A Time-Lag Analysis for Improving Communication among OSS Developers

Masao Ohira, Kiwako Koyama, Akinori Ihara,
Shinsuke Matsumoto, Yasutaka Kamei, and Ken-ichi Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama, Ikoma, Nara, Japan
{masao,kiwako-k,akinori-i,shinsuke-m,yasuta-k,matumoto}@is.naist.jp

Abstract. In the open source software (OSS) development environment, a communication time-lag among developers is more likely to happen due to time differences among locations of developers and differences of working hours for OSS development. A means for effective communication among OSS developers has been increasingly demanded in recent years, since an OSS product and its users requires a prompt response to issues such as defects and security vulnerabilities. In this paper, we propose an analysis method for observing the time-lag of communication among developers in an OSS project and then facilitating the communication.

Keywords: time-lag analysis, OSS, distributed development.

1 Introduction

Open source software (OSS) such as Linux and Apache is generally developed by globally distributed developers. Unlike commercial software development in a company, OSS development does not necessarily request developers to engage in development at a designated time and location. OSS developers may voluntarily decide whether they continue to dedicate themselves to OSS development or not.

In this OSS development environment, a time-lag occurs in communication among developers more than a little, because of differences of time zones among geographically-distributed developers with a variety of lifestyles. For instance, according to the geographical distribution of registered users at SourceForge which was reported by Robles and Gonzalez-Barahona [1], the top three regions by the number of registered developers at SourceForge are North America, West Europe, and China. Since the time-lag among those regions is at least more than five hours, it would not be easy to discuss among developers in real-time. Furthermore, even if developers reside in the same time zone, it is not still guaranteed that developers can communicate each other in real time, because each developer has no constraint on working hours.

The goal of our research is to construct a support mechanism for effective communication among geographically-distributed OSS developers. As a first step toward achieving the goal, in this paper we present an analysis method for helping

K. Nakakoji, Y. Murakami, and E. McCready (Eds.): JSAI-isAI, LNAI 6284, pp. 135–146, 2010.

OSS developers comprehend the whole picture of the communication time-lag occurred in a OSS project. The analysis method targets a mailing list archive as a data source, and consists of three kinds of analyses as follows;

1. analysis of a geographical distribution and activity time of OSS developers
2. analysis of a distribution of time required for information exchanges among OSS developers in different locations, and
3. analysis of appropriate timing for sending messages.

From a case study with Python project data, this paper explores the usefulness of the analysis method.

2 Analysis Method

This section describes data extraction, conversion and classification which are necessary in advance of performing our analysis.

2.1 Preparation

(1) Data extraction and conversion. The target data source for our analysis is archives of mailing lists which are used by OSS developers to exchange information. The reason we select mailing list archives as the target data for our analysis is because mailing lists are widely used in OSS projects. We consider that data of mailing list archives allows us to reveal the whole picture of the existence of the time-lag in many OSS projects.

In order to apply the analysis method to the target data, firstly we need to extract information of **posted date and time**, and **posted locations** from mailing list archives (i.e., from e-mail headers). In what follows, "posted date and time" means local date and time of a message's sender, and "posted locations" is presented as a time-lag between Coordinated Universal Time (UTC) and local time. For instance, "**UTC+9**" means the location of Japan because the standard time of Japan is nine hours prior to UTC.

Fig. 1 shows the procedure of data extraction and conversion. When a developer posts a message to a mailing list, the message is delivered to subscribed developers of the mailing list. Replying to the post, the other developers can discuss the message. Using such the post-reply relationship (i.e., thread structure) in a mailing list, we extract[1] information on posted/replied date and time, and locations (time zones) from mailing list archives. For instance, from a thread structure illustrated in Fig. 1 (a), we extract information of posted and replied messages as the table in Fig. 1 (b). Then we convert the information into post-reply relationships as the table in Fig. 1 (c) and calculate a time-lag from a difference between posted and replied date and time. Note that we suppose that message B replied to message A can be a posted message for message C.

[1] We do not collect data from posted messages with no replies.

(a) a thread in a mailing list

post A : _____
- reply B : Re:_____
  - reply C : Re:Re:_____
- reply D : Re:_____

(b) a list of information of posted/replied messages

ID	posted date and time	location (time zone)	Coordinated Universal Time (UTC)
message A	2009/04/29 10:33:53	UTC+9	2009/04/29 01:33:53
message B	2009/04/28 22:40:04	UTC−4	2009/04/29 02:40:04
message C	2009/04/29 09:12:30	UTC+3	2009/04/29 06:12:30
message D	2009/04/29 02:26:59	UTC−10	2009/04/29 12:26:59

(c) a relationship among posted and replied messages

info. of posted messages			info. of replied messages			time lag (hours)
ID	posted date and time	location	ID	replied date	location	
message A	2009/04/29 10:33:53	UTC+9	message B	2009/04/28 22:40:04	UTC−4	1.10
message B	2009/04/28 22:40:04	UTC−4	message C	2009/04/29 09:12:30	UTC+3	3.54
message A	2009/04/29 10:33:53	UTC+9	message D	2009/04/29 02:26:59	UTC−10	10.88

Fig. 1. Data extraction and conversion

(2) Classification of data. Several factors such as differences of time zones (i.e., countries and/or regions) and differences of developers' working hours may have an influence on time-lags between posted time and replied time. For instance, communication among developers living in different time zones might be prolonged due to differences of lifestyles (e.g., dinner time or sleeping time). Developers in the same time zone also might be difficult to communicate each other in real time, because each developer has no constraint on working hours.

In order to distinguish between the time-lag due to time zone differences and the time-lag due to lifestyle differences, the collected data described above is classified into data **within** and **over** the time-lag of 24 hours. Many of replied messages within 24 hours after a post would be affected by differences of time zones, while replied messages over 24 hours after a post would be generated by differences of developers' lifestyles and/or difficulty of the content of a posted message, rather than geographical differences among developers. For these reasons, our analysis method targets the data of posted and replied messages within the 24 hours time-lag.

2.2 Procedure

(1) Geographical distribution and activity time of OSS developers. In order to understand the existence of the communication time-lag in an OSS project, the analysis method firstly identifies a geographical distribution of developers of the project, counting the number of replied messages by each location (UTC−11∼UTC+12). The analysis method also identifies a distribution of the number of replied messages by local time in each location in order to understand working hours of developers by each location, since developers' working hour can

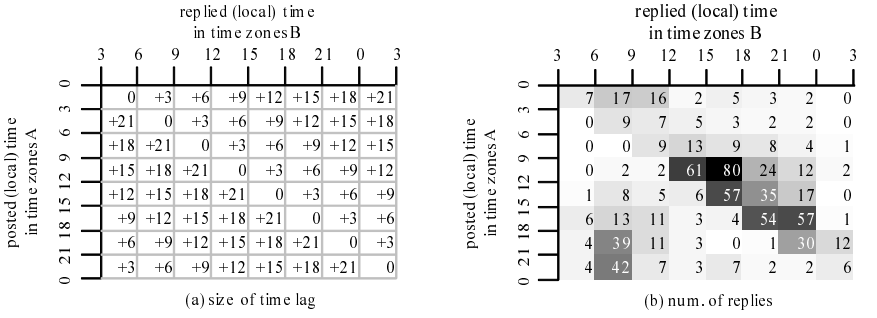

Fig. 2. Distribution of posted and replied time

differ even in the same location. By this means, we can identify active or inactive locations and working hours of OSS developers.

(2) Distribution of time required for information exchanges. In order to understand the communication time-lag due to the geographical (time zone) differences, the analysis method calculates distributions of time required for information exchanges among OSS developers in **different** locations and the **same** locations respectively. This helps us more clearly distinguish between the time-lag by the geographical differences and the time-lag by the differences of developers' lifestyles.

(3) Appropriate timing for sending messages. In order to identify the appropriate timing for communication which resolves communication time-lags as much as possible, the analysis method calculates the number of replied messages by each hour, using **posted (local) time** and **replied (local) time**. A numerical number in Fig. 2 (a) shows the size of a time-lag (hours) between time zones A and B. Fig.2 (b) shows the number of pairs of posted messages from time zone A and replied messages from time zones B. For instance, suppose that one developer in A post a message between 9 and 12, and other developer in B replies a message between 15 and 18. In this case, the time-lag is +3 hours and the number of post/reply pairs is 80.

Time zones A and B are fixed after selecting target locations for analysis. Time zones B in Fig. 2 is arranged as replied messages within an hour correspond to posted messages on the diagonal. In Fig. 2, the size of time-lag and the number of posted/replied messages are counted by three hours, but the length may be changed depends on analysis needs. Furthermore, the all cells in Fig. 2 (b) are gray-scaled according to the number of posted/replied pairs of messages, to grasp a big picture of time slots with a large or small number of replied messages.

Using Fig. 2 (a) and (b), it is possible to identify a time slot with a large or small time-lag. For instance, we can see that messages posted between 21 and 0 in time zones A (the bottom row in Fig. 2) tend to be replied after 6 hours. That is, to post messages from 21 to 0 would not be the appropriate timing for less time-lag communication.

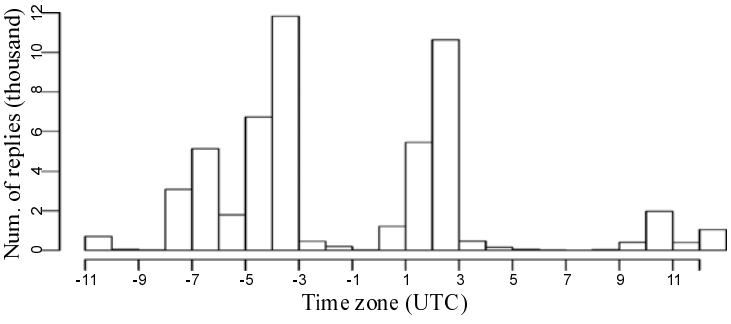

Fig. 3. Distribution of the number of replied messages by time zones

3 Case Study

This section describes a case study with a mailing list for developers in the Python project. Through the case study, we would like to confirm whether the analysis method can help us understand the existence of the time-lag in communication among OSS developers.

3.1 Python Project

Python[2] is an object oriented script language developed by OSS. It is very popular in Europe and the United States as well as Perl. Because it supports various platforms and provides rich documentations and libraries, it is used in a broad range of domains (e.g., Web programming, GUI-based applications, CAD, 3D modeling, formula manipulation, and so forth).

3.2 Target Data

We selected the mailing list archive called Python-Dev[3] which is for discussing development of Python such as new features, release and maintenance. We use the Python-Dev mailing list archive from April 1999 to April 2009, which have 89,301 messages. Excluding posted messages with no replies and messages with no information on posted/replied time and locations, posted and replied messages were 56,707. 51,830 of 56,707 messages were sent within 24 hours.

3.3 Analysis Results

(1) Geographical distribution and activity time. Fig. 3 shows a distribution of the number of replied messages by time zones. The X-axis and Y-axis respectively mean time zones and the number of replied messages. It indicates

[2] Python Programming Language, http://www.python.org/
[3] Python core developers ML, http://mail.python.org/mailman/listinfo/python-dev/

Table 1. Target locations for the case study of Python

region	time zone	location
North and South American continent	UTC−8∼ UTC−4	United States, Canada, West of Brazil, Chile, Bolivia, Mexico, etc.
European and African continent	UTC+0∼ UTC+3	Europe, Africa, Moscow, Iran, Saudi Arabia, etc.

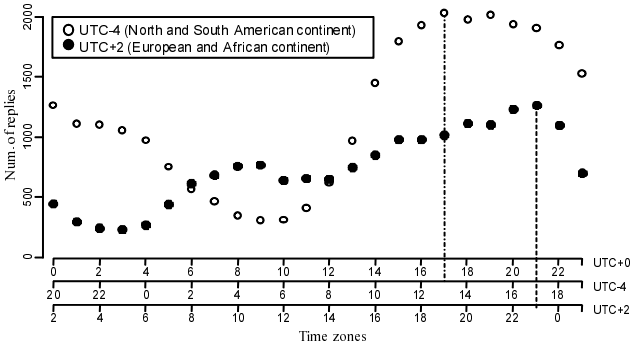

Fig. 4. Distribution of the number of replied messages by time slots (white circles: North and South American continent, black circles: European and African continent)

that a large number of messages are replied by developers from UTC−4 (East of the United States) and UTC+2 (central Europe) in the Python project. This result is not surprising at all, because Python is mainly used and developed by European and American developers. It would be natural that developers living in these locations actively communicated.

Many of countries in the locations of UTC−4 and UTC+2 is utilizing daylight-saving time. And countries around the countries in UTC−4 and UTC+2 also have many messages. So, we selected two regions around UTC−4 (the North and South American continent: UTC−8∼UTC−4) and UTC+2 (the European and African continent: UTC+0∼UTC+3) as the analysis target in this paper. Table 1 shows major countries included in these regions.

Fig. 4 shows transitions of replied messages by hour in the two regions which are determined from Fig.3. The X-axes shows time in the three time zones (UTC+0, UTC−4, UTC+2) and the Y-axis is the number of replied messages. It indicates that the maximum and minimum number of replied messages from the North and South American continent are attained respectively at 13 and 5 in the local time (UTC−4). Python developers in the North and South American continent seem to mainly communicate during daytime hours. In contrast, Python developers in the European and African continent actively communicate during nighttime hours, because the number of replied messages from the European and African continent is peaked at 23 in the local time (UTC+2). In this

Table 2. Statistics of time-lags by region (A: North and South American continent, E: European and African continent)

posted region → replied region	the number of replies	maximum (hours)	median (hours)	minimum (hours)
A → A	18,901	11.55	1.24	0.00
A → E	6,942	16.34	2.07	0.00
E → E	9,426	14.69	1.59	0.00
E → A	7,215	13.91	1.80	0.00

way, analyzing activity time of OSS developers by using the number of replied messages helps us understand the existence of the difference of working hours by region.

Although Fig. 4 provides an overview on the difference of working hours of OSS developers by region, however, it does not tell us anything about time-lags. In fact, developers in the both regions actively communicate each other from 12 to 23 in UTC+0. Communication time-lags might not exist in the regions. In contrast, developers in either one region or the other region does not actively communicate from 12 to 23 in UTC+0. Communication time-lags between developers living different locations might exist in this time period.

(2) Distribution of time required for information exchanges. Table 2 shows time spent to reply messages to the same and different time zones, the number of replied messages, and time-lags (maximum/median/minimum). A pair of a post from location X and a reply from location Y is represented as "X → Y".

The median hours of the time-lag among the same time zone was 1.24 hours for A → A and 1.59 hours for E → E. The median hours of the time-lag between the different time zones was 2.07 hours for A → E and 1.80 hours for E → A. Developers in the same time zone can expect to have a reply within 90 minutes, and developers between different time zones also can expect to have a reply within about 2 hours. Since the actual difference of the time-lag between the target regions is nearly 6 hours, we can consider that communication time-lags in the Python project are relatively small.

(3) Appropriate timing for sending messages. Fig. 5 (a), (b), (c) and (d) are distributions of the number of replied messages between two regions. For the simplicity, only gray-scaled figures without the number of replied messages are shown in Fig. 5.

We can see that the zero time-lag (i.e., dark gray cells near the diagonal line) is expected from 10 to 17 in Fig. 5 (a), from 9 to 17 in posted local time and from 15 to 23 in replied local time in Fig. 5 (b), from 16 to 23 in Fig. 5 (c), and from 16 to 23 in posted local time and from 10 to 17 in replied local time in Fig. 5 (d). For these time periods, developers would timely communicate each other.

In contrast, reply time seems to be delayed from 18 to 23 in posted local time in Fig. 5 (b) and from 7 to 13 in replied local time in Fig. 5 (d), because there

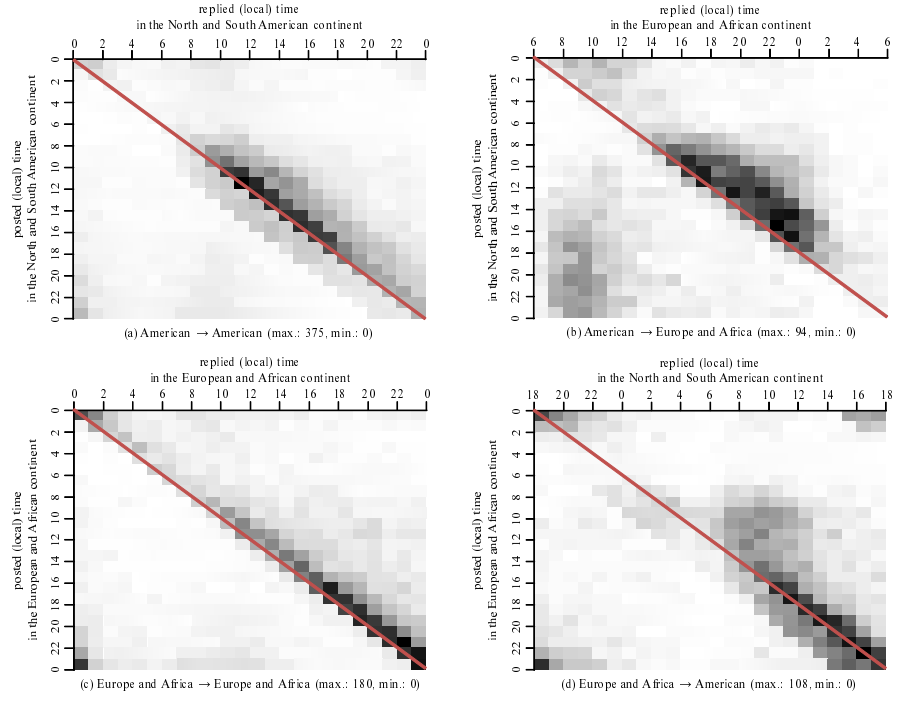

Fig. 5. Distributions of posted/replied local time between two regions

are darker cells a short distance away from the diagonal line. These two posted local time periods correspond to the time period from midnight to early morning i0 to 6) in replied locations, which means that developers in replied locations was sleeping at the posted time.

From the result of Fig. 5, in order to receive a quick reply, it would be desirable to post a message from 10 to 17 in the North and South American continent, and from 16 to 23 in the European and African continent. On the contrary, it is not appropriate timing to post a message from 18 to 23 in the North and South American continent, and from 7 to 13 in the European and African continent, since the time-lag is likely to occur.

In this way, our analysis method helps OSS developers know the appropriate timing so that they can resolve a time-lag of information exchange in an OSS project as much as possible.

4 Discussions

Opposite to what we expected before our case study, we have confirmed in Table 2 that the influence of the time-lag due to the time zone difference was relatively small in the Python project. One reason of this result might be that active time of Python developers is partly overlapping in the two regions. Although

there are about 6 hours time-zone difference between the two regions, the active time in the North and South American continent was different from that in the European and African continent as shown in Fig. 4. Therefore, active hours of Python developers in the two regions might overlap by coincidence from 10 to 17 in the North and South American continent (from 16 to 23 in the European and African continent). Another reason may be that the number of Python developers subscribed to the "Python-Dev" mailing list is sufficiently-large to quickly respond to a posted message at any time.

Our analysis method is not only useful in knowing the appropriate timing for communication among geographically-distributed OSS developers, but also useful in changing communication media used in an project. For instance, when a project replaces mailing lists with IRCs (Internet Relay Chat) as communication media, developers would be required to more precisely understand the appropriate timing for communication to resolve the time-lag. In that case, our method would help developers know the better timing for real-time communication.

OSS developers are not necessary to be geographically-distributed, but they may be at the same region or location. Though our analysis method mainly aims to understand the communication time-lag arising from time-zone differences, it can be used for the time-lag due to lifestyle differences of OSS developers in the same region or location. OSS developers have no constraint on their working hours and they can freely engage in OSS development. At the same region, some developers can work in the morning and other developers can develop OSS at midnight. Depending on the differences of lifestyles of developers, time-lags could happen even if they live close to each other. In this situation, our method can provide an insight on the differences of active time in the same region and help developers understand the appropriate timing for sending messages.

The analysis method also can be used for distributed development in a company. Working hours in a company are fixed to some extent, but it is not necessarily that a developer in one site can communicate with other developers in another site at a particular time. In the prior study [2], time zone differences are visualized to understand and exploit overlapping hours in a distributed environment. Our method can not only visualize the time zone differences, but also allows developers to understand the easiness of communication at a particular time period, using the number of replied messages (i.e., density of working activity at a particular time period).

In this paper, we introduce the time-lag analysis method toward improving the communication efficiency of geographically-distributed OSS developers. The analysis method targets mailing list archive data as communication logs to reveal the existence of communication time-lags. Although IRC communication is often used in OSS projects and they can be our analysis target, communication using IRC do not work when developers one wishes to talk are off-line. So, IRC communication logs are not likely to well-capture communication time-lags.

In this paper, we have conducted a case study of the Python project, using the "Python-Dev" mailing list archive. Python-Dev consists of mailing list archive data for about 10 years. So, it might be too large to show communication

time-lags among Python developers at the fine-grained level. Actually, we have observed that communication time-lags in the Python project were relatively small. We suspect that this results from the size population of developers (subscribers) of Python-Dev. In Python-Dev, a posted message must be read by a number of developers in the world and so it might be easy to have replies. In order to emphasize the existence of time-lags and its issues, in the near future, we need to analyze more specific situations such as the level of communication among module owners, reviewers and patch contributors.

5 Related Work

The issues on the communication time-lag or delay in OSS development have been intensively studied in relation to bug modification processes with bug tracking systems [3,4,5,6,7,8,9,10,11,12,13]. For instance, Wang et al. [12] proposed several metrics to measure the evolution of open source software. The metrics include the number of bugs in software, the number of modified bugs and so on. As a result of the case study using the Ubuntu project (one of Linux-based operating system distributors), the study found that about 20% of all the reported bugs were actually resolved and over ten thousand bugs were not assigned to developers. These findings indicate that it takes a long time to resolve all bugs reported into bug tracking systems and that it also takes a long time to start modifying bugs. The study, however, did not reveal the amount of time or communication time-lags to resolve bugs.

Mockus et al. [10] and Herraiz et al. [6] have reported studies on the mean time to resolve bugs in open source software development. Mockus et al. have conducted two case studies of the Apache and Mozilla projects to reveal success factors of open source software development. In the case studies, they analyzed the mean time to resolve bugs because rapid modifications of software bugs are generally demanded by users. As a result of the analysis, they have found that the mean time to resolve bugs were short if bugs existed in modules regarding to kernel and protocol, and existed in modules with widely-used functions. They also found that 50% of bugs with the priority P1 and P3 were resolved within 30 days, 50% of bugs with P2 were resolved within 80 days, and 50% of bugs with P4 and P5 were resolved within 1000 days. While [10,6] mainly focused on precise understandings of bug modification processes in open source software development, we are interested in the influence of communication time-lags among developers on the bug modification process.

The issues on differences of time-zone and/or geographical distance in distributed development rather have been discussed in terms of the context of corporate (proprietary) software development [14,15,16,17,18]. For instance, Harbsleb et al.[16] have compared single-site development with multi-sites development and then revealed that development in the distributed environment introduced the delay of development speed. In contrast, Bird et al. [19] analyzed the development of Windows Vista by comparing distributed teams with collocated teams from the aspect of the post-release failures of components. They have found a

slight difference in failures, but the difference has been less significant. Nguyen et al.[20] also reported the similar phenomena in the Eclipse Jazz project. Although the lessons learned from these studies on distributed software development provides us a lot of useful insights, they are partly applicable to geographically-distributed OSS development due to the differences of lifestyles of developers even in the same region or location. In this paper, we tried to tackle this unique feature of time-lags in OSS development.

6 Conclusion and Future Work

In this paper, we proposed a method for analyzing a communication time-lag among OSS developers. As a result of our case study of the Python developers' mailing list archive, we could confirm that our analysis method helps geographically-distributed OSS developers understand: (1) active time of developers are different from regions, (2) communication time-lags in the Python project is relatively small, and (3) there exists the appropriate timing for resolving communication time-lags as much as possible. In this paper, our analysis method targets communication time-lags in the two regions with the time zone difference. In the future, we need to analyze regions and/or locations without time zone differences in order to better understand the influence of lifestyle differences of developers on communication time-lags. As described before, we still need to analyze more specific situations of time-lags at the fine-grained level.

Acknowledgment

This research is being conducted as a part of the Next Generation IT Program and Grant-in-aid for Young Scientists (B)–20700028, 21–8995C20–9220 by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Robles, G., Gonzalez-Barahona, J.M.: Geographic location of developers at sourceforge. In: The 2006 international workshop on Mining Software Repositories (MSR 2006), pp. 144–150 (2006)
2. Laredo, J.A., Ranjan, R.: Continuous improvement through iterative development in a multi-geography. In: The 2008 IEEE International Conference on Global Software Engineering (ICGSE 2008), pp. 232–236 (2008)
3. Bettenburg, N., Just, S., Schröter, A., Weiss, C., Premraj, R., Zimmermann, T.: What makes a good bug report? In: The 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering (SIGSOFT 2008/FSE-16), pp. 308–318 (2008)
4. Colazo, J.A.: Following the sun: Exploring productivity in temporally dispersed teams. In: The Fourteenth Americas Conference on Information Systems (AMCIS 2008). Paper no. 240 (2008)

5. Godfrey, M.W., Tu, Q.: Evolution in open source software: A case study. In: 16th IEEE International Conference on Software Maintenance (ICSM 2000), pp. 131–142 (2000)
6. Herraiz, I., German, D.M., Gonzalez-Barahona, J.M., Robles, G.: Towards a simplification of the bug report form in eclipse. In: The 2008 international Working Conference on Mining Software Repositories (MSR 2008), pp. 145–148 (2008)
7. Ihara, A., Ohira, M., Matsumoto, K.: An analysis method for improving a bug modification process in open source software development. In: The joint international and annual ERCIM workshops on Principles of Software Evolution and Software Evolution Workshops (IWPSE-Evol 2009), pp. 135–144 (2009)
8. Kim, S., Pan, K., Whitehead, E.J.: Memories of bug fixes. In: The 14th ACM SIGSOFT international symposium on Foundations of software engineering (SIGSOFT 2006/FSE-14), pp. 35–45 (2006)
9. Kim, S., Zimmermann, T., Whitehead, E.J.: Automatic identification of bug-introducing changes. In: The 21st IEEE/ACM International Conference on Automated Software Engineering (ASE 2006), pp. 81–90 (2006)
10. Mockus, A., Fielding, R.T., Herbsleb, J.D.: Two case studies of open source software development: Apache and mozilla. ACM Transactions on Software Engineering and Methodology (TOSEM) 11(3), 309–346 (2002)
11. Śliwersk, J., Zimmermann, T., Zeller, A.: When do changes induce fixes? In: The 2005 International Workshop on Mining Software Repositories (MSR 2005), pp. 1–5 (2005)
12. Wang, Y., Guo, D., Shi, H.: Measuring the evolution of open source software systems with their communities. ACM SIGSOFT Software Engineering Notes 32(6), Article No.7 (2007)
13. Yilmaz, C., Williams, C.: An automated model-based debugging approach. In: The twenty-second IEEE/ACM international conference on Automated Software Engineering (ASE 2007), pp. 174–183 (2007)
14. Carmel, E.: Global software teams: collaborating across borders and time zones. Prentice Hall PTR, Upper Saddle River (1999)
15. Karolak, D.W.: Global Software Development: Managing Virtual Teams and Environments. Wiley-IEEE Computer Society Press, Los Alamitos (1999)
16. Herbsleb, J.D., Mockus, A., Finholt, T.A., Grinter, R.E.: An empirical study of global software development: distance and speed. In: The 23rd International Conference on Software Engineering (ICSE 2001), pp. 81–90 (2001)
17. Milewski, A.E., Tremaine, M., Egan, R., Zhang, S., Kobler, F., O'Sullivan, P.: Guidelines for effective bridging in global software engineering. In: The 2008 IEEE International Conference on Global Software Engineering (ICGSE 2008), pp. 23–32 (2008)
18. Sangwan, R., Bass, M., Mullick, N., Paulish, D.J., Kazmeier, J.: Global Software Development Handbook. Auerbach Series on Applied Software Engineering Series. Auerbach Publications, Boston (2006)
19. Bird, C., Nagappan, N., Devanbu, P., Gall, H., Murphy, B.: Does distributed development affect software quality? an empirical case study of windows vista. In: The 2009 IEEE 31st International Conference on Software Engineering (ICSE 2009), pp. 518–528 (2009)
20. Nguyen, T., Wolf, T., Damian, D.: Global software development and delay: Does distance still matter? In: The 2008 IEEE International Conference on Global Software Engineering (ICGSE 2008), pp. 45–54 (2008)