Measurement and defect modeling for a legacy software system

Joel Troster and Jeff Tian
IBM Software Solutions Toronto Laboratory, Mail Station 4T/881,
1150 Eglington Avenue East, North York, Ontario, Canada M3C 1H7

This paper analyzes the quality of a large-scale legacy software system using selected metrics. Quality measurements include defect information collected during product development and in-field operation. Other software metrics include measurements on various product and process attributes, including design, size, change, and complexity. Preliminary analyses revealed the high degree of skew in our data and a weak correlation between defects and software metrics. Tree-based models were then used to uncover relationships between defects and software metrics, and to identify high-defect modules together with their associated measurement characteristics. As results presented in tree forms are natural to the decision process and are easy to understand, tree-based modeling is shown to be suitable for change solicitation and useful in guiding remedial actions for quality improvement.

Keywords: Legacy software, software metrics, defect, data visualization, tree-based modeling.

1. INTRODUCTION

In today's competitive environment for software products, quality has become an increasingly important asset to software development organizations. To ensure product quality, various quality indicators have to be measured and monitored. In addition, various product attributes that can be linked to quality are often measured, in the hope that product quality can be improved by adjusting and controlling such product attributes. As a result, massive amounts of measurement data have been

*The following are trademarks or service marks in the United States and/or other countries:
  - IBM is a trademark of International Business Machines Corporation.
  - REFINE and Software Refinery are trademarks of Reasoning Systems Inc.
  - SAS is a trademark of the SAS Institute Inc.
  - S-PLUS is a trademark of the Statistical Sciences, Inc.

†E-mail: jtroster@vnet.ibm.com
collected in many software development organizations. These efforts are aimed at understanding the relationships between such product metrics and quality, identifying problematic areas, and guiding remedial actions for the product under development as well as for future products.

Recently, much emphasis has been put on improving the software development process, with the assumption that this will lead to improved product quality [Humphrey 1989]. However, a precursor to improved process is an understanding of the dynamics of the current process. Various process and product attributes need to be measured and analyzed, so that appropriate and timely improvement actions can be carried out.

A large proportion of the products developed and supported in the IBM Software Solutions Toronto Laboratory are so-called legacy systems. Such systems evolve over a long period of time from research prototypes and earlier versions. Characteristics of such legacy systems need to be carefully examined so that appropriate metrics can be selected to address the unique problems of assessing such systems. An additional benefit of measuring and analyzing such legacy systems is the improved understanding and control over the continuously evolving development process. With the use of this information, we can better anticipate changes and plan for future product updates and new releases.

In this paper, we study a large-scale legacy software system and attempt to uncover and understand relationships between defects and selected measurements of product and process attributes. Since the ultimate responsibility for process change and product improvement lies with the development teams, we attempt to analyze the data and present the results in a way that is easy to understand and useful to solicit changes. Visualization and correlation analysis were used to examine the trends and patterns in the measurement results. A statistical analysis technique called tree-based modeling [Clark and Pregibon 1992] was used to perform systematic analysis and modeling on the measurement results. This technique presents results in forms similar to decision trees and therefore is easy to interpret and conducive to change solicitation.

Tree-based modeling originated from the search for alternative techniques to classical statistical models (e.g., linear regression) to analyze data in the social sciences, where the data are often characterized by complicated and unexplained interactions and irregularities [Clark and Pregibon 1992]. Recently, this technique has been used with some success in analyzing various software engineering data [Selby and Porter 1988; Porter and Selby 1990; Tian et al. 1992; Tian and Henshaw 1994]. In this paper, we use tree-based models to analyze our measurement data, helping us to identify modules with high defects and their associated measurement characteristics. Modeling results can be used to improve our understanding of the relationships between quality and product characteristics, and to guide our remedial actions focused on those high-defect modules.

The remainder of the paper is structured as follows: Section 2 describes our product characteristics and metrics selection and implementation. Section 3 presents