Visual Data Mining of Novice Programmer Behaviour

Baxolile Mabinya, Charmain Cilliers & Jean Greyling
Department of Computer Science and Information Systems
Nelson Mandela Metropolitan University, PO Box 77000, Port Elizabeth, 6031
Tel: (041) 504 2080, Fax: (041) 504 2831
Email: {Baxolile.Mabinya, Charmain.Cilliers, Jean.Greyling}@nmmu.ac.za

Abstract –The many difficulties encountered by novice programmers while they program have an impact on their programming behaviour. Although it is easy to identify the individual events that novice programmers carry out while they program, it remains a daunting task to identify patterns that constitute behaviour from the events. Data mining provides effective techniques to identify and extract significant patterns from large data sets of event data. Visualisation techniques can present the mined patterns in a way that is easy to understand and interpret. The purpose of this paper is to discuss the possibility of applying effective data mining techniques and algorithms to visually model the behaviour of novice programmers while they learn to program.

Index Terms— novice programmer behaviour, visual data mining, programmer event data

I. Introduction

Students encounter many difficulties while they program in both procedural and object-oriented programming techniques [6]. Novice programmers react to these difficulties by performing a series of actions, resulting in a number of interrelated events.

A large amount of novice programmer events have been collected during the offering of an introductory programming course, where the programming environment used was Delphi.

Due to the volume and the structure of the programmer event data, it is difficult to identify behavioural patterns. A technique is therefore required to extract useful and interesting knowledge from large data sets. The process of data mining could provide effective techniques for uncovering meaningful behavioural patterns in the event data.

Patterns and trends are, however, sometimes difficult to understand and interpret. An effective way to deal with this issue is to visualise the data mining results. Visual data mining tools and techniques are effective in creating visualisations of mining models that discover patterns in data sets [7]

II. Programmer Event Data

Programmer event data are a set of facts that record the actions a programmer undertook while programming. Programmer event data could capture events such as the errors resulting from a compile event, the filename compiled, when the compile took place as well as the source of the compile.

A related study [5] focused on uncovering the behavioural patterns of novice programmers through a study of only their compilation events. The current study, however, also includes events such as the typing and the executing events.

The collected programmer event data consists of 2 tables. The session table captures data about a practical session and consists of 6170 records. The event table records data about the events that particular students performed while programming and consists of 135000 records. Event data was collected over a period of 4 months for 236 students.

Figure 1 shows an extract of data from the event table. The first record illustrates a Compile/Run event where a student encountered 15 errors. The second and third records illustrate the exact characters that a student typed at a specific point in time.

instance_id	type	FName	StartTime	EndTime	NumLines	NumErro	chars
128573	Compile/Run	***	2005/04/14 1	2005/04/14 1	(null)	15	(null)
128619	Typing	***	2005/04/14 1	2005/04/14 1	(null)	(null)	;;bB;REAL R
128672	Typing	***	2005/04/14 1	2005/04/14 1	(null)	(null)	0WRwriteln :
128674	Run	(null)	2005/04/14 1	2005/04/14 1	(null)	(null)	(null)
128675	FileSave	***	2005/04/14 1	(null)	32	(null)	(null)
128676	Compile/Run	***	2005/04/14 1	2005/04/14 1	(null)	15	(null)
128677	Run	(null)	2005/04/14 1	2005/04/14 1	(null)	(null)	(null)

Figure 1: Extract of programmer event data

III. Novice Programmer Behavioural Patterns

Research indicates that novice programmers exhibit numerous behaviours when they program [5]. It is speculated that a number of behavioural patterns could be extracted from the event data. These speculated patterns include:

- Novice programmers tend to compile less as they progress
- If a student's typing event(s) reduce the number of errors after the initial compilation, the student is likely to end up with a running program
- Students who encounter more errors than a specified threshold are likely to end up with an unsuccessful program
- Students who performed similar events for the same practical sessions end up with similar outcomes.

This study attempts to uncover and visualise behavioural patterns such as aforementioned speculated ones.

IV. Data Mining of Novice Programmer Behaviour

The success of the data mining process is highly dependent on the quality and suitability of the data being mined. Time and effort needs to be dedicated to ensuring that the data to be mined is suitable and ready for data mining. Data mining techniques could be effective in uncovering behavioural patterns in the event data. Visualisations can be used to effectively present the mined patterns.

A. Data Preprocessing

Data preprocessing is a technique that is applied prior to mining to consolidate and ensure that the input data is of highest quality [2].

Some of event data that has been collected is incomplete and inconsistent. For example, 28.5% of the practical sessions have no start/end times recorded. Also the ***FileName*** attribute which reveals the exact file that a student has been working on, is erroneous (See Figure 1). Other examples of missing and/or inconsistent data include:

- The combining of the compile and run events into a single event
- The occurrence of erroneous dates
- The inclusion of irrelevant users
- The splitting of typed text into batches of 30 characters

Data cleaning was applied to resolve the identified problems in the event data. That is, the critical data was consolidated by applying specific methods to guarantee the quality and reliability of the data. For example, the batches of characters were combined into their full meaningful text. Insignificant data was either ignored or removed. For example, the non-student users were simply removed from the event data.

B. Data Mining Techniques

Classification and clustering are suitable data mining techniques to uncover the speculated behavioural patterns. These techniques are classified according to the kinds of knowledge to be discovered, the type of databases to be mined and the kinds of methods to be adopted [1].

Classification aims to assign an object into a predetermined distinct category [3]. For example, novice programmers could be classified as good programmers according to the consistency of their behavioural patterns which lead to successful programs. Classification models can be implemented using decision trees, rule induction, regression models and even neural networks. Clustering is the grouping of similar objects into classes [3]. For example, students who follow a similar set of events for the same problem could be grouped together. Partitioning and density based methods are applicable to the clustering technique.

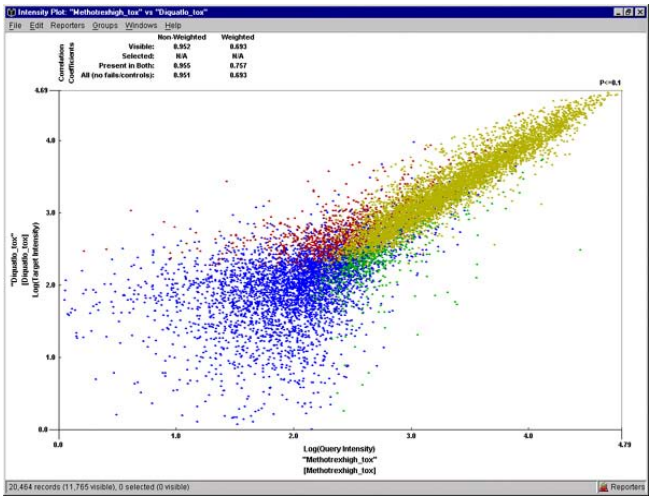

Figure 2:Intensity plot viewer [4]

C. Visualisation Techniques

Patterns are only interesting and useful when they can be interpreted. Visualising the output of a data mining algorithm helps understand the patterns far more succinctly than without any visualisation techniques [8].

Figure 2 is used to compare intensities in two-dimensional data. The technique could perhaps be used to indicate the clusters of students who performed similar events for a particular practical session. It could also be interesting to observe the progress of the different clusters of students over a period of time.

Tree structures could be employed to visualise how novice programmers are classified according to their behaviour (the events they perform while programming).

Depending on the data mining techniques and algorithms as well as the nature of the output produced, a suitable visualisation technique needs to be implemented for effective interpretation of novice programmer behaviour.

V. Conclusion and Future Work

The major motivation behind this research is to gain an understanding of how novice programmers program, what they do wrong and possibly explain why they struggle with programming. The visual data mining model of this project could be used as an analysis tool to identify and understand novice programmer behaviour.

Future work includes the design of the data warehouse, the application of data mining techniques on the event data, as well as the implementation of a suitable visualisation technique to visualise the mined patterns.

Acknowledgements

We would like to thank the Telkom Centre of Excellence and the Department of Computer Science and Information Systems at the Nelson Mandela Metropolitan University for making this research possible.

References

[1] HAN, J. (1996) Data mining techniques. *International Conference on Management of Data: Proceedings of the 1996 ACM SIGMOD international conference on Management of data.* Montreal, Quebec, Canada, ACM Press.

[2] HAN, J. & KAMBER, M. (2001) *Data mining: concepts and techniques,* San Fransisco, Morgan Kaufmann.

[3] HAND, D., MANNILA, H. & SMYTH, P. (2001) *Principles of Data Mining,* Cambridge, Massachusetts, Massachusetts Institute of Technology.

[4] INPHARMATICS, R. (2006) Intensity Plot Viewer. Rosetta Inpharmatics Inc

[5] JADUD, M. C. (2004) A first look at novice compilation behavior using BlueJ. *16th Workshop of the Psychology Interest Group.* Routledge.

[6] PILLAY, N. (2003) Developing Intelligent Programming Tutors for Novice Programmers. *The SIGCSE Bulletin,* 35.

[7] SOUKUP, T. & DAVIDSON, I. (2002) *Visual Data Mining*, John Wiley & Sons, Inc.

[8] WITTEN, I. H. & FRANK, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* San Francisco, Morgan Kaufmann Publishers.

Baxolile Mabinya received his BCom Honours degree (*Cum laude*) in 2005 from the Nelson Mandela Metropolitan University. He is presently doing his MCom in Computer Science and Information Systems at the Nelson Mandela Metropolitan University.