MINERAÇÃO DE REPOSITÓRIOS DE SOFTWARE LIVRE

por **Marco Aurélio Gerosa, Igor Scaliante Wiese, Gustavo Ansaldi Oliva e Maurício Finavaro Aniche**

A MINERAÇÃO DE REPOSITÓRIOS DE SOFTWARE OFERECE UM AMPLO LEQUE DE INFORMAÇÕES SOBRE PROJETOS DE DESENVOLVIMENTO E TEM SE TORNADO UMA IMPORTANTE ÁREA DE PESQUISA, CONTRIBUINDO PARA A MELHORIA DAS ATIVIDADES QUE ENVOLVEM A CONSTRUÇÃO, MANUTENÇÃO E EVOLUÇÃO DE SISTEMAS DE SOFTWARE.

POR QUE UM GRANDE SUPERMERCADO resolve abrir uma filial na cidade X em vez da cidade Y? As grandes empresas tomam decisões analisando dados históricos coletados de diversas fontes. Projetos de software também têm à disposição uma grande quantidade de dados produzidos durante as atividades de desenvolvimento e de apoio: milhares ou milhões de linhas de código são escritas, defeitos são relatados e resolvidos, discussões técnicas acontecem nas listas de e-mail e gerenciadores de requisitos etc. Analisar tais dados aumenta o entendimento sobre o projeto, os desenvolvedores e os processos que governam a evolução e manutenção do software e possibilita que decisões sejam tomadas de forma mais fundamentada. As análises também possibilitam que aprendamos mais sobre a Engenharia de Software em si.

Nesse contexto, a mineração de repositórios de software é uma área de pesquisa voltada para a recuperação, interligação e análise dos dados históricos produzidos durante o desenvolvimento de software e que estão armazenados em repositórios. Por exemplo, suponha que você é um desenvolvedor e fez uma mudança para uma determinada parte de um sistema. O que mais você tem que mudar? Com base na ideia de que os artefatos que mudaram juntos no passado tendem a mudar juntos no futuro, pesquisadores têm construído ferramentas que auxiliam na propagação de mudanças. Além disso, usa-se também a mineração para entender como o software evoluiu ao longo do tempo, compreender os efeitos da interação social entre os

desenvolvedores no processo de desenvolvimento e prever defeitos e esforço. Há ainda uma série de outras aplicações. Você pode encontrar uma lista mais extensa na principal conferência da área (msrconf.org). A mineração de repositórios de software é um campo de pesquisa relativamente novo e tem atraído o interesse de diversos pesquisadores. Hoje, uma grande parcela dos estudos empíricos em Engenharia de Software envolve algum tipo de mineração de repositórios.

A mineração de repositórios de software é um campo de pesquisa relativamente novo e tem atraído o interesse de diversos pesquisadores.

Projetos de Software Livre e Mineração de Repositórios

Grande parte do sucesso da área de mineração de repositórios deve-se à abundante disponibilização de dados viabilizada pelo movimento de software livre. Projetos de grande visibilidade, como o Linux, a Eclipse IDE e vários outros da Apache Software Foundation e da Free Software Foundation, assim como, mais recentemente, projetos hospedados no GitHub, são frequentemente escolhidos para realização de estudos científicos. O acesso aos dados normalmente é realizado por meio de ferramentas oferecidas por grupos de pesquisa, tais como o Metrics Grimoire (http://metricsgrimoire.github.io/) para coleta de dados de tarefas e modificações de artefatos de diferentes repositórios, ou por meio do uso de APIs oferecidas pelos próprios repositórios, como é o caso do GitHubAPI (https://developer.github.com/v3/). O próprio GitHub oferece acesso a dados de milhões de projetos abertos por meio de um banco de dados BigQuery do Google para facilitar análises de forma aberta (https://www.githubarchive.org/). Vale a pena visitar esse sítio e conferir os ganhadores das competições que têm sido promovidas pelo GitHub e ver bons exemplos de aplicações de mineração de repositórios.

Software Livre tem, portanto, se tornado uma importante fonte de informação e vem contribuindo para a evolução dessa recente linha de pesquisa. Afinal, se os pesquisadores precisam de muitos dados e projetos para conseguir explicar fenômenos da Engenharia de Software, o ecossistema de código aberto se torna uma grande oportunidade para viabilizar essa exploração.

Desafios da Mineração de Repositórios

Ahmed Hassan, no artigo "The Road Ahead for Mining Software Repositories", listou uma série de estudos e desafios enfrentados por pesquisadores de mineração de repositórios. Dentre os desafios, destacam-se a grande quantidade de dados não estruturados e de fontes heterogêneas de informação. Outro ponto de destaque está relacionado à escala. Alinhado ao fenômeno de BigData, estudos atuais têm demandado escalabilidade das técnicas de coleta e análise dos dados para grande quantidade de informações. Outro desafio é facilitar a reprodução das pesquisas realizadas para aplicação em contextos/projetos diferentes.

O desafio é transformar pesquisas empíricas em soluções aplicáveis no dia a dia.

No que diz respeito à aplicação, o assunto tem sido foco de discussões na comunidade. O desafio é transformar pesquisas empíricas em soluções aplicáveis no dia a dia dos engenheiros de software, auxiliando o processo de desenvolvimento do software e no contínuo aumento de qualidade.

Conclusão

Nos últimos anos, a importância da mineração de repositórios tem aumentado com a necessidade das empresas de tornarem-

-se mais orientadas a dados. Gerentes de projetos de software têm interesse em transformar os dados que antes eram somente armazenados nos diferentes repositórios de software em informações relevantes para fundamentar seu processo de tomada de decisão. Do ponto de vista acadêmico, a mineração de repositórios provê dados empíricos para avaliar o uso de técnicas e tecnologias do ponto de vista histórico e vêm sendo cada vez mais usada pelos principais grupos de pesquisa em Engenharia de Software. Dados de projetos de software livre vêm sendo largamente utilizados nesses estudos.

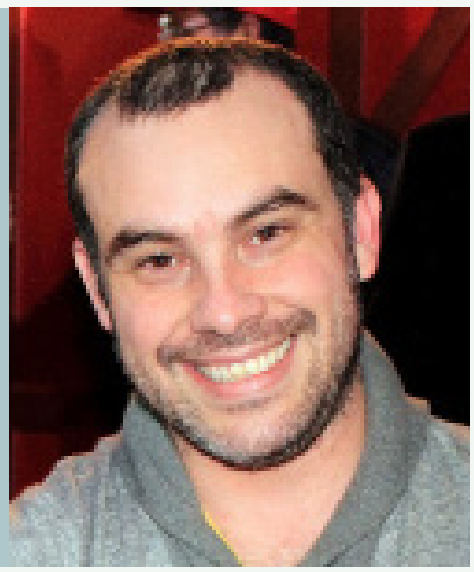
A mineração de repositórios de código veio para ficar e vai ajudar gerentes e desenvolvedores a tomarem decisões no dia a dia.

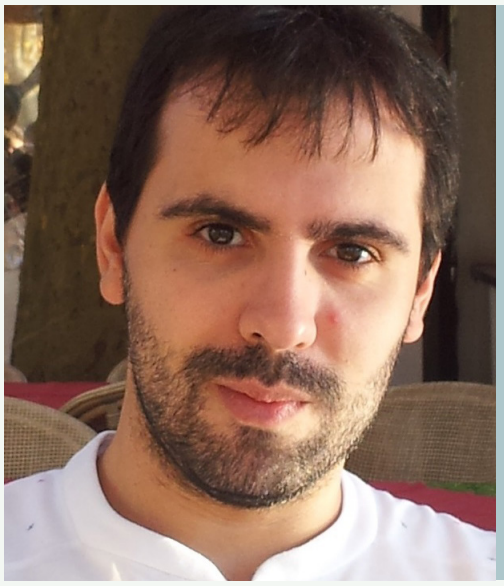
O grupo de pesquisa LAPESSC (http://lapessc.ime.usp.br) do Departamento de Ciência da Computação do IME/USP realiza pesquisas de mineração de repositórios focando em estudos de evolução, manutenção e design de software, dependências de artefatos, métricas de software, análise da colaboração de desenvolvedores e software social. Atualmente o grupo vem desenvolvendo uma ferramenta chamada MetricMiner (http://www.metricminer.org.br), cuja ideia é facilitar a vida de pesquisadores que precisam extrair e ligar informações oriundas de diversos repositórios de código.

Fique de olho: a mineração de repositórios de código veio para ficar e vai ajudar gerentes e desenvolvedores a tomarem decisões no dia a dia e pesquisadores a entenderem melhor a Engenharia de Software, acabando com muitos de seus mitos.

MARCO AURÉLIO GEROSA | Professor associado no Departamento de Ciência da Computação da Universidade de São Paulo. Sua pesquisa está na interseção entre Engenharia de Software e Sistemas Colaborativos, focando em engenharia de software experimental, mineração de repositórios de software, evolução de software e dimensões sociais do desenvolvimento de software. Recebe bolsa de produtividade do CNPq e coordena projetos de software livre que já receberam diversos prêmios. Para mais informações, visite www.ime.usp.br/~gerosa.

IGOR SCALIANTE WIESE | Doutorando pela Universidade de São Paulo (IME-USP) sob supervisão do professor Marco Gerosa. É professor do Departamento de Computação da Universidade Tecnológica Federal do Paraná - campus Campo Mourão. Cursou Doutorado-sanduíche na University of California - Irvine sob a orientação do professor David Redmiles. Atua na área de Engenharia de Software, em especial nos temas de mineração de repositórios de software, métricas de software e dependências de software.





GUSTAVO ANSALDI OLIVA | Doutorando pela Universidade de São Paulo (IME-USP) sob supervisão do professor Marco Gerosa. Sua principal pesquisa é em Engenharia de Software, com foco na recuperação, análise e visualização de dependências de software. Gustavo já recebeu bolsas da HP Brasil e da Comissão Europeia. Na indústria, Gustavo trabalhou na IBM Brasil como consultor e desenvolvedor de software por mais de 3 anos. Recentemente, cursou Doutorado-sanduíche na Queen's University sob a supervisão do professor Ahmed Hassan.

MAURÍCIO FINAVARO ANICHE | Doutorando pela Universidade de São Paulo (IME-USP) sob supervisão do professor Marco Gerosa. É também instrutor e desenvolvedor pela Caelum, renomada instituição de ensino brasileira. Atua na área de engenharia de software e suas principais linhas de pesquisa são Test-Driven Development e Métricas de Código. É criador do MetricMiner, ferramenta de código aberto, que ajuda pesquisadores a minerar repositórios.