Newcomers Withdrawal in Open Source Software Projects: Analysis of Hadoop Common Project

Análise da Desistência de Novatos em Projetos de Software Livre: Caso do Projeto Hadoop Common

Igor Steinmacher, Igor S. Wiese, Ana Paula Chaves
Coordenação de Ciências da Computação
Universidade Tecnológica Federal do Paraná (UTFPR)
e-mail: {igorfs, igor, anachaves}@utfpr.edu.br

Marco Aurélio Gerosa
Departamento de Ciência da Computação - IME
Universidade de São Paulo (USP)
e-mail: gerosa@ime.usp.br

***Abstract*— Collective production communities, like open source projects, are based on volunteers collaboration and require newcomers for their continuity. Newcomers face difficulties and obstacles when starting their contributions, resulting in a large withdrawal and consequent low retention rate. This paper presents an analysis of newcomers withdrawal, checking if the dropout is influenced by lack of answer, answers politeness and helpfulness, and the answer author. We have collected five years data from the developers mail list communication and task manager (Jira) discussions of Hadoop Common project. We observed the users' communication, identifying newcomers and classifying questions and answers content. For the study conducted, less than 20% of newcomers became long term contributors. There are evidences that the withdrawal is influenced by the respondents and by the type of response received. However, the lack of answer was not evidenced as a factor that influences newcomers withdrawal in the project.**

Keywords-newcomer; communication; collaboration; open source software; withdrawal

***Resumo*— Comunidades de produção coletiva, como as que mantêm projetos de software livre, demandam a colaboração de voluntários e necessitam de novatos para sua continuidade. Os novatos enfrentam dificuldades e obstáculos ao iniciar sua interação em um projeto, resultando em uma grande desistência e consequente baixa taxa de retenção. Este trabalho apresenta uma análise da desistência de novatos, observando se a ausência de resposta, respostas que não os ajudam ou o autor das respostas, impactam nessa desistência. Foram coletados cinco anos de comunicação da lista de e-mails de desenvolvedores e de discussões do gerenciador de tarefas (Jira) do projeto Hadoop Common. Observamos a comunicação dos usuários, identificamos os novatos e classificamos o conteúdo das perguntas e respostas. Verificamos que menos de 20% dos novatos que ingressaram continuaram contribuindo no projeto. Há indícios de que a desistência é influenciada pelos autores das respostas e o pelo tipo de resposta recebida. Entretanto, a ausência de resposta não mostrou ser fator relevante na desistência dos novatos no projeto.**

Keywords-novatos; comunicação; colaboração; software livre; evasão

I. INTRODUÇÃO

Projetos de Software Livre são conduzidos por voluntários, o que demanda a constante entrada e retenção de novos contribuintes [1]. Entretanto, os primeiros passos em projetos de software livre podem oferecer diversos obstáculos. Degenais et al. [2] compara novatos em projetos de software a exploradores que precisam se orientar em um ambiente hostil. Por um lado, os novatos precisam aprender aspectos sociais e técnicos sozinhos, explorando as informações existentes em listas de e-mails, repositórios de código fonte e gerenciadores de tarefas [3]. Por outro, não é fácil acessar essas informações devido ao grande volume, à falta de ferramentas para navegar nos repositórios, e à dificuldade de fazer as conexões entre os itens relacionados logicamente em fontes diferentes [5].

Em um estudo anterior [4], apresentamos relatos de desenvolvedores que tentaram iniciar a participação em dois projetos de software livre conhecidos. Os desenvolvedores indicaram que a ausência de informação e de orientação durante a realização dos primeiros passos causa vários desconfortos, desencorajando mais contribuições. Para reduzir esse tipo de problema, geralmente, os novatos postam suas dúvidas e solicitam ajuda na escolha de tarefas em fóruns e listas de e-mail existentes ou enviam e-mails para os desenvolvedores que têm papel mais central no projeto (e.g. *owners*, líderes de projeto) [1, 6]. Por serem meios de livre acesso, as pessoas muitas vezes utilizam esses meios como forma de iniciar suas contribuições no projeto. Entretanto, o recebimento de respostas que não oferecem o encaminhamento correto ou respostas mal educadas pode resultar na desistência dos novatos.

Diante do cenário apresentado, é importante realizar observações em diferentes comunidades de software livre para entender a forma de interação e as demandas de novatos quando iniciam a participação nesse tipo de projeto. Esse entendimento possibilitará a criação de mecanismos e ferramentas para auxiliar a retenção de novatos em projetos de software livre, possibilitando a recomendação de informações e de possíveis mentores que guiem os primeiros passos dos novatos. Esse ferramental técnico e conceitual também poderá ser estendido para outras comunidades que

978-0-7695-4890-6/12 $26.00 © 2012 IEEE
DOI 10.1109/SBSC.2012.16

dependam de produção coletiva por meio de trabalho voluntário, como enciclopédias virtuais e outros sistemas de mídia social.

Este artigo apresenta um estudo que tem por objetivo verificar se a ausência de resposta, polidez e utilidade das respostas, ou autores das respostas recebidas por novatos na lista de e-mails e no gerenciador de tarefas influenciam na decisão de continuar no projeto. Buscamos entender as razões para a desistência dos novatos baseado nas primeiras interações com o projeto. Para isso, examinamos a questão de pesquisa:

> *Ausência de resposta, polidez, utilidade ou o tipo de autor das respostas influencia na permanência de novatos em um projeto de software livre?*

Para responder à questão de pesquisa, foram definidos três objetivos específicos, a saber:

- Verificar se os novatos recebem respostas;
- Observar quem são os autores das respostas às dúvidas dos novatos; e
- Classificar as respostas recebidas pelos novatos.

Para o presente estudo foi escolhido para observação o projeto Hadoop Common , hospedado no repositório da Apache Software Foundation. Para a análise, foram utilizados dados de lista de discussão de desenvolvedores e do gerenciador de tarefas (Jira) e da lista de discussão de usuários.

O restante do trabalho está organizado da seguinte forma: Na Seção II são apresentados alguns trabalhos relacionados. Na Seção III é apresentado o método da pesquisa, detalhando cada um dos seus passos. Na Seção IV, os resultados obtidos são apresentados e discutidos. Na Seção V são apresentadas as ameaças à validade. As conclusões e trabalhos futuros são descritos na Seção VI.

II. TRABALHOS RELACIONADOS

Vários trabalhos na literatura analisam a entrada de novatos em comunidades de produção coletiva, que incluem estudos sobre novatos no Wikipedia [12, 13, 14] e em projetos de software [1, 2, 4, 6, 7]. Especificamente para projetos de software livre, estudar novatos é importante, pois, de acordo com Jensen et al. [7], eles são potenciais futuros contribuintes que são vitais para o crescimento e sobrevivência de projetos.

O presente artigo segue a linha de outros trabalhos publicados anteriormente, que estudam os passos iniciais e as dificuldades enfrentadas por novatos em projetos de software livre [6, 7]. Park e Jensen [1] estudam as necessidades dos novatos por informação. Os autores mostram que ferramentas de visualização de informações apoiam os primeiros passos de novatos na aprendizagem sobre um projeto de Software Livre, ajudando-os a encontrar informações mais rapidamente.

Von Krogh et al. [6] conduziram um estudo sobre o projeto FreeNet, por meio de entrevista com desenvolvedores, análise de histórico de e-mails, repositório de código fonte e documentos do projeto. Os autores propuseram um roteiro de entrada (*joining script*) para os desenvolvedores que desejem entrar naquela comunidade virtual. Uma de suas contribuições indica que os novatos frequentemente ficam observando o projeto antes de iniciar sua participação, para, em seguida, interagir. Apesar de estudar o processo de entrada em projetos de software livre, não há uma preocupação com a análise das razões de desistência, verificando apenas o comportamento daqueles que se tornam desenvolvedores do projeto.

Nakakoji et al. [10] estuda quatro projetos de software livre para analisar a evolução da comunidade desses projetos. O estudo apresenta oito papéis encontrados para os membros de um projeto e os apresenta na forma de uma cebola, o chamado *onion patch*. Outros autores conduziram seus estudos baseados no modelo *onion patch* [7, 8, 9]. A teoria por trás desse modelo diz que novatos geralmente iniciam pelas camadas mais externas do modelo e vão em direção ao centro de acordo com seus objetivos. Esses artigos tratam da entrada e evolução da participação de membros em comunidades de software livre, mas nenhum dos estudos se preocupa com as razões da desistência dos novatos.

O trabalho apresentado por Jensen et al. [7] é o que mais se aproxima do presente artigo em termos de método da pesquisa e objetivos. Os autores fazem uma análise de quatro listas de e-mails de projetos de software livre, visando verificar se os e-mails de novatos são respondidos rapidamente, se o sexo (gênero) e a nacionalidade dos novatos interferem no tipo de resposta recebida e na continuidade dos novatos, e, por fim, se o tratamento a novatos é diferente na lista de usuários e na lista de desenvolvedores. O presente artigo tem por objetivo se aprofundar nas razões pelas quais os novatos desistem.

III. MÉTODO DE PESQUISA

Para conduzir esta pesquisa, foram utilizados dados do projeto Hadoop Common[1]. O projeto foi escolhido por ser um projeto de sucesso, já consolidado, e com uma comunidade ativa e bem organizada. Além disso, os dados do gerenciador de tarefa[2] (Jira) e listas de e-mails estão disponíveis e podem ser coletados livremente.

Os dados utilizados foram obtidos da lista de discussões de desenvolvedores (common-dev) e dos comentários provenientes do gerenciador de tarefas do projeto. Foram coletados os dados de e-mails, tarefas e comentários, com data entre janeiro de 2006 e dezembro de 2010. As análises da lista de e-mails e da ferramenta Jira foram realizadas separadamente. As seções a seguir descrevem detalhes da coleta dos dados.

A. Coleta de Dados do Jira

Para coletar os dados do gerenciador de tarefas (Jira), construímos uma ferramenta para extrair os dados relativos às tarefas e armazená-los em um banco de dados relacional local. A extração é feita acessando a página web que apresenta a tarefa e variando a URL, que tem o formato:

[1] http://hadoop.apache.org/common
[2] https://issues.apache.org/jira/browse/HADOOP

"*https://issues.apache.org/jira/browse/<nome_projeto>-<número_tarefa>*. O sistema recebe como parâmetro o nome do projeto e varia o valor do número da tarefa sequencialmente.

O extrator interpreta cada página HTML e extrai as seguintes informações para cada tarefa relatada: descrição; usuário relator (*reporter*); encarregado (*assignee*); data de criação; data de fechamento; prioridade; status atual; e comentários (com autor, data e mensagem). Para a análise, foram considerados os usuários que aparecem como relatores, encarregados ou que tenham comentado qualquer tarefa.

B. *Extração dos dados de e-mails*

Para extrair os dados de e-mails, primeiramente foram obtidos os arquivos mbox (formato de armazenamento de coleções e-mails) de cada mês do período investigado. Esses arquivos contêm todos os e-mails, incluindo cabeçalho e mensagem, enviados para a lista no mês correspondente.

As informações das mensagens contidas nos arquivos mbox foram mineradas analisando os cabeçalhos para adquirir informações do conteúdo da mensagem, assunto, ID da mensagem, remetente e identificador da cadeia de mensagens (*In-reply-to*), que identifica a árvore de discussão (*thread*) a qual a mensagem pertence. Essas árvores foram reconstruídas verificando o campo *in-reply-to* do cabeçalho bem como o assunto do e-mail (examinando os prefixos "*Re*:", "*Fwd*:") e o campo *references* do cabeçalho, para diminuir as chances de perda de mensagens relativas a uma discussão.

Por fim, os e-mails foram armazenados em um banco de dados local contendo todos os detalhes das mensagens extraídas. Para a análise, foram desconsideradas as mensagens enviadas automaticamente na criação, comentário ou mudança de estado de uma tarefa no Jira. Tais mensagens foram identificadas verificando se o remetente tinha endereço jira@apache.org ou se o assunto iniciava com o identificador "[jira]". A separação das mensagens foi importante, uma vez que as análises foram feitas considerando os ambientes isoladamente.

C. *Análise dos Dados*

Foram obtidos 60 meses de discussões da lista de e-mails dos desenvolvedores, que contêm 7891 árvores de discussão com 37095 respostas, resultando em um total de 45076 mensagens. Também foram coletados 60 meses de tarefas criadas no gerenciador de tarefas Jira, totalizando 6793 tarefas com 53664 comentários.

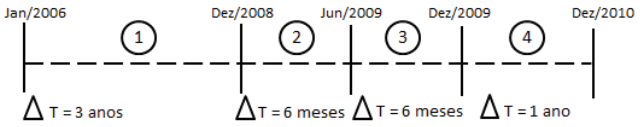

Figura 1. Linha do tempo para a coleta de dados

Como pode ser observado na Figura 1, o período de análise foi dividido em 4 intervalos. Inicialmente, foi realizada uma consulta no intervalo 1 para identificar as pessoas que contribuíram com o projeto durante os 3 primeiros anos de registro. Para a análise, esses usuários são considerados membros já pertencentes ao projeto.

No intervalo 2, foram identificadas as pessoas que iniciaram suas contribuições nesse período, e não haviam aparecido no intervalo 1. Essas pessoas foram consideradas novatas no contexto deste estudo.

No intervalo 3, foram verificados quais novatos voltaram a aparecer nos 6 meses seguintes e quais não retornaram. Os usuários novatos que não apareceram no intervalo 3 foram considerados desistentes.

Por fim, no intervalo 4, foram constatados aqueles usuários que continuaram a contribuir no próximo intervalo de 1 ano. Os usuários que apareceram nos intervalos 2, 3 e 4 foram considerados novatos que continuaram no projeto.

Para os novatos desistentes, foi conduzida uma análise manual das mensagens enviadas e das respostas recebidas, a fim de classificá-las de acordo com a ausência de resposta, os autores das respostas e com o tipo da resposta enviada aos novatos. A definição dos possíveis tipos de mensagens foi realizada adaptando o método definido por Qu et al. [11] para o domínio do estudo. O objetivo foi criar uma classificação e revisar o modelo independentemente. O método utilizado para criar a classificação é ilustrado na Figura 2.

Primeiramente, foi sugerida uma classificação inicial para as mensagens, que foi discutida pelos autores e resultou em um modelo inicial. Em seguida, foi examinada uma amostra aleatória de mensagens enviadas. Esse conjunto de mensagens foi categorizado por dois pesquisadores, de acordo com o modelo definido. Os dois pesquisadores discutiram as inconsistências e entraram em consenso com relação à classificação e aos critérios. As subseções a seguir descrevem alguns detalhes para cada um dos objetivos específicos definidos para responder à questão de pesquisa.

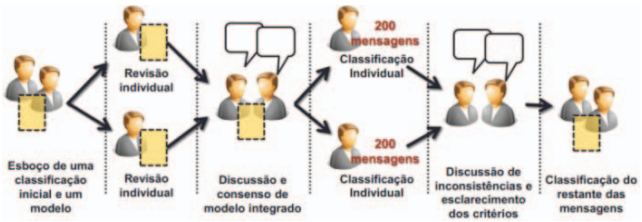

Figura 2. Método utilizado para definir o modelo de classificação de mensagens

1) *Verificar se os novatos recebem respostas*

Uma vez que os novatos foram identificados, verificou-se o recebimento de respostas e o tempo decorrido até o recebimento da primeira resposta. Para o gerenciador de tarefas, foi necessário observar se o novato era relator ou apenas participava da discussão enviando comentários. Se o novato era o relator, os comentários realizados por outros usuários na tarefa foram considerados respostas.

Já para a lista de e-mail, verificou-se se as mensagens dos novatos foram as primeiras mensagens da árvore de discussão. Em caso positivo as mensagens foram classificadas como pergunta. Caso elas fizessem parte de uma discussão já existente, a mensagem era classificada

como parte das respostas. Para as mensagens classificadas como perguntas, foi verificada a existência de respostas na mesma discussão.

2) Observar quem são os autores das respostas

Para as mensagens que receberam respostas, analisaram-se os e-mails e nome de usuário dos membros que respondiam às mensagens. Em seguida, os e-mails e nomes de usuário foram classificados de acordo com o período de aparição do membro e com a quantidade de mensagens enviadas anteriormente, dividindo-os em três categorias:

- Membros centrais: apareceram no intervalo 1 e estão entre os 10% mais participativos;
- Novato: não apareceu no intervalo 1 e apareceu no intervalo 2;
- Outros membros: apareceram no intervalo 1 e não estão entre os 10% mais participativos.

Isso foi feito para verificar se a 'experiência' anterior do membro que responde à pergunta tem influência na qualidade da resposta e na permanência do novato.

3) Classificar as respostas recebidas pelos novatos

As respostas recebidas pelos novatos foram também classificadas de acordo com os tipos de respostas definidos pelo método apresentado na Figura 2. Assim, cada mensagem foi classificada em um dos seguintes tipos:

- No Tópico/Ajuda: quando a resposta está no tópico do problema levantado e auxilia na resolução do problema;
- Indiferente: quando a mensagem não é esclarecedora, não apresentam tom receptivo e normalmente indicam um link externo para o usuário esclarecer sua dúvida;
- Não Útil/Fora do Tópico: quando a resposta é fora do tópico e não contribui na solução do problema;
- Não Útil/Outra questão: quando a resposta é um novo problema, criando uma discussão diferente da inicial;
- Outros: quando não é possível classificar nos tipos anteriores, por exemplo, anúncio de produtos ou mensagens não entendíveis.

D. Questionário Enviado aos Desistentes

Após a análise estatística foi realizada uma pesquisa via e-mail com os novatos que haviam desistido, para complementar o entendimento das razões de interação e desistência do projeto. Para tanto foi enviado um questionário aos novatos que deixaram o projeto nos intervalos 2 e 3. As questões enviadas e os resultados podem ser verificados na Seção IV.D.

IV. RESULTADOS OBTIDOS

Nesta seção, são apresentados e discutidos os resultados obtidos. Primeiramente, é apresentada uma análise geral sobre as informações obtidas, e, em seguida, são apresentadas discussões específicas para cada uma das questões de pesquisa.

A Tabela I apresenta o resultado da análise dos dados da lista de e-mails dos desenvolvedores. Para cada intervalo, é apresentada a quantidade de usuários encontrados, a porcentagem dos usuários existentes em cada intervalo pelo total encontrado, e a porcentagem dos novatos que ficaram no projeto no período seguinte e, posteriormente, aqueles que continuaram contribuindo. No segundo intervalo, 67 novatos apareceram e 20 deles (29,85%) continuaram no próximo intervalo. Após um ano, somente 12 (17,91%) continuaram contribuindo.

TABELA I. USUÁRIOS QUE CONTRIBUÍRAM NA LISTA DE E-MAILS

	# usuários	*% dos existentes*	*% dos novatos*
Existentes (intervalo 1)	677		
Novatos (intervalo 2)	67	9,90%	
Ficaram (intervalo 3)	20	2,95%	29,85%
Continuaram (intervalo 4)	12	1,77%	17,91%

Uma análise semelhante foi feita sobre os dados do gerenciador de tarefas. Os resultados são apresentados na Tabela II. Foram considerados usuários que relataram ou que comentaram/discutiram em alguma tarefa. É possível verificar que surgiu uma quantidade maior de novatos, quando comparado à lista de e-mails.

TABELA II. USUÁRIOS QUE CONTRIBUÍRAM SENDO RELATOR OU ESCREVENDO COMENTÁRIOS NO GERENCIADOR DE TAREFAS

	# usuários	*% dos existentes*	*% dos novatos*
Existentes (intervalo 1)	483		
Novatos (intervalo 2)	127	26,29%	
Ficaram (intervalo 3)	30	6,21%	23,62%
Continuaram (intervalo 4)	17	3,52%	13,39%

A Tabela III apresenta uma visão geral da evolução da participação dos novatos na lista de e-mails do projeto. A tabela apresenta a quantidade de novatos que seguiram contribuindo com o projeto e a quantidade de mensagens enviadas por esses novatos em cada intervalo. A primeira observação que se pode fazer é que, apesar de haver um grande índice de desistência de novatos no intervalo 2, a quantidade de mensagens enviadas pelos membros que permaneceram no projeto aumentou.

Observa-se também que o intervalo 2 apresenta uma maior quantidade de perguntas que os intervalos 3 e 4, tanto em termos absolutos quanto se for observada a relação com a quantidade de novatos. A razão disso é que as primeiras interações dos usuários com a lista são feitas para esclarecer dúvidas, configurar ambiente ou solicitar ajuda para dar os passos iniciais do projeto.

TABELA III. EVOLUÇÃO DA PARTICIPAÇÃO DOS NOVATOS NA LISTA DE E-MAILS DE DESENVOLVEDORES NOS INTERVALOS 2, 3 E 4.

	Intervalo 2			*Intervalo 3*			*Intervalo 4*		
	Msgs	*Nova-tos*	*Msgs/ novato*	*Msgs*	*Nova-tos*	*Msgs/ novato*	*Msgs*	*Nova-tos*	*Msgs/ novato*
Perguntas feitas	68	47	1,45	18	9	2,00	6	5	1,20
Respostas a outras discussões	56	24	2,33	160	15	10,67	55	12	4,58
Respostas à própria discussão	56	20	2,80	17	6	2,83	12	4	3,00
TOTAL	180	67	2,69	195	20	9,75	73	12	6,08

Quanto às respostas às discussões iniciadas por outros membros, no intervalo 2 foram enviadas apenas 56 mensagens em 32 discussões distintas, com a colaboração de 24 novatos. Poucos novatos (9) participaram de mais de uma discussão e somente 8 novatos escreveram mais de uma resposta em discussões de terceiros. No terceiro intervalo, 15 novatos responderam a 160 mensagens que foram enviadas por terceiros em 94 discussões. Isso mostra que, após um período inicial na lista, os novatos que seguiram no projeto passaram a contribuir mais em discussões iniciadas por terceiros, e a auxiliar na solução de problemas. Já no intervalo 4, houve um decréscimo nas respostas enviadas pelos novatos que continuaram.

A diminuição de mensagens enviadas pelos novatos que continuaram (intervalo 4) pode ser observada em todas as linhas da tabela. Não foi encontrada uma razão para tal redução durante a análise manual. Entretanto, pode haver uma relação com o processo de crescimento dos membros dentro do projeto, que passam a contribuir de outras formas, como, por exemplo, respondendo a questões ou corrigindo defeitos. Um fato a observar é que, embora os novatos que continuaram tenham enviado apenas 55 mensagens respondendo a terceiros, todos os 12 novatos apareceram nessas respostas.

Também foi analisada a evolução dos novatos encontrados no gerenciador de tarefas Jira. Os resultados são apresentados na Tabela IV. Diferentemente do que ocorreu na lista de e-mails, observa-se que a média de mensagens por novatos (msgs/novato) aumentou a cada intervalo de tempo analisado. Esse aumento é verificado na quantidade de tarefas relatadas e de comentários enviados nas discussões. Esse crescimento deve-se à visibilidade e à confiança que os usuários vão adquirindo a cada vez que contribuem de forma positiva.

TABELA IV. EVOLUÇÃO DA PARTICIPAÇÃO DOS NOVATOS NO GERENCIADOR DE TAREFA NOS INTERVALOS 2, 3 E 4

	Intervalo 2			*Intervalo 3*			*Intervalo 4*		
	Msgs	*Nova-tos*	*Msgs/ novato*	*Msgs*	*Nova-tos*	*Msgs/ novato*	*Msgs*	*Nova-tos*	*Msgs/ novato*
Tarefas relatadas	154	78	1,97	61	15	4,07	76	10	7,60
Comentários a outras tarefas	420	107	3,93	308	17	18,12	356	14	25,43
Comentários à própria tarefa	421	55	7,65	260	18	14,44	331	11	30,09
TOTAL	995	127	7,83	629	30	20,97	763	17	44,88

A análise inicial evidenciou a pequena proporção de novatos que continuaram tanto no ambiente de lista de e-mails quanto no gerenciador de tarefas. Para investigar mais sobre as possíveis razões da desistência dos novatos, as próximas seções apresentam discussões relacionadas aos objetivos específicos definidos para este estudo.

A. Os Novatos Recebem Respostas?

A Tabela V apresenta os dados relativos ao recebimento de respostas por novatos na lista de e-mails de desenvolvedores e à desistência ou continuidade desses novatos. A observação mostra que 47 novatos enviaram mensagens no intervalo 2. Os novatos enviaram 68 perguntas à lista de e-mails. Um total de 38 novatos obtiveram respostas a 47 e-mails enviados, entretanto 7 tratavam-se de respostas enviadas pelo próprio novato. Obteve-se, então, 34 novatos (72,34%) que foram respondidos em um total de 40 árvores de discussão. Dentre as pessoas que receberam respostas, o tempo médio para recebimento foi de 1,32 dias, e 19 delas foram respondias no mesmo dia.

TABELA V. RESPOSTAS *VS.* DESISTÊNCIA DOS NOVATOS NOS INTERVALOS 2, 3 E 4 CONSIDERANDO PARTICIPAÇÃO NA LISTA DE E-MAILS

	# pessoas	*Desistiram*	*Aparecem nos intervalos 3 e 4*
Sem resposta	13	11	2
Com resposta	34	30	4

Dentre os 13 novatos que não receberam resposta alguma para suas questões, onze desistiram (84,62%) e os outros dois (15,38%) seguiram contribuindo com o projeto. Dentre aqueles que tiveram alguma pergunta respondida, 30 desistiram (88,24%) e quatro continuaram (11,76%) no projeto. Portanto, não existe um indício numérico de que o não recebimento de respostas influencia a desistência dos novatos no projeto.

A análise manual das mensagens possibilitou verificar que algumas mensagens não são respondidas, pois as perguntas foram feitas fora do contexto da lista de e-mails. Por exemplo, foram encontradas dúvidas de instalação e configuração do Hadoop enviadas à lista de desenvolvedores, quando deveriam ser enviadas à lista de usuários.

Também foi possível observar que a maior parte das perguntas são prontamente respondidas, e os autores das dúvidas agradecem pela resposta, e, ainda assim, o novato deixou de participar da comunidade logo após a resposta, mesmo que fosse útil. Nesse caso percebe-se que a pessoa que enviou e-mail não tinha intenção de prosseguir no projeto, mas resolver algum problema que estava enfrentando momentaneamente.

A Tabela VI apresenta os dados relativos ao recebimento de comentários em tarefas relatadas por novatos no Jira e à desistência ou continuidade desses novatos. Foram 78 novatos relatando tarefas durante o intervalo 2, dentre os quais, 71 (91%) receberam comentários. Apenas sete novatos não receberam comentários em oito tarefas enviadas. Ao analisar manualmente essas tarefas, percebeu-se que seis delas foram redirecionadas para o projeto MapReduce, cujo grau de atividade é menor do que o do projeto Hadoop Common. Com isso, considera-se que apenas duas tarefas não receberam comentários.

TABELA VI. PERGUNTAS, RESPOSTAS E DESISTÊNCIA DOS NOVATOS NOS INTERVALOS 2, 3 E 4 CONSIDERANDO A PARTICIPAÇÃO NO JIRA

	# pessoas	*Desistiram*	*Aparecem nos intervalos 3 e 4*
Sem comentários	7	6	1
Com comentários	71	55	16

Pode-se perceber que a receptividade no Jira é muito boa. Até mesmo as tarefas que reportavam algo que não faz parte

do escopo daquela ferramenta ou que relatavam problemas já reportados anteriormente, eram comentadas dando o direcionamento correto aos usuários.

Assim, pode-se afirmar que o Jira é um ambiente em que os novos membros são bem recebidos, e receber comentários nessa ferramenta não é um fator que influencia a desistência ou permanência no projeto Hadoop Common.

B. Quem são os Autores das Respostas aos Novatos

A Figura 3 apresenta um diagrama de Venn mostrando a relação entre perguntas feitas por novatos e membros que as responderam no contexto da lista de e-mails de desenvolvedores durante o intervalo 2. Na figura cada um dos conjuntos representa o tipo dos autores que responderam questões disparadas por novatos. Os autores de respostas foram classificados de acordo com os tipos apresentados na Seção III.C.3. Os valores apresentados no interior dos conjuntos representam a quantidade de discussões em que um determinado tipo de membro participou.

Pode-se perceber que quem mais responde às perguntas enviadas pelos novatos são os membros centrais do projeto. Dentre as discussões iniciadas por desistentes, 21 (63,63%) tiveram participação dos membros centrais, das quais sete foram respondidas apenas por membros centrais. Onze discussões (34,38%) iniciadas por desistentes não tiveram participação de membros centrais nas respostas. Durante a leitura das mensagens enviadas por e-mail, foi possível verificar que, em alguns casos, novatos respondendo a novatos trazem sim influência negativa, que poderia resultar na desistência. Por exemplo, em uma discussão na qual um novato solicitava auxílio na escolha de um defeito para iniciar sua contribuição, outro novato respondeu. A resposta dele dizia que apenas usuários que eram *committers* poderiam trabalhar em defeitos. Em outro caso, o novato solicitava informações sobre a arquitetura e qual seria o meio mais simples para iniciar no projeto. Dois outros novatos enviaram mensagens na discussão também dizendo que desejavam contribuir.

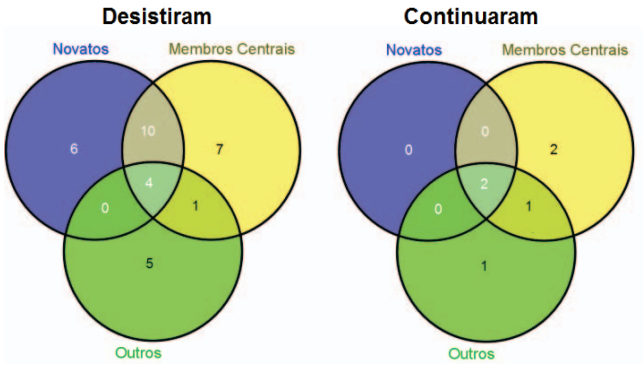

Figura 3. Análise de quem respondeu às perguntas dos novatos que desistiram e continuaram na lista de e-mails

Observando a Figura 3, também é possível verificar que a maioria das discussões iniciadas por novatos que desistiram receberam respostas de membros centrais. Ao analisar o texto dos e-mails, percebe-se que grande parte dessas discussões diz respeito a dúvidas técnicas pontuais de um usuário ou de configuração de um ambiente específico, sendo as mensagens inclusive replicadas para a lista de usuários. Esses novatos, algumas vezes, apenas participam da lista para resolver seus próprios problemas, sem terem a real intenção de colaborar com o projeto.

Observando as respostas às mensagens dos novatos que continuaram, percebe-se que não houve discussões com participação exclusiva de novatos. Apenas 2 árvores de discussão apresentaram participação de novatos. Essas discussões foram analisadas, sendo discussões com seis e nove mensagens, trocadas por cinco pessoas diferentes em cada discussão. Os dois novatos que aparecem continuaram no projeto e suas mensagens colaboram com a discussão.

Os resultados obtidos com a análise do gerenciador de tarefas Jira são apresentados na Figura 4. Pode-se perceber que existe uma maior quantidade de respostas aos novatos, tanto para os que continuaram, quanto para os que desistiram. No Jira, 39 relatos (45,34%) de novatos desistentes não tiveram comentários de membros centrais, enquanto para os que continuaram o número cai para 11 (28,20%).

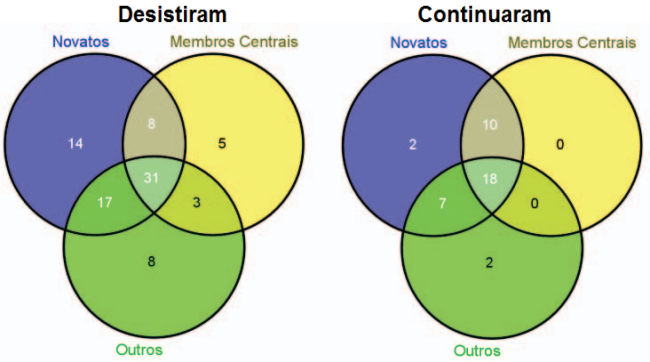

Figura 4. Análise de quem comentou tarefas de novatos que desistiram e continuaram no gerenciador de tarefas

De modo geral, percebe-se maior contribuição dos membros centrais, e mais respostas de novatos no Jira do que na lista e-mails. Isso ocorre, pois a ferramenta estimula as discussões, que são mais contextualizadas e focadas, e terão resultado efetivo no projeto. Notou-se ainda que, durante as discussões, os novatos recebem auxílio dos membros centrais durante o envio de contribuições (relatos de defeitos, melhorias ou envio de componentes de atualização).

C. Que Tipos de Respostas os Novatos Recebem?

Durante a leitura das discussões iniciadas por novatos, as respostas a cada uma delas foram classificadas a fim de verificar o impacto da resposta na desistência. A Tabela VII apresenta o resultado da classificação de acordo com o modelo definido seguindo o método apresentado na Seção III.C.2. Pode-se verificar que as respostas "Não Úteis" ou "Indiferentes" foram dadas apenas a desistentes. Foram nove respostas classificadas em um desses tipos. Esse pode ser um indício de que o tipo da resposta pode ter influência sobre a desistência dos novatos. Algumas outras evidências disso serão apresentadas na Seção IV.E.

TABELA VII. TIPO DE RESPOSTA QUE OS NOVATOS RECEBERAM

Tipo da Resposta	*Desistiram*	*Continuaram*
Ajuda / No Tópico	20	7
Não Útil / Outra questão	5	0
Não Útil / Fora do tópico	3	0
Indiferente	1	0
Outro	4	0

Os dados da Tabela VII mostram ainda que, mesmo recebendo respostas que auxiliaram na sua dúvida, alguns novatos desistiram do projeto. A inspeção manual das mensagens possibilitou identificar que as questões enviadas pelos novatos continham dúvidas não relacionadas à maneira de contribuir com o projeto ou a dúvidas técnicas relativas a uma contribuição. As mensagens eram relacionadas a necessidades específicas do usuário, como, por exemplo, a integração de uma tecnologia proprietária ao Hadoop, verificação da existência de versão de biblioteca em uma *release*, divulgação de resultados de testes utilizando grades de computadores, dúvida de uso e instalação do Hadoop e incompatibilidade com uma máquina virtual Java. Dessa forma, percebe-se que alguns usuários tinham intenção de esclarecer dúvidas pontuais, e não necessariamente tinham interesse em permanecer contribuindo.

Para o gerenciador de tarefas Jira, as observações feitas a partir das análises manuais conduzidas mostraram que as respostas eram no tópico, contextualizadas e provia informações úteis. Por se tratar de um ambiente controlado, não houve alterações ou padrões a tratar no contexto dessa questão. Algumas exceções apareciam quando, por exemplo, uma tarefa estava relatando um problema de configuração ou instalação. Nesses casos, os usuários respondiam redirecionando o novato ao fórum correto.

D. Questionário Conduzido com os Novatos Desistentes

Nesta seção apresentamos detalhes do questionário enviado via e-mail aos novatos que desistiram do projeto. Ao fim da análise dos dados foi enviado um pequeno questionário aos 55 novatos que deixaram o projeto, com as seguintes questões:

> *1. Do you remember sending an email to hadoop-common-dev mail list? [Y/N]*
> *2. At that time, were you interested in keep contributing to Hadoop project? [Y/N]*
> *2a. In case you answered YES to question 2, why did you give up?*
> *2b. In case you answered NO to question 2, what was the goal of the messages sent to developers list?*
> *3. Have you contributed to the project after June 2009? [Y/N]*
> *4. Have you contributed to other Open Source project BEFORE 2009? [Y/N]*

Dos e-mails enviados, 10 entregas falharam e 13 foram respondidos. As respostas recebidas estão sumarizadas na Tabela VIII. Outras 31 pessoas não responderam ao e-mail. Um usuário respondeu ao e-mail sem respostas ao questionário. Ele apenas informou que continuava no projeto e havia se tornado *committer* recentemente. Esse é o caso do usuário que apresentamos no início desta seção.

TABELA VIII. RESPOSTA DO QUESTIONÁRIO DOS DESENVOLVEDORES QUE DEIXARAM DE CONTRIBUIR COM O PROJETO HADOOP COMMON

	Questão 1	*Questão 2*	*Questão 3*	*Questão 4*
Sim	13	11	1	7
Não	0	2	12	6

Percebe-se que os 13 respondentes se lembram da interação com a lista e explicitaram a intenção de continuar no projeto (11 responderam positivamente à questão 2) e mais da metade dos desistentes tinham experiência anterior em outros projetos. Tal questão se desdobrava em duas outras, de acordo com a opção escolhida, para entendermos as razões da desistência. As respostas foram analisadas e o resultado é apresentado na Tabela IX.

As respostas classificadas com os tipos *ii*, *iii* e *iv* mostram que uma das possíveis causas para desistência dos usuários foi a receptividade do projeto. Das 13 pessoas que responderam que tinham intenção de participar do projeto, seis enviaram respostas relacionadas à recepção. Destacamos duas respostas à questão 2a que mostram claramente a insatisfação dos desistentes: um deles respondeu *"... meu problema era como começar a contribuir... se eu tivesse alguém para segurar minha mão, isso poderia ter ajudado..."*; o outro foi direto ao assunto, *"eles não responderam à minha pergunta"*.

TABELA IX. CLASSIFICAÇÃO DAS RESPOSTAS ÀS QUESTÕES ENVIADAS AOS DESENVOLVEDORES QUE DEIXARAM O PROJETO HADOOP COMMON

Tipo da Resposta	*Resposta à questão 2*	
	Sim	*Não*
i. Era usuário e só quis esclarecer dúvidas	0	2
ii. Pergunta não respondida ou resposta não agradou	2	0
iii. Falta de ajuda em escolher tarefa	3	0
iv. Diz não ter sido aceito pelo projeto	1	0
v. Mudou de foco ou empresa	4	0
vi. Voltou depois ao projeto	1	0

Ainda dentre as pessoas que tinham interesse em contribuir percebe-se que um usuário voltou a contribuir com o projeto. Ele informou que passou algum tempo sem trabalhar com Hadoop e, desde 2011, voltou a contribuir com o projeto respondendo a dúvidas e discutindo na lista.

E. Análise da contribuição dos novatos que continuaram no projeto

Para cada novato que continuou no projeto (apareceu nos intervalos 2, 3 e 4), foi investigada a participação em outros meios (lista de usuários, Jira, repositório de códigos fontes). Isso possibilita uma validação cruzada entre as diferentes fontes de dados e a observação da evolução dos novatos que permaneceram no projeto por mais de um ano após sua primeira aparição.

A Tabela X apresenta o padrão de contribuição dos 24 desenvolvedores que continuaram no projeto. Dentre eles, sete novatos participaram da lista de desenvolvedores, ao menos por um período. Esses desenvolvedores também participaram da lista de usuários. Somente um desenvolvedor

não contribuiu no gerenciador de tarefas, o que mostra que os novatos que começaram contribuindo apenas na lista de desenvolvedores, contribuíram também em outros ambientes, inclusive enviando componentes de atualização (*patches*). Dois desses desenvolvedores se tornaram *committers* do projeto. Importante mencionar que esses desenvolvedores também não se limitaram ao projeto Hadoop e participaram de outros projetos da Apache Software Foundation.

Para os desenvolvedores que iniciaram sua contribuição a partir do gerenciador de tarefas, pôde-se perceber que a grande parte contribuiu com a lista de desenvolvedores e usuários. Nenhum deles se tornou *committer* oficial do projeto, muito embora tenham contribuído com mais componentes de atualizações. Uma inspeção identificou que esses usuários enviaram componentes de atualização para tarefas pontuais, normalmente tarefas que eles mesmos relataram como defeitos, ou como uma nova funcionalidade.

TABELA X. FORMAS DE COLABORAÇÃO DOS USUÁRIOS QUE CONTINUARAM NO PROJETO

	Apareceram como novatos apenas na Lista de Desenv.	***Apareceram como novatos apenas no Jira***	***Apareceram como novatos em ambos***
# membros	**7**	**12**	**5**
Contribuíram na lista de Desenv.	7 (100%)	10 (83,3%)	5 (100%)
Contribuíram na Lista de Usuários	7 (100%)	9 (75%)	5 (100%)
Contribuíram no Jira	6 (85,7%)	12 (100%)	5 (100%)
É *committer*	2 (28,6%)	0 (0%)	1 (20%)
Continuam até 2012	3 (42.9%)	7 (58,3%)	3 (60%)
Trabalhou em outros projetos	6 (85,7%)	12 (100%)	5 (100%)

Também foram avaliados os desenvolvedores que começaram contribuindo na lista de desenvolvedores e no gerenciador de tarefas simultaneamente. Esses desenvolvedores são a menor quantidade, mas é importante observar que todos participam de outros projetos, da lista de usuários, e três dos cinco continuaram no projeto até maio de 2012.

V. AMEAÇAS À VALIDADE

A presente seção discute as ameaças à validade que podem ter influenciado o estudo. As próximas três subseções apresentam as ameaças à validade interna, externa e de construção. Os riscos à **validade interna** estão relacionados com influências que ocorrem nas variáveis independentes, sem a ciência do pesquisador, causando risco de uma possível conclusão sobre um relacionamento entre o tratamento e o resultado [15]. Os riscos à **validade externa** são condições que limitam a capacidade de generalização dos resultados a uma população mais ampla [15]. Os riscos à **validade de construção** dizem respeito à generalização do resultado do experimento com os conceitos ou teorias que o apoiam [15].

A. Validade Interna

Apesar da coleta de dados ter sido realizada para um período relativamente grande, foram encontrados poucos novatos que permaneceram no projeto. Esse pequeno número de novatos e de mensagens enviadas por eles pode influenciar os resultados pela falta de densidade de dados existentes.

Outros fatores poderiam ter sido considerados como razões de desistência. Para reduzir essa ameaça, foi enviado o questionário aos desistentes. Entretanto, a baixa taxa de respostas ao questionário não possibilitou uma avaliação mais completa dos dados.

B. Validade Externa

A validade do presente estudo limita-se ao projeto Hadoop Common. As conclusões e discussões apresentadas são específicas para o projeto. Para um resultado com conclusões para uma população mais ampla, é necessário analisar uma amostra significativa de projetos e períodos de verificação. Como trabalho futuro, pretende-se realizar uma pesquisa em um ecossistema que pode trazer resultados mais genéricos e confirmar os resultados apresentados nesse estudo.

C. Validade de Construção

As medidas utilizadas neste artigo podem não ser a melhor maneira de mostrar os resultados, podendo ser interpretada de maneiras diferentes. Não foram encontrados trabalhos que ofereçam outros meios para medição ou que possibilitem comparação ou confirmação dos resultados obtidos.

A classificação manual das perguntas de novatos e das respostas está sujeita a erros, por ter sido realizada por humanos. Não é possível garantir que a classificação tenha abrangido todas as situações de perguntas e respostas. Para reduzir essa ameaça as perguntas e respostas foram analisadas por dois pesquisadores que discutiam até chegar a um consenso sobre a classificação.

As janelas de tempo escolhidas podem ter afetado as observações. Alterar o tamanho dos períodos de tempo ou alterar o início e fim dos intervalos podem gerar observações diferentes. Usuários que fizeram submissões pontuais em diferentes intervalos e podem ter sido classificadas erroneamente, como novatos, podem ter causado algum viés na análise.

Podem existir questões feitas por novatos que não tenham sido classificadas como perguntas na lista de desenvolvedores. Isto porque foram consideradas como perguntas apenas a primeira mensagem enviada na árvore de discussão. Entretanto, pode haver casos de interações que foram iniciadas como respostas a discussões já existentes.

Usuários podem ter dois *logins* no gerenciador de tarefas ou participar da lista de desenvolvedores com dois e-mails diferentes. Pode, ainda, haver casos em que o mesmo usuário está cadastrado no gerenciador de tarefas e na lista de desenvolvedores com e-mails diferentes. Para reduzir esta ameaça, o coletor de e-mails tenta combinar endereços diferentes utilizados por um mesmo usuário. Foi também

realizada uma análise manual para combinar e-mails e usuários do Jira.

VI. CONCLUSÃO

Este trabalho apresentou uma análise da desistência de novatos em projetos de software livre, observando o projeto Hadoop Common. O método aplicado possibilitou entender algumas das razões para a desistência, com base nas primeiras interações dos novatos na lista de desenvolvedores e no gerenciador de tarefas.

Foi possível evidenciar que a taxa de novatos que continuaram é pequena, sendo de 18% na lista de e-mails e de 13% no gerenciador de tarefas. Dentre os fatores que influenciam a desistência, há indícios de que o recebimento de respostas inadequadas e a experiência da pessoa que responde interferem na decisão dos novatos. Em contrapartida, concluiu-se que a falta de resposta não é um fator de grande influência.

Na lista de e-mails, após leitura das discussões, pode-se concluir que novatos com intenção de iniciar no projeto, mas que têm questões respondidas por outros novatos, têm uma tendência maior a desistir. Percebeu-se, durante a análise manual, haver novatos que respondiam a questões de maneira equivocada ou que meramente replicavam a intenção de entrar no projeto. Entretanto, para uma análise mais realista seria necessário entrevistar cada um dos membros que desistiram, a fim de entender as razões da entrada e desistência. Um contato com esses novatos foi realizado e os resultados são apresentados na Seção IV.E.

A classificação das respostas a novatos mostrou ainda que as mensagens negativas ou de direcionamento podem influenciar a desistência. Algumas evidências desse comportamento foram obtidas com as respostas ao questionário enviado aos desistentes. Seis das 13 respostas recebidas (46,15%) mostraram novatos insatisfeitos com as respostas recebidas, pois não conseguiram encontrar ajuda necessária para dar os primeiros passos. As respostas ao questionário mostraram ainda que fatores externos podem ocasionar a desistência. Nesse caso, houve quatro pessoas (28,57%) que informaram ter saído do projeto devido à mudança de foco (projeto, empresa).

Com relação aos novatos que continuaram no projeto, a análise mostrou que os membros tendem a contribuir mais e diversificar suas ações. Treze dos 24 dos novatos que continuaram (54,17%) estão ativos até maio de 2012 no gerenciador de tarefa Jira, dos quais três tornaram-se *commiters*. Dos 24 novatos que continuaram, 23 (95,83%) participaram de outros projetos dentro do Hadoop. Entretanto, a participação nas listas de e-mail (mesmo dos novatos que iniciaram a participação por esse meio) decresceu com o tempo.

Percebeu-se, pela leitura das discussões via e-mail iniciadas por desistentes, que grande parte dos novatos não tinha intenção de seguir no projeto. Muitos deles se relacionaram apenas uma vez com a lista de e-mails para esclarecer uma dúvida pontual, recebiam respostas corretas, parte das vezes agradecia e não mais voltavam.

Uma classificação mais rigorosa e uma análise automatizada do conteúdo das mensagens e do padrão das discussões devem ser realizadas para esclarecer o real impacto das mensagens na desistência ou no interesse de um novato participar de uma tarefa.

Este trabalho foi o primeiro passo para entender como os desenvolvedores colaboram em projetos de software livre, e, fundamentalmente, como os novatos se comportam nessas comunidades. Entender esse comportamento é importante para criar mecanismos de recomendação e recepção desses novatos e aumentar a sua retenção nos projetos.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pela Fundação Araucária sendo parte do projeto "Análise de redes sociais de desenvolvedores para predição de bugs em projetos de software" e do Doutorado Interinstitucional entre IME/USP e UTFPR. O pesquisador Marco Aurélio Gerosa é bolsista produtividade do CNPq.

REFERÊNCIAS

[1] Y. Park and C. Jensen, "Beyond pretty pictures: Examining the benefits of code visualization for Open Source newcomers", in 5th IEEE International Workshop on Visualizing Software for Understanding and Analysis (VISSOFT), 2009, pp. 3-10.

[2] B. Dagenais, H. Ossher, R.K.E Bellamy, M.P. Robillard and J.P. de Vries, "Moving into a new software project landscape", in 2010 ACM/IEEE 32nd International Conference on Software Engineering, 2010, pp. 275-284.

[3] W. Scacchi, "Understanding the requirements for developing open source software systems", in IEE Proceedings Software, v. 149, no. 1, 2002, pp. 24-39.

[4] I. Steinmacher, I.S. Wiese, M.A. Gerosa, "Recommending Mentors to Software Project Newcomers". In Proc. of 3rd International Workshop on Recommendation Systems for Software Engineering (RSSE '12). IEEE CS, 2012, pp. 63-67.

[5] G.C. Cubranic, C. Murphy, K. Singer, K.S. Booth, "Hipikat: a project memory for software development", IEEE Trans. Softw. Eng., v. 31, no. 6, 2005, pp. 446-465.

[6] G. von Krogh, S. Spaeth, K.R. Lakhani, "Community, joining, and specialization in open source software innovation: a case study", Res. Policy, v. 32, no. 7, 2003, pp. 1217-1241.

[7] C. Jensen, S. King, V. Kuechler, "Joining Free/Open Source Software Communities: An Analysis of Newbies' First Interactions on Project Mailing Lists", in Proceedings of the 44th Hawaii International Conference on System Sciences, 2011, pp. 1-10.

[8] C. Jensen, W. Scacchi, "Role Migration and Advancement Processes in OSSD Projects: A Comparative Case Study". Proceedings of the 29th International Conference on Software Engineering, 2007, pp. 364-374.

[9] C. Jergensen, A. Sarma, P. Wagstrom, "The onion patch: migration in open source ecosystems". In Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European conference on Foundations of Software Engineering, 2011, pp. 70-80.

[10] K. Nakakoji, Y. Yamamoto, Y. Nishinaka, K. Kishida, Y. Ye, "Evolution patterns of open-source software systems and communities". In Proceedings *of* the International Workshop on Principles of Software Evolution (IWPSE '02). ACM, New York, NY, USA, 2002, pp. 76-85.

[11] Y. Qu, C. Huang, P. Zhang, J. Zhang, "Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake". In Proceedings of the Conference on Computer supported cooperative work (CSCW '11), 2011, pp. 25-34.

[12] P. Vora, N. Komura, "The n00b Wikipedia Editing Experience". In Proceedings of the 6th International Symposium on Wikis and Open Collaboration (WikiSym '10), 2010, Article No 36 , 3 pp.

[13] J. Thom-Santelli, D. Cosley, G. Gay, “What do you know?: experts, novices and territoriality in collaborative systems”. In Proceedings of the 28th international conference on Human factors in computing systems (CHI '10). 2010, pp. 1685-1694.

[14] A. Halfaker, A. Kittur, J. Riedl, “Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work”. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11), 2011, pp. 163-172.

[15] C. Wohlin, P. Runeson, M. Höst, “Experimentation in Software Engineering: An Introduction”. Kluwer Academic Publisher, 2000.