Techniques for Identifying the Country Origin of Mailing List Participants

Ran Tang
Dept. of Elec. and Comp. Eng.
Queen's University
Kingston, Ontario, Canada
ran.tang@queensu.ca

Ahmed E. Hassan
School of Computing
Queen's University
Kingston, Ontario, Canada
ahmed@cs.queensu.ca

Ying Zou
Dept. of Elec. and Comp. Eng.
Queen's University
Kingston, Ontario, Canada
ying.zou@queensu.ca

Abstract — **Developer mailing lists play a central role in facilitating communication in open source projects. Participants from different countries and across diverse time zones discuss and resolve important design decisions or conflicts. A good understanding of the social structure of these mailing lists helps in managing these projects and in shaping their implementation structure (i.e., design and architecture). In this paper, we present a technique to determine the country of a mailing list participant. A case study on the developer mailing list for two large open source projects (i.e., PostgreSQL and GTK+) shows that our technique outperforms prior technique.**

Keywords — Global software development; Mining Software Repositories

I. INTRODUCTION

Mailing lists are one of the most important communication channels for open source projects. Globally distributed participants use mailing lists to communicate various aspects of the projects. Design discussions, project decisions, and requirement gathering are frequently fulfilled through the mailing lists due to the globally distributed pool of participants. The discussions and decisions are archived and preserved in mailing list repositories. By mining these repositories, we can understand the social structure of open source projects. Such social structure impacts and affects the technical structure (i.e., design and architecture) of the software. For instance, Conway's conjecture hypothesizes that the code ownership architecture serves as a predictor of the concrete (i.e., as built) architecture. Work by Bowman and Holt [3] shows that this conjecture holds for large commercial and open source projects, like Mozilla and Linux.

By studying the social structure in the mailing lists we can identify the location of live design knowledge in projects and the flow of such knowledge across the participants of the project. Such understanding is instrumental in global software development with managers working on distributing projects across a global pool of developers [4]. In particular, we wish to study how people from different countries and diverse time zones participate in mailing lists and how they interact. Prior work [8] shows that the participant pool of the mailing list for open source projects is international. However, prior work uses estimation techniques to derive an aggregate breakdown of the countries of the participants. Due to the limited number of identified participant countries, the output of prior techniques cannot be used to perform detailed analysis of the interaction on the mailing list.

In this paper, we present a technique to determine the country of a mailing list participant. Using this technique we can analyze the participation and interaction patterns. Our technique can determine the country of 67% of participants on a mailing list. This represents an 80% improvement over prior techniques (e.g.,[8]).

Organization of the paper. The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents our technique to identify the country of a participant. Section 4 presents our case study. Finally, Section 5 concludes the paper.

II. RELATED WORK

Prior work on studying the global pool of participants in open source projects can be categorized into two groups: research based on surveys and research based on mining repositories.

Research based on surveys. The surveys are conducted to identify the countries of participants involved in open source projects. Robles et al. [13] surveyed over 5,500 respondents and showed that a majority of open source developers are from Europe. Similar results are also reported in the survey conducted by Ghosh [7] and David et al. [5].

Research based on mining repositories. Dempsey et al. [6] analyzed the top-level domain name of the email address (e.g., .ca, .com) of the participant to identify the country of the participant. However the study did not compensate for the US bias resulting from the wide use of generic domains (e.g., .com). Prior techniques do not map .com addresses to any country. Therefore, the participants from the US may be underrepresented in the analysis. Studies in [11] [13] show that the developer pool becomes more European-based over time. Robles and Gonzalez-Barahona [8] [12] used a technique to identify countries of participants in SourceForge [16] open source projects. Email address and time zone information in the user profile are analyzed to infer the country. The mailing list was also studied using a similar technique.

1095-1350/09 $26.00 © 2009 IEEE
DOI 10.1109/WCRE.2009.46

Received: from postgresql.org ([200.46.204.86])
by localhost (mx1.hub.org [200.46.204.183]) (amavisd-maia, port 10024)
with ESMTP id 30975-04 for <pgsql-hackers-postgresql.org@postgresql.org>; Wed,
22 Oct 2008 23:40:39 -0300 (ADT)
Received: from dolly.its.queensu.ca (outgoing.QueensU.CA [130.15.241.183])
by postgresql.org (Postfix) with ESMTP id 5EE0864FC5F for
<pgsql-hackers@postgresql.org>; Wed, 22 Oct 2008 23:40:39 -0300 (ADT)
Received: from RanPC (DU129.N46.QueensU.CA [**130.15.46.129**])
by mta01.its.queensu.ca
(Sun Java System Messaging Server 6.2-7.05 (built Sep 5 2006))
with ESMTPA id <0K9600FVZ4NOIH00@mta01.its.queensu.ca> for
pgsql-hackers@postgresql.org; Wed, 22 Oct 2008 21:40:37 -0400 (EDT)
Date: Wed, 22 Oct 2008 21:40:33 **-0400**
From: Ran Tang <**8rt4@queensu.ca**>
Subject: [HACKERS] Can anyone explain to me how the "ps_OuterTupleSlot" in
PlanState is being used in implementing HashJoin?
To: pgsql-hackers@postgresql.org
Message-id: <3CB8AFE7D4C3400B980E1B91E441ED87@RanPC>
Hello there,
Can anyone explain to me how the "ps_OuterTupleSlot" in JoinState is being used in implementing HashJoin?
I looked at the source code. When it find a tuple in outer relation. It store the outer tuple to field : node->js.ps.ps_OuterTupleSlot
But I can't find how this field is being used in the following processing.
Thanks,
Ran

Figure 1. An example email header and body

However, the time zone information in the mailing list does not contain specific country information. Therefore, the analysis of the time zone can only derive the origins of participants to specific time zone regions instead of particular countries.

III. IDENTIFYING THE COUNTRY OF A PARTICIPANT

Prior work (e.g., [6], [8]) primarily used the top-level domain name of an email address to identify a participant's country. Such work can identify a limited number of participants since many participants use email addresses that are not specific to any country, such as "hotmail.com". This technique in turn limits our ability to perform detailed analysis of the interaction on the mailing list. To overcome this limitation, we use the IP address of the email sender to enhance prior techniques. Using the IP address of the email sender, we can identify the country of a larger number of participants. We can then study the mailing list repository to understand participation patterns.

The IP address information is stored in the **Received field** in the header of an email. The Received field contains tracking information generated by each mail server which has routed the message. Generally, a message is relayed through multiple mail servers before it reaches its destination. There may be multiple Received fields in an email header. Each Received field represents one hop in the relay. In each Received field, the IP addresses or domain names of the receiver and sender are recorded. The first hop is between the sender's computer and the first mail server. As a result, the sender's IP address and/or domain name are recorded in the Received field for the first hop. As shown in Figure 1, the Received field for the first hop contains the IP address of the sender (i.e., 130.15.46.129). It can be used to determine the true location of a participant instead of primarily relying on the top-level domain name of email address as done by prior approaches.

Figure 2 shows a high level overview of our technique which consists of four steps: email header extraction, top-level domain analysis, sender IP address analysis and conflict resolution.

A. *Email header extraction*

We first find all sender IP addresses for each email address (e.g., a@b.fm). Each IP address represents one location where the participant sent an email. The extracted email address and the IP addresses are used as the input for the following two analyses: top-level domain analysis and sender IP address analysis. For the example shown in Figure 2, five IP addresses are collected for the participant a@b.fm.

B. *Top-level domain analysis*

We examine the top-level domain for each email address. A top-level domain can be classified into three categories:

1) Country code domains are reserved for a country. For example, .ca is reserved for Canada. A country code is a good indicator of a participant's country with high confidence.

2) Resalable country code domains are registered to specific countries, but are often used for commercial

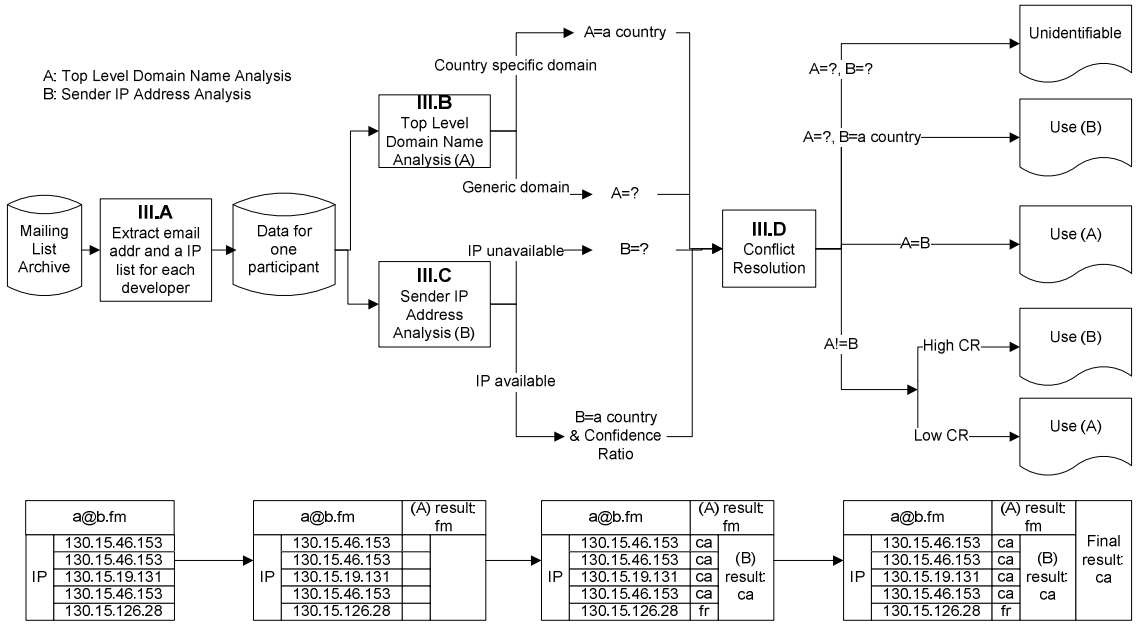

Figure 2. An overview of our technique to identify the country of a participant (Each step is marked with the corresponding Section numbers in the text)

purposes worldwide. For example, .fm is the country code domain for the Federated States of Micronesia, a small island-state in the South Pacific. However, .fm is actually used for radio stations in other countries. Such domains are less reliable to indicate the country of a participant. Examples of such domains are: .fm, .im, .it, and .tv.

3) Generic domains are not specific to a particular country. Generic domains include .org, .net, .name, .info, .com and .biz. These domains do not indicate the country of a participant. However, other generic domain names, such as .mil, .edu and .gov, are specific to the US and can be used to denote the country of a participant.

For the example shown in Figure 2, the top-level domain analysis assigns the participant, with the a@b.fm email address, to the Federated States of Micronesia.

C. Sender IP address analysis

Using an email address from a generic domain, we cannot determine the country of a participant. We further examine the sender IP address for each email sent from that email address. We use IP2Location database [10] to map each sender IP address to a country. A participant might travel to other countries while sending emails using the same email address. Therefore, the sender's IP address may vary depending on the location of the sender. However, a participant is likely to send most of his/her emails from a single country. We locate the most frequent country from which a participant sends emails and assign the participant to that country. For the example shown in Figure 2, 4 out of 5 sender IPs are mapped to Canada and one is mapped to France. Our sender IP address analysis would indicate that the country of the participant is Canada. We measure the confidence of the identified country using a Confident Ratio (CR), defined in formula (1). CR ranges from 0 to 1. The higher the CR, the more confident we are about the identified country.

$$\text{Confidence Ratio} = \frac{D}{T} \qquad (1)$$

D is the number of messages sent from the most frequent country. T is the total number of sent messages.

TABLE I. BREAKDOWN OF EMAIL ADDRESSES BY TYPE OF TOP-LEVEL DOMAINS

	# of addrs w/ country code	# of addrs w/ resalable country code	# of addrs w/ generic domain	Total # of addrs
PostgreSQL	1,591(34%)	159(3%)	2,992(63%)	4,742
GTK+	955(35%)	97(4%)	1,682(61%)	2,734

Table I shows a breakdown of the usage of top-level domains in the participants' email addresses in our studied projects. We observe that a large number (roughly 63%) of participants use generic top-level

domain names. Such domain names cannot be resolved by prior techniques (e.g., [6], [8]). Furthermore, 3-4% of participants use resalable country codes. These participants are highly likely to be incorrectly mapped by prior techniques.

D. Conflict resolution

For each email address, both the top-level domain analysis and sender IP address analysis might produce different countries. For the example shown in Figure 2, the top-level domain analysis indicates that the Federated States of Micronesia is the participant's country, while the sender IP address analysis shows that the participant is from Canada. We define the following heuristics to resolve the discrepancies:

- If none of the analyses produces a result, we mark the participant's country as unidentifiable.
- If only one of the analyses produce a result, we use the result as the country for the participant. For example, when a participant uses generic domain such as .com in their email address, only the sender IP address analysis can produce a result.

The following cases may happen when both analyses produce results.

- If both analyses produce identical results, we use that result.
- For participants with a resalable country code domain in their email addresses, we use the result of the sender IP address analysis. The IP address is a better indicator of a participant's country than the domain name of the email address, since the resalable country code domain name can be used by participants throughout the globe.
- When the Confidence Ratio is lower than 0.7 using the sender IP address analysis, we apply the country determined by the top-level domain analysis. The threshold of 0.7 represents the majority (i.e., 70%) of the messages sent from one country in our studied projects.

IV. Case Study

To demonstrate the effectiveness of our technique, we compare the performance of our technique with the prior techniques. In this section, we introduce two studied projects and discuss the results produced by our technique.

A. Studied mailing list

We conducted a case study using the mailing list repositories of two long-lived open source projects. We use the developer mailing lists for the PostgreSQL [15] (postgresql-hackers) and GTK+ [9] (gtk-devel-list) projects in our case study. The PostgreSQL is a relational database management system. GTK+ is a toolkit for creating cross platform graphical user interfaces. Both projects involve a large pool of international developers who interact through the mailing lists. We chose the projects from two different domains: database management and graphic user interface development. Our objective is to determine if our results hold across domains and projects. Table II presents descriptive statistics about both projects.

TABLE II. Statistics on studied mailing lists

	Studied Period	**# of Participants**	**# of Threads**
PostgreSQL	1999-2008	4,742	23,104
GTK+	1999-2008	2,734	7,481

TABLE III. Performance of our technique and prior technique [8]

		Prior technique	**Our technique**
PostgreSQL	Participants	37%	67%
	Messages	51%	87%
GTK+	Participants	38%	68%
	Messages	28%	78%

B. Result

We apply our technique and reapply prior technique [8] on the studied mailing list. Table III summarizes the performance of our technique relative to the prior technique in identifying the country of a participant or a message on the mailing list. For example, our technique can identify roughly 67% of the country of a participant in both mailing lists in contrast to prior techniques which can only identify 37%. This represents approximately 80% improvement over prior technique. It indicates that our technique is able to identify the country of as much as 87% of the messages on the mailing list.

V. Conclusion

Mailing lists reflect the social structure of open source projects. Timely and open discussions of participants from around the world and across diverse time zone and regions ensure the smooth evolution of projects. Communication on these lists shed light about the spread and flow of knowledge for a project. Through a case study on two large and long-lived open source projects: PostgreSQL and GTK+, we show that our technique can identify more participants' country than prior technique. This improvement shed light on investigating the impact of having participants from around the world communicating on the mailing list.

TABLE IV. COUNTRY COMPOSITION OF POSTGRESQL MAILING LIST PARTICIPANTS

Country	Participants (%)	Messages (%)
United States	1037(32.6%)	76723(57.8%)
Germany	228(7.2%)	7237(5.5%)
Canada	160(5.0%)	7602(5.7%)
United Kingdom	144(4.5%)	8584(6.5%)
Australia	108(3.4%)	4862(3.7%)
Russian Federation	98(3.1%)	2578(1.9%)
India	97(3.0%)	574(0.4%)
France	97(3.0%)	1621(1.2%)
Italy	92(2.9%)	424(0.3%)
Brazil	90(2.8%)	424(0.3%)
Japan	89(2.8%)	3979(3.0%)
Netherlands	66(2.1%)	722(0.5%)
China	54(1.7%)	210(0.2%)
Poland	51(1.6%)	326(0.2%)
Czech Republic	48(1.5%)	940(0.7%)
Austria	47(1.5%)	3247(2.5%)
Sweden	44(1.4%)	2974(2.2%)
Hungary	41(1.3%)	271(0.2%)
Spain	37(1.2%)	227(0.2%)
Denmark	28(0.9%)	209(0.2%)
New Zealand	28(0.9%)	1,024(0.8%)
Other	492(15.5%)	7,891(6.0%)

TABLE V. COUNTRY COMPOSITION OF GTK+ MAILING LIST PARTICIPANTS

Country	Participants (%)	Messages (%)
United States	517(27.8%)	4,623(19.9%)
Germany	189(10.2%)	6,670(28.7%)
France	124(6.7%)	1026(4.4%)
United Kingdom	120(6.5%)	3,111(13.4%)
Sweden	64(3.4%)	800(3.4%)
Australia	63(3.4%)	708(3.0%)
Canada	57(3.1%)	429(1.8%)
Italy	55(3.0%)	260(1.1%)
India	53(2.9%)	173(0.7%)
Netherlands	50(2.7%)	268(1.1%)
Spain	42(2.3%)	216(0.9%)
China	41(2.2%)	1469(6.3%)
Finland	32(1.7%)	864(3.7%)
Russian Federation	29(1.6%)	242(1.0%)
Brazil	27(1.5%)	147(0.6%)
Japan	25(1.3%)	110(0.5%)
Austria	23(1.2%)	65(0.3%)
Belgium	23(1.2%)	134(0.6%)
Czech Republic	22(1.2%)	113(0.5%)
Norway	21(1.1%)	99(0.4%)
Other	283(15.2%)	1,733(7.5%)

REFERENCES

[1] Bird, C., Gourley, A., Devanbu, P., Gertz, M., and Swaminathan, A. Mining email social networks in Postgres. International Workshop on Mining Software Repositories, Shanghai, China, May 22 - 23, 2006, pp. 185-186.

[2] Bird, C., Gourley, A., Devanbu, P., Swaminathan, A., Hsu, G. Open Borders? Immigration in Open Source Projects, Fourth International Workshop on Mining Software Repositories pp. 6, 2007.

[3] Bowman, I. T. and Holt, R. C. Software architecture recovery using Conway's law. Conference of the Centre For Advanced Studies on Collaborative Research, Toronto, Ontario, Canada, November 30 - December 03, 1998, pp. 6

[4] Cherry, S. and Robillard, P. N. Communication Problems in Global Software Development: Spotlight on a New Field of Investigation. International Workshop on Global Software Development, International Conference on Software Engineering, Edinburgh, Scotland, May 24, 2004, IEEE, pp. 48-52.

[5] David, P.A., Waterman, A., Arora, S. FLOSS-US. The free/libre/open source software survey for 2003. Technical Report, Stanford Institute for Economic and Policy Research, Stanford, CA, 2003.

[6] Dempsey, B.J., Weiss, D., Jones, P., Greenberg, J. Who is an open source software developer? Communications of the ACM Volume 45, Issue 2, 2002, pp. 67–72.

[7] Ghosh, R.A., Glott, R., Krieger, B., Robles, G. Survey of developers (free/libre and open source software: survey and study). Technical Report, International Institute of Infonomics, University of Maastricht, The Netherlands, 2002.

[8] Gonzalez-Barahona, J. M., Robles, G., Andradas-Izquierdo, R. and Ghosh, R. A., Geographic origin of libre software participants, Information Economics and Policy, Volume 20, Issue 4, December 2008, pp. 356-363.

[9] GTK+ developer mailing list, http://mail.gnome.org/mailman/listinfo/gtk-devel-list/, last accessed on April 10, 2009.

[10] IP2Location Database, http://www.ip2location.com/, last accessed on April 09, 2009.

[11] Lancashire, D. Code, culture and cash: the fading altruism of open source development. First Monday, Volume 6, No. 12, December 2001.

[12] Robles, G. and Gonzalez-Barahona, J. M. Geographic location of developers at SourceForge. International Workshop on Mining Software Repositories, Shanghai, China, May 22 - 23, 2006, ACM, pp. 144-150.

[13] Robles, G., Scheider, H., Tretkowski, I., Weber, N. Who is doing it? A research on libre software developers. Technical Report, Technische Universitt, Berlin, Berlin, Germany, 2001.

[14] Padmanabhan, V., Subramanian, L. An investigation of geographic mapping techniques for Internet hosts. In Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, San Diego, CA, August 27- 31, 2001, ACM, pp. 173-185.

[15] PostgreSQL developer mailing list, http://archives.postgresql.org/pgsql-hackers/, last accessed on April 10, 2009.

[16] SourceForge open source project repository, http://sourceforge.net, last accessed on April 10, 2009.