Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00011

CẢI TIẾN TRỌNG SỐ KẾT HỢP KỸ THUẬT RÚT TRÍCH ĐA ĐẶC ĐIỂM TRONG VIỆC DÒ TÌM NHỮNG BÁO CÁO LỖI TRÙNG NHAU

Nhan Minh Phúc[1], Nguyễn Hoàng Duy Thiện[2], Dương Ngọc Vân Khanh[3]

[1,2,3] Khoa Kỹ thuật và Công nghệ, Trường Đại học Trà Vinh

nhanminhphc@tvu.edu.vn, thiennhd@tvu.edu.vn, vankhanh@tvu.edu.vn

TÓM TẮT: *Đối với các phần mềm mở như Firefox, Eclipse, Subversion,... họ thường có hệ thống kho lưu trữ những báo cáo lỗi do người dùng gửi đến. Những báo cáo lỗi này giúp cho hệ thống xác định được những lỗi khác nhau của phần mềm, điều này làm cho việc bảo trì phần mềm tốt hơn. Do số lượng người dùng ngày càng tăng, do đó số lượng báo cáo lỗi được phát hiện cũng ngày càng nhiều. Điều này dẫn đến tình huống có nhiều báo cáo lỗi được gửi đến kho xử lý mà những báo cáo lỗi này đã được những người dùng khác nhau báo cáo trước đó, điều này được gọi là báo cáo lỗi trùng nhau. Để giải quyết vấn đề này, một lập trình viên được phân công phụ trách việc xử lý lỗi cần phải gắn nhãn các báo cáo lỗi này theo cách thủ công dưới dạng các báo cáo lỗi trùng nhau. Tuy nhiên, trong thực tế có quá nhiều báo cáo lỗi trùng được gửi hàng ngày, nếu cứ thực hiện công việc nhận biết thủ công sẽ tốn nhiều thời gian và công sức. Để giải quyết vấn đề này, gần đây, một số kỹ thuật đã được đề xuất để tự động phát hiện các báo cáo lỗi trùng lặp, tuy nhiên kết quả chính xác chỉ chiếm khoảng 36-89 %, lý do vì hai báo cáo của cùng một lỗi có thể được viết theo nhiều cách khác nhau, do đó việc cải tiến về tính chính xác của quá trình phát hiện trùng lặp đang là chủ đề được nhiều sự quan tâm của các nhà nghiên cứu gần đây. Trong bài báo này, chúng tôi giới thiệu một mô hình đa đặc điểm kết hợp với sự cải tiến trọng số từ CFC (Class-Feature-Centroid) để phát hiện các báo cáo lỗi trùng nhau chính xác hơn. Chúng tôi đã tiến hành thực nghiệm trên ba kho phần mềm chứa lỗi lớn từ Firefox, Eclipse và OpenOffice. Kết quả cho thấy rằng kỹ thuật của chúng tôi có thể cải thiện tốt hơn từ 8-11 % khi so với các phương pháp được so sánh.*

Từ khóa*: Duplication detection, bug reports, CFC-27, feature weighting.*

I. GIỚI THIỆU

Do sự phức tạp trong quá trình xây dựng nên hầu hết các phần mềm thường vẫn còn nhiều lỗi sau khi hoàn chỉnh. Những lỗi phần mềm đôi khi dẫn đến thiệt hại nhiều triệu USD [4]. Vì vậy việc xử lý lỗi trở thành một trong những vấn đề quan trọng cần thực hiện thường xuyên trong việc bảo trì phần mềm. Để giúp quản lý lỗi phần mềm và làm cho hệ thống đáng tin cậy hơn, những công cụ quản lý lỗi được xây dựng và ứng dụng vào các hệ thống lớn như Bugzilla, Eclipse,… những công cụ này cho phép người dùng sử dụng phần mềm như "tester" và gửi báo cáo lỗi mà họ phát hiện được đến hệ thống quản lý lỗi, thông tin này sau đó được tiếp nhận và xử lý để hoàn thiện độ tin cậy của phần mềm hơn.

Mặt dù mang lại nhiều lợi ích trong việc cung cấp hệ thống báo cáo lỗi, nhưng nó cũng gây ra nhiều thách thức. Một trong những thách thức đó là cùng một lỗi được phát hiện bởi nhiều người dùng, khi đó có nhiều người gửi cùng một báo cáo lỗi đến hệ thống, gây nên tình trạng gọi là trùng lặp báo cáo lỗi. Điều này làm mất nhiều thời gian và công sức cho người phân loại, nghĩa là khi một báo cáo mới được gửi đến, họ phải kiểm tra xem báo cáo lỗi này đã được gửi đến trước đó chưa. Theo thống kê [2], [3] mỗi ngày có ít nhất 300 báo cáo lỗi được gửi đến hệ thống quản lý lỗi của Mozilla, số lượng này được xem là quá nhiều cho công việc phân loại. Vì vậy việc xây dựng một hệ thống tự động phân chẳng hay một báo cáo lỗi vừa được gửi đến đã được báo cáo trước đó hay chưa. Đây là chủ đề đang được các nhà nghiên cứu quan tâm hiện nay. Để giải quyết vấn đề những báo cáo lỗi trùng nhau, hiện tại trong cộng đồng nghiên cứu có hai hướng. Hướng thứ nhất khi có một báo cáo lỗi mới được gửi đến, sau đó xây dựng mô hình xử lý và kết quả trả về danh sách những báo cáo lỗi gần giống nhất với báo cáo lỗi được gửi đến trong top K. Phương pháp này được công bố bởi [3], [4], [5], [1]. Hướng thứ hai được công bố bởi Jalbert and Weimer [6] như sau, khi có một báo cáo mới được gửi đến, họ sẽ thực hiện việc phân loại thành hai nhóm, trùng nhau hay không trùng nhau, nghiên cứu này còn được gọi phân loại báo cáo lỗi bằng cách gán nhãn những báo cáo lỗi trùng nhau, và những báo cáo lỗi không trùng nhau. Theo thống kê bởi [9] hướng thứ nhất nhận được nhiều sự quan tâm hơn của các nhà nghiên cứu, ly do ngoài việc trả về kết quả top K danh sách báo cáo lỗi trùng nhau, nó gần như bao gồm luôn hướng thứ hai là phân loại báo cáo lỗi. Trong bài báo này chúng tôi cũng nghiên cứu theo phương pháp thứ nhất.

ID:000002; **CreationDate**:Wed Oct 10 20:34:00 CDT 2001; **Reporter**:Andre Weinand
Summary: Opening repository resources doesn't honor type.
Description:Opening repository resource always open the default text editor and doesn't honor any mapping between resource types and editors. As a result it is not possible to view the contents of an image (*.gif file) in a sensible way.

Hình 1. Một báo cáo lỗi trong dataset Eclipse

II. BÁO CÁO LỖI TRÙNG NHAU

Một báo cáo lỗi thông thường là một tập tin bao gồm vài thuộc tính như tóm tắt lỗi (summary), mô tả lỗi (description), dự án (project), người gửi (submitter), bình luận (comment),… Mỗi thuộc tính chứa những thông tin