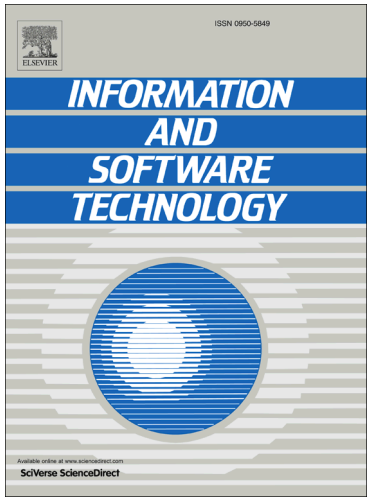

Automated Classification of Software Change Messages by Semi-supervised Latent Dirichlet Allocation

Ying Fu, Meng Yan, Xiaohong Zhang, Ling Xu, Dan Yang, Jeffrey D. Kymer

PII: S0950-5849(14)00134-7
DOI: http://dx.doi.org/10.1016/j.infsof.2014.05.017
Reference: INFSOF 5474

To appear in: *Information and Software Technology*

Received Date: 12 September 2013
Revised Date: 21 May 2014
Accepted Date: 22 May 2014

Please cite this article as: Y. Fu, M. Yan, X. Zhang, L. Xu, D. Yang, J.D. Kymer, Automated Classification of Software Change Messages by Semi-supervised Latent Dirichlet Allocation, *Information and Software Technology* (2014), doi: http://dx.doi.org/10.1016/j.infsof.2014.05.017

Automated Classification of Software Change Messages by Semi-supervised Latent Dirichlet Allocation

Ying Fu, Meng Yan, Xiaohong Zhang[1], Ling Xu, Dan Yang, Jeffrey D. Kymer
School of Software Engineering, Chongqing University, Chongqing (401331), PR China.

Context: Topic models such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) have demonstrated success in mining software repository tasks. Understanding software change messages described by the unstructured nature-language text is one of the fundamental challenges in mining these messages in repositories.
Objective: We seek to present a novel automatic change message classification method characterized by semi-supervised topic semantic analysis.
Method: In this work, we present a Semi-supervised LDA based approach to automatically classify change messages. We use domain knowledge of software changes to make labeled samples which are added to build the semi-supervised LDA model. Next, we verify the cross-project analysis application of our method on three open-source projects. Our method has two advantages over existing software change classification methods: First of all, it mitigates the issue of how to set the appropriate number of latent topics. We do not have to choose the number of latent topics in our method, because it corresponds to the number of class labels. Second, this approach utilizes the information provided by the label samples in the training set.
Results: Our method automatically classified about 85% of the change messages in our experiment and our validation survey showed that 70.56% of the time our automatic classification results were in agreement with developer opinions.
Conclusion: Our approach automatically classifies most of the change messages which record the cause of the software change and the method is applicable to cross-project analysis of software change messages.

KEY WORDS: Software repositories mining; semi-supervised topic modeling; LDA; change message

1. INTRODUCTION

[1] Corresponding author: xhongz@cqu.edu.cn (X. H. Zhang), School of Software Engineering, Chongqing University, Huxi Town, Shapingba, Chongqing, PR. China 410331.