A Study on Characteristics of Open Source Software Development Projects in the Areas of Engineering and Games

Pervis Fly, James Sims, and Hyunju Kim
Department of Computer Science
Jackson State University
Jackson, MS 39217, USA

ABSTRACT

This paper investigates characteristics of Open Source Software (OSS) development projects based on the number of participants and their roles and activities in the areas of Games and Science/Engineering. We utilized Principle Component Analysis (PCA) and clustering over the data from SourceForge.net, and the features that we considered and analyzed for this study include number of users, number of developers, number of message postings, number of file releases, and project subcategories. For the categories of Games and Engineering, the three most active subcategories within each were identified, and the result indicates that most of the Games and Engineering projects have similar characteristics in terms of the selected features and are small with about five users. The result also shows that a group of Engineering projects seem highly specialized and require more subject knowledge in executing the projects than other Games projects.

Keywords: Open Source Software Projects, Games, Engineering, PCA, Clustering.

1. INTRODUCTION

Open Source Software (OSS) has made a significant impact on the software development industry. OSS is software that provides its source code and allows users to modify and redistribute the updated software. The original vision for OSS was to improve the reliability and quality of the software: by making the source code available worldwide, fixing defects is faster and easier.

According to Raymond [10], there are several benefits from adopting the open source approach, including better-quality software and software of higher reliability, and lower development and maintenance costs. Independent peer review of software enhances its quality and provides a rapid release schedule, which ensures that user requests are addressed quickly. In addition, according to International Data Corporation's *Worldwide OSS and Billing 2011-2016 Forecast*, the global OSS and billing market is expected to grow from $25.4 billion in 2011 to up to $29.7 billion in 2016 [5].

Due to these benefits and demands, there are a growing number of studies of OSS communities, software processes, development tools and techniques, and OSS licenses for better understanding OSS activities and products. Among these studies, those on OSS development communities are important to understand OSS participants' activities and OSS evolution. These also provide insight into OSS people and products so as to better utilize OSS in diverse ways.

This paper investigates characteristics of OSS development projects based on the number of participants and their roles and activities in the two different areas: Games and Science/Engineering (Engineering for short). Since OSS development heavily depends on people, by analyzing the people's behaviors in terms of their development activities, we expected to study characteristics of the OSS projects in these two different areas. In addition, it has been reported that the growth of OSS depends on the growth of the community, which again depends on the number of participants and their activities.

For the features and related data, we used SourceForge Research Data Archive (SRDA) [13] and employed a model-based clustering technique to classify the OSS projects. Based on the resulting clusters, we analyzed the projects with respect to the selected features of participants and their activities. Section 2 of this paper presents related work, and section 3 explains the research data and methods. Section 4 discusses the findings from the result, and section 5 concludes this paper.

2. RELATED WORK

Since the OSS development process is different from that of traditional software development, many researchers and case studies have focused on unique aspects of OSS process and participants. Research in [11,12] studied social processes, interrelationships, organizational contexts of OSS, and other related issues in OSS communities. They analyzed communications and messages within project email lists, discussion boards, bulletin boards, news postings, etc. These resources contain diverse information about OSS requirement analysis, implementation, testing, and software evolution.

Research in [1,2,8,14,15,16] studied characteristics of OSS participants and social structures/networks of OSS communities. Research in [8] studied four major OSS projects and identified eight different roles of the participants. They found that the structure of an OSS community heavily depends on the members and their roles. The study in [14] analyzed email activities of 120 projects to understand how decentralization within the OSS communities leads to hierarchies. It applied weighted-network analysis on the email activities and identified possible reinforcement effect between a few core developers, which affects the mechanisms behind the OSS community's social self-organization. Prior research by [7] studied the activities of participants in regards to their roles within OSS projects.

Findings from the previous research indicate that an OSS community's success and evolution are closely related to the participants' roles and population. Specifically, a study in [2] and a series of research conducted by the University of Notre Dame [1,15,16] focused on OSS individuals' activities, roles, and mechanisms/structures of the communities. Research in [2] analyzed mailing lists and bug trackers of SourceForge's 116 projects, where each project has more than seven developers and more than 100 bugs in the bug tracking system. The result showed that the size of the core groups is less than seven in each project.

SourceForge (http://www.sourceforge.net) provides free hosting to OSS projects and supports more than 300,000 registered OSS projects as of August 2011. Also, SourceForge provides information about the projects to OSS research groups via SRDA and FLOSSmole [3]. The research conducted by the University of Notre Dame extracted their experimental data from SRDA and studied various aspects of OSS participants and communities. Our study also utilized data from SourceForge.

As shown from the previous studies, analyzing activities of OSS participants is important to understand OSS communities as social organizations. In many cases, OSS communities assign a formal role to each participant within an OSS project according to his/her expertise and contribution to the project. In [7], these formal roles were studied with respect to participants' activities: the activities were clustered according to the activities' characteristics, and the resulting clusters were analyzed with respect to the participants' roles to see how these formal roles were distinguished from each other by activity patterns.

In this paper, the OSS projects themselves are studied, with clusters being formed based on the characteristics of each project. The resulting clusters are then analyzed to determine the overall differences for each project as well as whether there were any unique characteristics between Games and Engineering projects. While the previous work utilized OSS data dated from March 2007 to March 2008, our experimental data is dated up to May 2011, which would provide up-to-date findings about the OSS projects.

3. EXPERIMENTAL DATA AND DATA CLUSTERING

3.1 Data Collection

SourceForge provides information about its users and the projects every month to SRDA in order to support OSS research and scholars. SRDA does not include personal user information or SourceForge's specific functional information. The archive contains more than 100 tables, and also provides query forms for data extraction from the tables. By using the query forms, we extracted our experimental data, which dated up May 2011.

Each observation in our experimental data includes a project ID, the number of users for the given OSS project, the subset of users with the Developer role, the number of message postings, the number of file releases that the project has made, and the project subcategory.

The project ID uniquely identifies the project. The number of users for the given OSS project is important as certain projects may have a tendency to involve more users than others. The Developer role was also included as the developer is in charge of making file changes and releases. We considered the number of Developers as well as the number of file releases in categorizing the projects. The number of message postings can show how active the project has been. Any users, including developers or other roles can post messages. However, the message postings in SourceForge have a special constraint: each project has the option of whether or not to have a forum. Thus, in our experimental data, those projects with no forum were considered to have zero messages. Similarly, those projects with a forum, but did not produce any messages have zero for the number of message postings.

The project subcategory is based on the three most active subcategories within each of Games and Engineering. We identified these subcategories by counting the number of projects under each subcategory. A project ID may appear multiple times in our dataset if the OSS project was involved in multiple categories or subcategories. Thus, the observation also includes the number of subcategories in which the project was included.

The query process extracted 34,421 records that belonged to the Games category and 32,374 records that belonged to the Engineering category, for a total of 66,795 observations. To work on a sample of the data that would most likely mirror the characteristics of the categories as a whole, we decided to use "active" projects for further analysis. A project was considered "active" if it has three or more users: the higher that number, the more active the project. This reduced the total number of records from 66,795 to 12,678, which shows that over 80% of the initially sampled projects have fewer than three users.

As aforementioned, the three most active subcategories within the two categories were identified. The three most active subcategories in Games were Role-playing, Simulation, and Board Games, whereas the three most active subcategories in Engineering were Mathematics, Simulations, and Bioinformatics. Table 1 shows the number of projects belonging to each subcategory. The overall total number of projects that fit these criteria was 4,774.

3.2 Clustering Games and Engineering Projects

The 4,774 observations were first analyzed with a Principal Components Analysis (PCA) to evaluate their characteristics. PCA is known as a statistical technique that identifies patterns in data and presents the data in such a way to highlight the similarities and differences [6]. Thus, we conducted PCA on the observations first so that the characteristics of the data within the observations are manifested, which would result in better performance in the following clustering step.

For the PCA, the "*prcomp*" function of R language [9] was applied, and the data was analyzed by a singular value decomposition of the observations. We assigned the number of users, developers, messages, and file releases for each given observation and the number of subcategories in which the observation is listed (MultiS for short) as features for PCA. As a result, each observation consists of these five features (dimension is equal to 5). MultiS was considered as a feature because it implies interdisciplinary characteristics of the project.

After conducting the PCA process over the observations, we performed model-based clustering on the output data from the process [4]. Model-based clustering assumes that the data were generated by a model, and the clustering process presents the model with clusters of similar observations. In our experiment, "*mclust*" function of R was used in clustering, which produces clusters according to the Bayesian Information Criterion (BIC) for an Expectation-Maximization (EM) algorithm.

In order to identify any cluster that has observations from a single subcategory or category, the process of PCA followed by clustering was repeated as follows:

- Step 1: Conduct PCA followed by clustering over the set of observations, OBS_{new}.
- Step 2: Identify clusters that have observations from a single category or subcategory and set them to OBS_{same}.
- Step 3: $OBS_{new} = OBS_{new} - OBS_{same}$
- Step 4: Go to Step 1.

We assumed that the observations of each OBS_{same} cluster imply characteristics of the corresponding category or subcategory because those are similar to each other within the cluster as well as are distinct from the observations that do not belong to that specific cluster.

4. ANALYSIS OF CLUSTERS AND DISCUSSION

The first iteration of the data produced 32 clusters, and only one cluster had the observations from a single category: the cluster contained 12 observations, and these were all Engineering projects. Table 2 shows details of the observations. *mean* (*u*) indicates the mean value of users, and similarly the *mean*()s indicate the mean values of developers (*d*), messages (*m*), file releases (*r*), and MultiS (*MS*). Also, *%*(*d*) indicates the percentage of developers within the cluster.

Table 1. The top three subcategories in each of Games and Engineering

Criterion	Games	Engineering
Most active	Role-playing (1,331 projects: 50.98% of all Games projects)	Bioinformatics (780 projects: 36.06% of all Engineering projects)
2nd most active	Simulation (772 projects: 29.57%)	Simulations (741 projects: 34.26%)
3rd most active	Board Games (508 projects: 19.45%)	Mathematics (642 projects: 29.68%)
Total (% of all projects)	2,611 projects (54.69%)	2,163 projects (45.31%)

Table 2. Analysis of Engineering projects from the first iteration

Size (%)	mean (u)	mean (d)	% (d)	mean (m)	mean (r)	mean (m)/mean (r)	mean (MS)
12 (0.25)	4.8	3.3	68.4	16.5	72.0	0.2	3

Even though this cluster is small, the observations show several interesting characteristics of Engineering projects, as follows:

- All the projects are interdisciplinary: each project is related to all of the Engineering's three subcategories, which are Bioinformatics, Simulations, and Mathematics.
- The user groups have a relatively high percentage (68.4%) of developers. The overall averaged percentage of developers is 49.0%.
- On the other hand, the user groups are relatively small (4.8 users) compared to 6.4 users, which is the overall averaged mean value of users.
- The users produced many more file releases than message postings. Comparing to this group's 0.2 ratio of *mean* (*m*) to *mean* (*r*), the overall averaged mean ratio is 4.8. This indicates that the projects as a whole produced 0.2 message per file release.

Since this cluster was identified as a result from the first iteration, the observations within the cluster can be assumed to have characteristics that make them distinguished well from others. Thus, the finding indicates that there is a small group of Engineering projects that are interdisciplinary and have a relatively small number of users, but most of the users actively produce file releases rather than message postings. This implies that the user groups understand their projects well and thus, there might be little communication bottleneck.

Our second iteration produced 45 clusters, and one of them contained all Engineering projects as shown in Table 3.

Those Engineering projects have similar characteristics to the ones from the first iteration. The projects tend to be related to more than one subject, and the user groups are small, but they produce a lot more file releases than messages. Interestingly enough, the projects produced almost zero (0) message: this implies that the projects were well-understood among the users.

From the third iteration, a total of 24 clusters were identified, and the cluster in Table 4 contained observations only from the Games category. Compared to the Engineering projects from the previous iterations, the user groups of the Games projects are large. Also, the users produced a lot more messages than the Engineering's user groups. This implies that the Games projects had active communications among the users. The percentage of developers shows that the 2/3 users on average are developers.

Another finding from the third iteration is that about 75% of the all Games and Engineering projects were similar to each other in terms of the selected features. An additional 10% of the all projects were also found in a single cluster as shown in Table 5.

When considering the two clusters in Table 5 together, the size of user groups is about 5, and the percentage of developers is about 40. They produced around 15 to 16 messages and fewer than 10 file releases. Most of the projects belong to a single subcategory, which indicates that the projects are not interdisciplinary.

Table 3. Analysis of Engineering projects from the second iteration

Size (%)	mean (u)	mean (d)	% (d)	mean (m)	mean (r)	mean (m)/mean (r)	mean (MS)
8 (0.17)	3.5	1.8	50.0	0.2	139.0	0.01	2

Table 4. Analysis of Games projects from the third iteration

Size (%)	mean (u)	mean (d)	% (d)	mean (m)	mean (r)	mean (m)/mean (r)	mean (MS)
6 (0.13)	66.8	47.2	70.6	579.0	196.7	2.9	1.67

Table 5. The largest two clusters after the third iteration

	CL_A	CL_B
Size (%)	3538 (74.42)	480 (10.10)
mean (u)	4.5	4.7
mean (d)	1.9	1.8
% (d)	43.0	39.2
mean (m)	16.1	14.6
mean (r)	4.5	9.8
mean (m)/ mean (r)	3.6	1.5
mean (MS)	1	2
Games projects (%)	2013 (56.90)	248 (51.67)
Engineering projects (%)	1525 (43.10)	232 (48.33)

In summary, about 85% of the Engineering and Games OSS projects have similar patterns in terms of the user group, message posting, and file release activities. However, some OSS projects seem to have unique characteristics inherited from the subject area, either Engineering or Games. Those characteristics can be described as follows:

- Some Engineering projects have a relatively small size of users, but they are all actively participated in the project.
    - There was little communication among the users. This might be because the users had good understandings about the project as well as their tasks.
    - The projects were interdisciplinary, and thus they seem to be highly specialized and targeting to a specific user group.

- Some Games projects have large user groups, and seem to have active communications among the users.
    - They produced many more messages and file releases than most of the other projects. This might indicate that the games from these projects are popular.
    - They also had a much higher percentage of developers than the average. This indicates that most of the users are actually developers of the code, and it might be because much of the programming does not require specialized subject knowledge compared to the Engineering projects aforementioned.

5. CONCLUSION

In this paper, we studied characteristics of OSS projects in the areas of Engineering and Games. We utilized the up-to-date, real-world OSS project data provided by SourceForge and the iterative process of PCA and clustering in order to analyze the data. The result indicates that most of the Engineering and Games projects have similar patterns in terms of user group, message posting, and file release. The detailed data also show that most of the projects are small with about five users.

However, based on the findings, it is also implied that some Engineering projects seem highly specialized and require more subject knowledge in executing the projects than other Games projects. The results from this study are expected to help build a better understanding of the OSS projects in the two areas and better utilize the code products from them: we plan to utilize the findings in locating appropriate OSS that can be used as seeds for students' development projects.

Acknowledgement: This study was supported by CREU (Collaborative Research Experience for Undergraduates) of CRA-W (The Computer Research Associations' Committee on the Status of Women in Computing Research).

6. REFERENCES

[1] S. Christley and G. Madey, "Analysis of Activity in the Open Source Software Development Community", In **Proc. of the 40th Hawaii International Conference on System Sciences**, 2007, pp. 166b.

[2] K. Crowston, K. Wei, Q. Li, and J. Howison, "Core and Periphery in Free/Libre and Open Source Software Team Communications", In **Proc. of the 39th Hawaii International Conference on System Sciences**, 2006, pp. 118.1.

[3] FLOSSmole: Collaborative Collection and Analysis of Free/Libre/Open Source Project Data, http://ossmole.sourceforge.net/

[4] C. Fraley and A. Raftery, “Model-Based Clustering, Discriminant Analysis, and Density Estimation”, **Journal of the American Statistical Analysis**, vol. 97, 2002, pp. 611-631.

[5] IDC, **Worldwide OSS and Billing 2011-2016 Forecast**, http://www.idc.com.

[6] I. T. Jollife, **Principal Component Analysis**, Springer, 1986.

[7] H. Kim, A. Ezeala, and Y. Park. “Analysis of Participants’ Roles in Open Source Software Communities with Model-Based Clustering”, In **Proc. of SERP**, 2009, pp. 141-146.

[8] K. Nakakoji, Y. Yamamoto, Y. Nishinaka, K. Kishidaand, and Y. Ye, “Evolution Patterns of Open-Source Software Systems and Communities”, In **Proc. of International Conference on Software Engineering**, 2002, pp. 76-85.

[9] The R Project for Statistical Computing, http://www.r-project.org/

[10] E. Raymond, “The Cathedral and the Bazaar”, http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/, 2000.

[11] W. Scacchi, “Free/Open Source Software Development: Recent Research Results and Emerging Opportunities”, In **Proc. of ESEC/FSE**, 2007, pp. 459-468.

[12] W. Scacchi, “Understanding the Requirements for Developing Open Source Software Systems”, **IEE Proceedings- Software**, vol. 149, no. 1, 2002, pp. 24-39.

[13] SourceForge Research Data Archive, http://zerlot.cse.nd.edu/

[14] S. Valverde, G. Theraulaz, J. Gautrais, V. Fourcassie, and R. Sole, “Self-Organization Patterns in Wasp and Open Source Communities”, **IEEE Intelligent Systems**, vol. 26, no. 2, 2006, pp. 36-40.

[15] J. Xu, Y. Gao, J. Goett, and G. Madey, “A Multi-Model Docking Experiment of Dynamic Social Network Simulations”, In **Agents 2003**, http://www.nd.edu/~oss/Papers/Agents2003_Docking.pdf.

[16] J. Xu, Y. Gao, S. Christley, and G. Madey, “A Topological Analysis of the Open Source Software Development Community”, In **Proc. of the 38th Hawaii International Conference on System Sciences**, 2005, pp. 198.1.