customer's requirements specification or request for proposal, a system specification, a software requirements specification. If you are [re-]estimating a project in later phases of the project's lifecycle, design documents can be used to provide additional detail. Don't let the lack of a formal scope specification stop you from doing an initial project estimate. A verbal description or a whiteboard outline are sometimes all you have to start with. In any case, you must communicate the level of risk and uncertainty in an estimate to all concerned and you must re-estimate the project as soon as more scope information is determined.

Two main ways you can estimate product size are:

1) By analogy. Having done a similar project in the past and knowing its size, you estimate each major piece of the new project as a percentage of the size of a similar piece of the previous project. Estimate the total size of the new project by adding up the estimated sizes of each of the pieces. An experienced estimator can produce reasonably good size estimates by analogy if accurate size values are available for the previous project and if the new project is sufficiently similar to the previous one.

2) By counting product features and using an algorithmic approach such as Function Points to convert the count into an estimate of size. Macro-level "product features" may include the number of subsystems, classes/modules, methods/functions. More detailed "product features" may include the number of screens, dialogs, files, database tables, reports, messages, and so on.

1.3. Estimating effort

Once you have an estimate of the size of your product, you can derive the effort estimate. This conversion from software size to total project effort can only be done if you have a defined software development lifecycle and development process that you follow to specify, design, develop, and test the software. A software development project involves far more than simply coding the software – in fact, coding is often the smallest part of the overall effort. Writing and reviewing documentation, implementing prototypes, designing the deliverables, and reviewing and testing the code take up the larger portion of overall project effort. The project effort estimate requires you to identify and estimate, and then sum up all the activities you must perform to build a product of the estimated size.

There are two main ways to derive effort from size:

1) The best way is to use your organization's own historical data to determine how much effort previous projects of the estimated size have taken. This, of course, assumes

(a) Your organization has been documenting actual results from previous projects

(b) That you have at least one past project of similar size (it is even better if you have several projects of similar size as this reinforces that you consistently need a certain level of effort to develop projects of a given size)

(c) That you will follow a similar development lifecycle, use a similar development methodology, use similar tools, and use a team with similar skills and experience for the new project.

2) If you don't have historical data from your own organization because you haven't started collecting it yet or because your new project is very different in one or more key aspects, you can use a mature and generally accepted algorithmic approach such as Barry Boehm's COCOMO model or the Putnam Methodology to convert a size estimate into an effort estimate. These models have been derived by studying a significant number of completed projects from various organizations to see how their project sizes mapped into total project effort. These "industry data" models may not be as accurate as your own historical data, but they can give you useful ballpark effort estimates.

1.4. Estimating schedule

The third step in estimating a software development project is to determine the project schedule from the effort estimate. This generally involves estimating the number of people who will work on the project, what they will work on (the Work Breakdown Structure), when they will start working on the project and when they will finish (this is the "staffing profile"). Once you have this information, you need to lay it out into a calendar schedule. Again, historical data from your organization's past projects or industry data models can be used to predict the number of people you will need for a project of a given size and how work can be broken down into a schedule.

If you have nothing else, a schedule estimation rule of thumb [McConnell 1996] can be used to get a rough idea of the total calendar time required:

Schedule in months = 3.0 * (effort-months) 1/3

Opinions vary as to whether 2.0 or 2.5 or even 4.0 should be used in place of the 3.0 value – only by trying it out will you see what works for you.

1.5. Estimating Cost

There are many factors to consider when estimating the total cost of a project. These include labor, hardware and software purchases or rentals, travel for meeting or testing purposes, telecommunications (e.g., long distance phone calls, video-conferences, dedicated lines for testing, etc.), training courses, office space, and so on.

Exactly how you estimate total project cost will depend on how your organization allocates costs. Some costs may not be allocated to individual projects and may be taken care of by adding an overhead value to labor rates ($ per hour). Often, a software development project manager will only estimate the labor cost and identify any additional project costs not considered "overhead" by the organization.

The simplest labor cost can be obtained by multiplying the project's effort estimate (in hours) by a general labor rate ($ per hour). A more accurate labor cost would result from using a specific labor rate for each staff position (e.g., Technical, QA, Project Management, Documentation, Support, etc.). You would have to determine what percentage of total project effort should be allocated to each position. Again, historical data or industry data models can help.
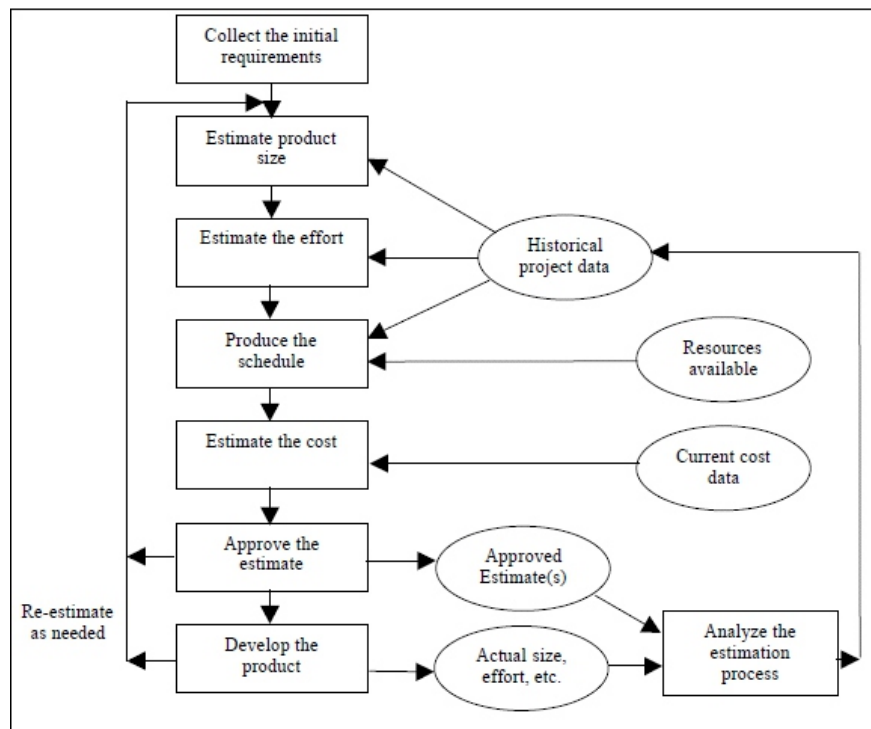
CONCLUSION

Today, almost no model can estimate the cost of software with a high degree of accuracy. This state of the practice is created because

(1) there are a large number of interrelated factors that influence the software development process of a given development team and a large number of project attributes, such as number of user screens, volatility of system requirements and the use of reusable software components.

(2) The development environment is evolving continuously.

(3) The lack of measurement that truly reflects the complexity of a software system. To produce a better estimate, we must improve our understanding of these project attributes and their causal relationships, model the impact of evolving environment, and develop effective ways of measuring software complexity.

Based on these approaches we have categories the do Requirements and collected the 30 software developed projects from different streams, we evaluated the Function Point and match with the actual effort based on Project Kickoff meeting started and the Project Closure submitted. Function Point methodology with an enhanced approach, focus on type of the domain of the Software Projects. The assessment of a Project based on Functional and Non Functional Requirements should be the first step for a Project manager before going to find the Cost or effort of the project. With new types of applications, new development paradigms and new development tools, cost estimators are facing great challenges in applying known estimation models in the new millenium. Historical data may prove to be irrelevant for the future projects. The search for reliable, accurate and low cost estimation methods must continue. Several areas are in need of immediate attention. For example, we need models for development based on formal methods, or iterative software process.

REFERENCES

1. Mohd Sadiq , S. K. Jain, *"Stakeholder Identification Method in Goal Oriented Requirements Elicitation Process"*, 2014 IEEE.

2. Mohd Sadiq , Tanveer Hassan, *"An Extended Adaptive Software Development Process Model"*, 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)

3. Goether, Wolfhart B., Elizabeth K. Bailey, Mary B. Busby, *Software Effort and Schedule Measurement: A framework for counting Staff-hours and reporting Schedule Information*, CMU/SEI-92-TR-021, 1992

4. Humphrey, Watts, *A Discipline for Software Engineering*, Addison-Wesley, 1995

5. Dreger, Brian, *Function Point Analysis*, Prentice-Hall, 1989

6. Garmus, David & David Herron, *Measuring the Software Process*, Yourdon Press, 1996

7. Jones, Capers, *Applied Software Measurement: Assuring Productivity and Quality*, McGraw-Hill, 1991

8. Jones, Capers, *What are Function Points*, 1997, http://www.spr.com/library/0funcmet.htm

9. International Function Point Users Group (IFPUG) web site: http://www.ifpug.org A function point FAQ: http://ourworld.compuserve.com/homepages/softcomp/fpfaq.htm

10. C.F. Kemerer, *"An empirical validation of software cost estimation models"*, Communications of the ACM, vol. 30, no. 5, May 1987, pp. 416-429.

11. S. S. Vicinanza, T. Mukhopadhyay, and M. J. Prietula, *"Software-effort estimation: anexploratory study of expert performance"*, Information Systems Research, vol. 2, no. 4, Dec.1991, pp. 243-262.

12. D. V. Ferens, and R. B. Gumer, *"An evaluation of three function point models of estimation of software effort"*, IEEE National Aerospace and Electronics Conference, vol. 2, 1992, pp. 625- 642.

13. G. R. Finnie, G. E. Wittig, *AI tools for software development effort estimation*, Software Engineering and Education and Practice Conference, IEEE Computer Society Press, pp. 346-353, 1996.

14. DeMarco, Tom, *Controlling Software Projects*, Prentice-Hall, 1982

15. F. J. Heemstra, *"Software cost estimation"*, Information and Software Technology, vol. 34, no.

16. D. R. Jeffery, and G. C. Low, *"Calibrating estimation tools for software development"*, Software Engineering Journal, July 1990, pp. 215-221.

17. J. Hihn and H. Habib-Agahi, *"Cost estimation of software intensive projects: a survey of current practices"*, International Conference on Software Engineering, 1991, pp. 276-287.

18. D. R. Jeffery, G. C. Low, and M. Barnes, *"A comparison of function point counting techniques"*, IEEE Trans on Soft. Eng., vol. 19, no. 5, 1993, pp. 529-532.

19. M. Shepperd and C. Schofield, *"Estimating software project effort using analogy"*, IEEE Trans. Soft. Eng. SE-23:12, 1997, pp. 736-743.

20. L. C. Briand, K. El Eman, F. Bomarius, *"COBRA: A hybrid method for software cost estimation, benchmarking, and risk assessment"*, International conference on software engineering, 1998, pp. 390-399.

21. T. Mukhopadhyay, S. S. Vicinanza, and M. J. Prietula, *"Examining the feasibility of a case based reasoning model for software effort estimation"*, MIS Quarterly, vol. 16, no. 2, June 1992, pp. 155-172.

22. K. Srinivasan and D. Fisher, *"Machine learning approaches to estimating software development effort"*, IEEE Trans. Soft. Eng., vol. 21, no. 2, Feb. 1995, pp. 126-137.

Design and Analysis of White Hat – Using Seo Techniques

Dr. Jitender Rai

(Associate Professor), Dept. of Computer Science, Tecnia Institute of Advanced studies, Rohini, (GGSIPU) Delhi Email:jitender12rai@gmail.com,

A B S T R A C T

Search Engine Optimization (SEO) becomes an integral part for online marketing and can be a crucial parameter for influencing success for any business. As such, Business people and organizations are seeking new technology and strategies to better their SEO efforts and to achieve the desired results as swiftly as possible. The purpose of SEO is to improve the organic ranking of a website with leading search engines. This article introduces and discusses the concept of white hat search engine optimization (SEO). This research works on the latest algorithms of search engines especially for Google without getting sand boxed. Techniques used in this article are safe and practically possible to increase the search traffic. This is done by "On page" and "Off page" Search engine optimization techniques [1].

General terms:- SEO (Search Engine optimization), traffic, on page, off page,

Keywords: White hat SEO; web crawler; search results; penguin, hummingbird; on page SEO: keywords, internal links, structure, HTML mark-ups, SEO elements; off page SEO: external links, social signals.

I. INTRODUCTION

Search engine optimization is the process by which one can increase the visibility of website or webpage in search results [1] in order to increase the visitors of that website or on the particular page [2]. Purpose of this research is to get higher ranking in search results using white hat SEO techniques.

Working of Search Engines

When any user searches phrase or a word then the search engine generate the list of website links where that particular phrase or word were found [3]**.** In order to show the list of websites in search results the search engine perform crawling, indexing, processing, ordering and then retrieving the results[3]. Search engines use "Web Crawlers" that crawls the webpages of websites, and index it on the basis of certain algorithm. Like Google uses Googlebot [4] for crawling the webpages, and uses algorithm like Panda, Penguin, Hummingbird and Pigeon for processing the results by removing spam results like a website that contain no relevant information but due to malpractices it comes in results, And arranging them according to relevancy criteria. The latest algorithm is introduced by Google is Pigeon [5] in 22nd July 2014 to increase the visibility of local listing in search results[6]. The whole story is that the algorithm is continuously updating to decrease the spam results which makes difficult to establish the ranking of website.
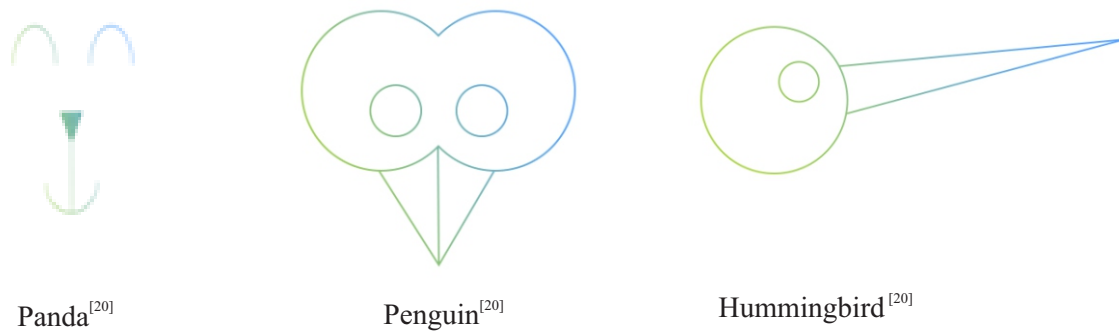
Panda[20]   Penguin[20]   Hummingbird[20]

FIGURE-1

Search Results

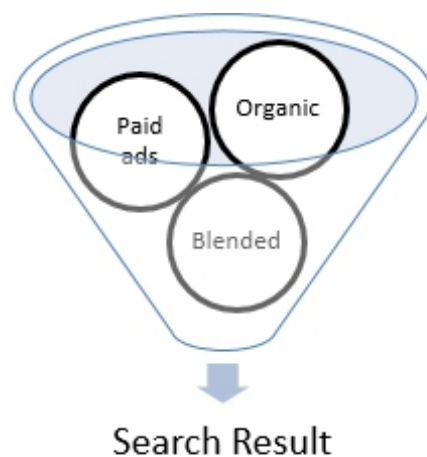The search results are organized in three different categories i.e.,

Search Result

FIGURE-2

1. **Organic results**[7]: The results that are come from search engine ranking algorithm. This type of results is also called natural results. This is the area where search engine optimization techniques are applied. In this article we primarily discuss about organic result.

2. Paid ads[7]: These are the paid results that often come on the top of the search results. In this area you have full control of content of that ad.

3. **Blended results**: These are the videos, images, maps, locations, news, directions, movie, sports, ecommerce, local business, public transport, etc. These types of results are automatically generated by search engines on the basis of the location of searcher. For example: Type "hotels" in Google, if you are in India than the search engine show hotel in India. These types of search contain local listing.
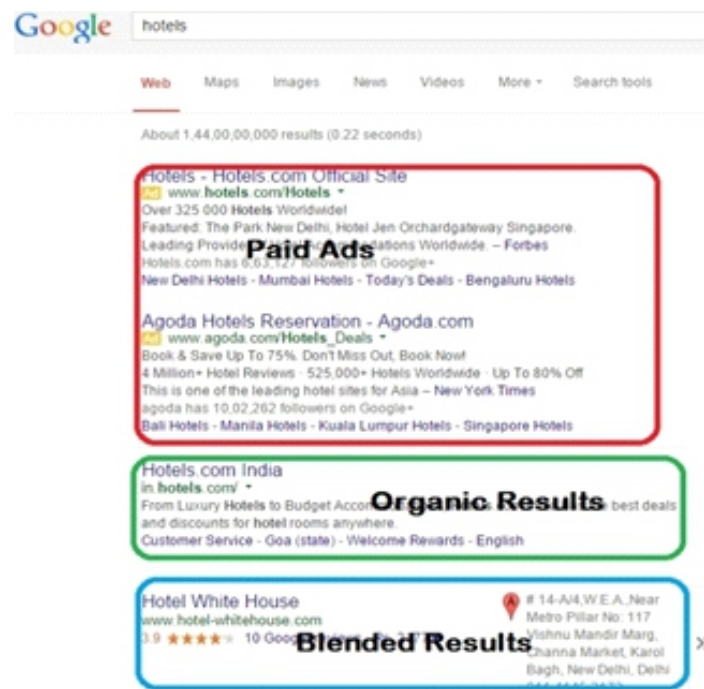
FIGURE-3

II. BACKGROUND

In the mid of 1990 the internet has had a revolutionary impact on the culture and commerce. In 1993 when there is about 600 of websites the first directory introduced by O'Reilly & Associates i.e, GNA[8] (Global Network Navigator),In 1994 "**yahoo and WebCrawler**", then in 1995 "**Altavista**"[9] one that show outline of content in webpage. Afterwards in 1998 "**Google**" [10] by Larry Page and Sergey Brin were the most successful among these. There are over 1,195,768,082(approx.) websites at the time and to deal with such a large number Google introduces many algorithms in order to give user better and fast results. Major search results are affected by the "**panda**"[11] algorithm introduced in 2011 that emphasis more on the content and the simplicity of website rather than backlinks, and then comes "**Penguin**"[12] in 2012 that affect the visibility of various website that violates the Webmaster Guideline. In 2013,"**Hummingbird**" [14] algorithm is introduced which more emphasis on the backlinks, age of website and webmaster Guidelines. In 2014,"**Pigeon**" which emphasis on increasing the ranking of local listing in a search[15]. So there is peak to peak competition to get better ranking.

III. SEO TECHNIQUES

Following types of techniques are used for ranking, i.e.

White Hat SEO: Use techniques which follow the rules and guidelines that are given by search engines. These methods may not harm the ranking of a website. These techniques are long lasting not easily sandboxed. These search engine optimization techniques may not suddenly increase the ranking of website or webpage but effect of this may appear slowly. This technique includes original content writing with high quality keywords[16], important html meta tags, heading tags, inbound links or internal links that help search engine to recognize the structure of the website.

Black Hat SEO : Use techniques that challenge or abolish the rules and guidelines of search engines [1]. These techniques are comparatively takes little time than white hat SEO techniques and gave higher results also but now search engines are very much aware of such scam and fraud website they caught you easily and penalize you. These methods may cause sudden increase your rank in search results but as per consequences you were easily penalized by search engines. It may harm the ranking of your website because of unusual behaviour of website. It includes auto generated pages, fraud backlinks creator, keyword stuffing[1], hidden text etc.

Grey Hat SEO: This is in between of white hat and black hat SEO techniques [2]. But these techniques may not create much impact on your ranking and also may sandbox by search engine.

IV. PROCESS OF SEO

Before starting the process of SEO first thing keep in your mind is that search engines are not human. The first step of Search Engine Optimization is "Competitor Analysis".

Competitor Analysis: As there is more than 1,195,768,082 websites to compete with such large number we have to analyse our competitor. Find the weakness of your competitor and work accordingly to get high ranking. For example, if you want your website on the top of the results by typing "hotels in delhi" then there is more than lakhs of results and to get on the top you have to analysis the strength and weakness of competitor in terms of number of backlinks, content ratio, keywords, and age of website. There are various tools available in internet like Woorank, Majestic, Alexa etc., that gave you some idea about the strategy of your competitor.

In order to improve the ranking of website we take both on page and off page optimization into consideration:

A. On Page SEO[17]

1. **Content**: Content of webpage should be 65% to 75% unique and in high quality, And the text to html ration is in between 25% - 75%. Search engines crawlers visit the site and create a copy of content in their database, if they found any other webpage containing the same content then that webpage penalized.

2. **Page Structure**: Like a newspaper webpage structure containing all the important HTML tags like title, heading (h1, h2,.., h6), bold, strong etc.

Example:

```
<TITLE>title</TITLE>
<H1>heading</H1>,
<STRONG>important keyword</STRONG>
```

3. **Keyword Research**: Now days search engines like Google doesn't gave extra attention to keywords, So Keyword research become a very crucial task that tells the starting point of webpage. A good keyword is that match with the content of page. There are lots of free tools available online like Google Adword, Bing.

Process of SEO

FIGURE-4

4. **Alt tags**: Alt tag is very important it tells the search engine content behind the image when the move the cursor over the image [1].

Example:

<IMG TITLE="image" SRC="image.jpg" ALT="image" />

5. **Meta Tags**: Use of meta tags like description, robots, keywords etc., is very important because it tells the search engine what type of content is present in the webpage. Meta description should be in between 70 – 160 characters. Example:

<META NAME="description" CONTENT=" "/>: description tag that tells the content of particular webpage.

<META NAME="keyword" CONTENT=" "/>: keyword tag tells the search engine what type of content is based on particular word.

6. **Robots meta tag**: The robots meta tag tells search engine how to an individual page should be indexed and served to users in search results [21].

Example:

<META NAME="ROBOTS" CONTENT = "NOINDEX"> : Tells search engine to not to index particular page by not showing the link on result page.

<META NAME="ROBOTS" CONTENT="INDEX"> : Tells search engine to index particular page by showing the link on result page.

Note: By default CONTENT="INDEX,FOLLOW".

7. **Internal Links**: Internal link are also dominant to tell the structure of website to search engine . Internal links are the links that point to your website from other website. They are useful for establishing site architecture and spreading link juice.
Example:
<A HREF="http://www.same-domain.com/" TITLE="Keyword Text">Keyword Text</A>

8. **Sitemaps**: A Sitemap is list of internal URLs (Unified Resource Locator) of website in XML (Extensible Markup Language) [18]. It used by webmaster tool to get the important information about the URL of website. But there is no surety that an XML Sitemap will get your pages crawled and indexed by search engines but having one certainly increases your probability [19].

Example:

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

<url>

<loc>http://www.site.com/saleyourbooks.html</loc>

<changefreq>weekly</changefreq>

</url>

</urlset>

B. Off Page SEO

9. **Backlinks**: : Backlinks are the links on the other website that are pointing to your website. Links form other websites help the search engine to determine website is relevant or not. Building links is not an easy task because fraud backlinks doesn't increase your ranking. Now the search engine very much aware of fraud backlinks. According our research if one website have very small number of backlinks and other have large, then it is not necessary that the large backlinks website rank well. If a website have small number of quality of backlinks compare to other website which having large number of poor quality backlinks then in this case the website having good quality of backlinks may rank well. To create good quality of backlinks make a list of all available blogs of websites you have than create a page on it that tells about the website that you want to create backlink for. In this webpage make links that point to your website url.

For Example:

The world best hotels.

And to check the backlinks there are many tools are available online like Majestic, Ahrefs etc.

10. **Social Sites**: Social sites play vital role in attracting visitors like: Facebook page, Google plus etc. And some of them support "follow links" for creating backlinks like Google plus. The site that support "follow links" may create good quality of backlinks.

11. **Local Search**: As Pigeon algorithm of Google more emphasis on local listing that allow users to submit geographically constrained searches against a structured database of local business listings. These type of search results are also called blended results.

12. **Webmaster tools**: These are the tools that help to detect the problems and the errors of website. These tools tells how the results in the search engines are appears and how, which, when, from where the visitor interact with your website. For example Google webmaster tools that helps to access search statistics on Google, tells latest data regarding incoming and internal links, sends the notification if your site has any crawling error or malware and shows top queries and impressions.

V.   EVALUATION OF WHITE HAT AND BLACK HAT TECHNIQUE

Parameter	White Hat	Black Hat
Time consumption	More	Less
Efficiency	More efficient	Less efficient method easily get penalize
Effect come into existence	Take long time	Take lesser time
Rules and Guidelines	Follow	Abolish
Properties	Original content writing with high quality keywords [16], important html meta tags, heading tags, inbound links or internal links that help search engine to recognize the structure of the website.	Auto generated pages, fraud backlinks creator, keyword stuffing [1], hidden text etc.

Table-1

FIGURE-5

APPLICATIONS

Search Engines are unique in a way since they provide targeted traffic—people, who are looking for what you offer and Search Engines, provide a smoother road way to make it happen. SE makes any business successful by reaching more number of people.

CONCLUSION

It is very common for layman user to access any websites through search engines, so ranking of a website is very much essential for directing more traffic towards any particular website. The higher the result of any website in organic result of any search operation, the more chances that user will view that website.

White hat SEO ensures the use of techniques and strategies which targets human audience rather than search engines. It uses meta tags, keywords and their analysis, links and content for human readers.

FUTURE SCOPE

When any user performs any search at any search engines, every one demands fast result. Even 1 or 2 seconds delay can cause dissatisfaction among users, so most of the search engine tries to produce results faster. When a user searches on SE, besides it requires SE to crawl and index billions of documents and two parameters namely relevance and importance is very important to influence SE. There is a need of some new strategies or hybrid method to make the entire process much faster than existing technique.

REFERENCES

- *Gurpreet Singh Bedi, Ms.Ashima Singh "Analysis of Search Engine Optimization (SEO) Techniques", in*

- *International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 3, March 2014 page 563*

- *Search engine optimization" available electronicallyat http://en.wikipedia.org/wiki/Search_engine_optimization*

- *Dr. KhannaSamratVivekanandOmprakash "CONCEPT OF SEARCH ENGINEOPTIMIZATION IN WEB SEARCH ENGINE", in International Journal of Advanced Engineering Research and Studies IJAERS/Vol. I/ Issue I/October-December, 2011/235- EXPLAINED" available electronically athttp://www.icrossing.com/sites/default/files/Google-Hummingbird-Explained-iCrossing-POV.*

- *"Crawling &Inexing" available electronically at http://www.google.co.in/insidesearch/howsearchworks/crawling-indexing.html*

- *" Google Algorithm Change History" available electronically at http://moz.com/google-algorithm-change*

- *"The technology behind Google's great results" available electronically at http://www.google.co.in/techn-ology/pigeonrank.html*

- *Cristian Danescu-Niculescu-Mizil‡1, Andrei Z. Broder†, EvgeniyGabrilovich†, VanjaJosifovski†, Bo Pang† "Competing for Users' Attention: On the Interplay between Organic and Sponsored Search Results"*

- *"Global Network Navigator" available electronically at http://en.wikipedia.org/wiki/Global_Network_Navigator*

- *" Altavista" available electronically at http://en.wikipedia.org/wiki/AltaVista*

- *"Google" available electronically at http://en.wikipedia.org/wiki/Google*

- *AntrikshaSoni, UgrasenSuman, "Counter Measures against Evolving Search Engine Spamming Techniques," in IEEE, 2011. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5942084*

- *"Google Penguin," available electronically at http://www.wordtracker.com/blog/google-penguin-what-itwas-and-what-it-wasnt*

- *David Shapiro, Doug Platts, Magico Martinez "ICROSSING POV: GOOGLE HUMMINGBIRD*

- *"The technology behind Google's great results" available electronically at http://www.google.co.in/technology/pige-onrank.html*

- *Fuxue Wang, Yi Li, Yiwen Zhang, "An empirical study on the search engine optimization technique and its outcomes," in Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Aug 2011.*

- *Dr S. Saravanakumar, K Ramnath, R Ranjitha and V.G.Gokul "A New Methodology for Search Engine Optimization without getting Sandboxed", in International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012 page 472.*

- *" Sitemaps" available electronically at http://en.wikipedia.org/wiki/Sitemaps*

- *"All About XML Sitemaps" available electronically at http://blog.woorank.com/2013/03/all-about-xml-sitemaps/*

- *"Google Penguin, Panda, & Hummingbird" available electronically at http://www.searchenginejournal.com/seo-guide/google-penguin-panda-hummingbird/*

- *"Robots meta tag and X-Robots-Tag HTTP header specifications" available electronically at https://developers.go-ogle.com/webmasters/control-crawl-index/docs/robots_meta_tag*

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);

2. Informative contribution (editorial, commentary, etc.);

3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.

The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005

Notes: