Success Rules of OSS Projects using Datamining 3-Itemset Association Rule

Andi Wahju Rahardjo Emanuel[1], Retantyo Wardoyo[2], Jazi Eko Istiyanto[3], Khabib Mustofa[4]

**[1] Bachelor Informatics, Faculty of Information Technology, Maranatha Christian University
Bandung, West Java, Indonesia**

**[2, 3, 4] Department of Computer Science and Electronics, Gadjah Mada University
Yogyakarta, Indonesia**

Abstract

We present a research to find the success rules of 134,549 Open Source Software (OSS) Projects at Sourceforge portal using Datamining 3-Itemset Association Rule. Seventeen types of OSS Project's data are collected, classified, and then analyzed using Weka datamining tool. The Datamining 3-Itemset Association Rule is used to find the success rules of these projects by assuming that the success of these projects are reflected by the number of downloads. The result are formulated into 9 success rules that may be used as guidelines by future initiators of OSS Project and other developers to increase the possibility of success of their projects.

Keywords: *Open Source Software Project, Datamining Association Rule, Success Rule, sourceforge.net*

1. Introduction

Open Source Software (OSS) is one of the current trends in Information Technology, especially in the field of Software Engineering. Once thought only as the sharing playground for researchers, academics and programmer enthusiasts during their spare time, this “methodology” is evolved into one of the mainstream software development methodology challenging the already established software engineering disciplines. Some success stories about this OSS Projects such as Apache Web Server, Linux Operating System, Openoffice.org productivity suite, Mozilla Web Browser, and many more. Despite the apparent success stories relating to OSS projects, there are many more projects using this scheme which are failed. Some approaches or guidelines need to be discovered to assist an initiators and contributors of OSS Projects in increasing the chance of success for the project. We believe that these approaches / guidelines could be found by studying the existing small to medium sized OSS Projects to find their success rules. In our previous research by gathering OSS Project's information from Sourceforge portal and using Datamining 2-Itemset Association Rule already found 6 success factors [5]. This research is further exploration from this research in which we are using 3-Itemset Association Rule to find additional or more specific success rules.

This paper is organized as follows: Section 2 describes the current studies on OSS Project’s success factors, Section 3 describes the theoretical background of OSS Projects and Datamining Association Rule. The Datamining processes are described in Section 4 with the interpretation of the result into the OSS success factors is shown in Section 5. The conclusion is described in Section 6.

2. Current Studies on OSS Project Success

Many studies have been conducted to identify the key success factors of OSS Projects. One approach of the study is by studying the processes of many large and successful projects, such as the study on Debian GNU/Linux [12], FreeBSD [4], Apache Web Server [10], OpenBSD [7], Apache against Mozilla [9], Arla against Mozilla [2], and some 15 popular OSS Projects [8]. This approach may provide excellent examples about how large OSS project works; however, these large and successful OSS Projects already have established process and organization involving large many developers and other stakeholders that are difficult to be implemented by small and medium sized OSS Projects. The study of small and medium sized OSS Projects that considered successful are more relevant compared to the study on large and mature OSS Projects since all of these projects are usually start from small size.

In our previous research, by gathering OSS Project's information from Sourceforge portal by using Datamining 2-Itemset Association Rule, we have found 6 success factors [5]. Further elaboration used in this research is by using 3-Itemset Association Rule to find more detail or additional success rules that contribute to the success of OSS Projects. The subject of the research is still the small to medium sized OSS Projects hosted in one of the most popular web portal which is Sourceforge. At the time of this research (January, 2010), this portal had 160,141 registered projects; and in this research, 134,549 OSS Projects are selected and their data are extracted and analyzed using Datamining 3-Itemset Association Rule.

3. Theoretical Background

3.1 Open Source Software Projects

Open Source is a software development methodology based on several distinct characteristics:

- The source code of the application is freely available for everybody to download, improve and modify [11].
- People who contribute to the development of the Open Source projects is forming a group called Open Source Community which is voluntary [3].
- The primary concern of the developers in Open Source Software Projects are building features and fixing bugs [6].

In order to develop software application in OSS Project, a project initiator may use the service from OSS Development portal such as sourceforge.net, launchpad.net, Google code, etc. The sourceforge.net portal is chosen since it covers more than 70% of total OSS Projects from these popular portals [5].

3.2 Datamining Association Rule

Datamining is a technique to find hidden structure and relationship in a large number of population [1]. The knowledge about these structure and relationship is discovered by using two methods which are predictive (predicting unknown value or future value of a variable), and descriptive (finding human readable patterns). In this research, the descriptive method is selected in this research since it is intended to find the human readable patterns from all the data being collected from Sourceforge.net. In this method, there are several techniques / rules that may be chosen which are Classification, Segmentation / Clustering, Association, etc. The Association Rule is selected since it is able show the dependency between one parameter to another parameter of some large collections of data.

The Association Rule will find dependency rule which will predict the appearance of an itemset (Consequent) based on the appearance of other itemset / itemsets (Antecedent). The 3-Itemset Association Rule has two Antecedents connected with logical AND, and a single Consequent. The 3 Itemset may be stated as:

$$\{X, Y\} \Rightarrow \{Z\} \qquad \text{.............................(1)}$$

Where X is the first Antecedent (Antecedent1) and Y is the second Antecedent (Antecedent2), and Z is the Consequent. X and Y must appear at the same time as the cause and Z is the result with some certainty values called Support and Confidence. The value of Support ({(X,Y),Z}) shows the number of transaction containing item X AND Y and item Z against total population, whereas the value of Confidence((X,Y)=>Z) shows the probability of the occurrence of item Z if a transaction containing item (X,Y) and Z. In this research, the rules are of in the interest if it has the minimum Support of 10% and the minimum Confidence of 50%. The value of 10% for the Support is selected since it will represent significant proportion of the entire population, and the value of 50% in Confidence is selected since it also represent half or more in terms of probability of occurrence.

4. Datamining Processes

Table 1 shows the data description of the OSS project recorded from sourceforge.net portal. There are 17 types of data parameters being recorded for each of the OSS Projects.

Table 1: OSS Project Data Description

No	*Parameter*	*Type*	*Remark*
1	Name	Text	The name of the project
2	Audience	Text	Intended audience
3	Database	Text	Database environment used
4	Description	Text	Project description
5	Developer	Text	User name of developer
6	Development Status	Text	Status of project development
7	Download	Integer	Number of download
8	Filename	Text	Name of downloadable file from project's front page
9	File Size	Text	Size of downloadable filename
10	License	Text	Applicable license
11	Operating	Text	Applicable operating system

No	Parameter	Type	Remark
	System		
12	Programming Language	Text	Programming language used
13	Review	Text	Review from user
14	Thumb	Integer	Recorded thumb up and thumb down from user
15	Topic	Text	Applicable topic for the project
16	Translation	Text	Available language translation
17	User Interface	Text	Applicable user interface

The parameters such as audience, database, developer, development status, license, operating system, programming language, review, topic, translation, and user interface are having zero to many parameters. The count of these parameters is also considered as the parameters used during datamining process. The total parameters being recorded are more than 27 parameters if it includes the count of these parameters.

4.1 Data Collection Process

The data collection process of OSS Projects from Sourceforge was conducted by creating custom-made PHP script crawler. The collecting process was conducted in three phases:

- Recording summary of projects (from link http://sourceforge.net/softwaremap) to record most of the parameters shown in table 1.
- Recording more detail information by crawling each individual project link page to record the developer, project description, filename, file size, number of thumbs (up and down), and reviews.
- Filling the missing information, finding and deleting duplicates and then generating count data from multiple-value parameters.

The collection process was taking about 9 weeks to complete starting early January 2010 through the end of February 2010. Out of 160,141 OSS Projects registered from the portal, the crawler was able to collect data from 134,549 unique projects stored in 27 tables with total 3,115,085 records.

4.2 Data Classification Process

The next process was the classification of the data. Most of the parameters need to be classified in some categories with enough number of population in order to gain meaningful rules.

Audience: There are 121,095 OSS Projects (90%) of the recorded projects that list the audiences of its software project. There more than 23 distinct values of project's audience which are then classified into three classes which is ‘Specific Audience’ (42.54%), ‘Developers’ (29.10%), and ‘Common Users’ (28.36%).

Database Environment: There are only 30,335 OSS Projects (22.55%) of recorded projects are using at least one database. The OSS Project with database are classified as ‘MySQL’ (31.25%), ‘SQL-based’ (27.17%), ‘API-based’ (20.48%), ‘Text-based’ (15.42%), and ‘Other’ (5.68%).

Project Description: Each OSS Project has a description to state the purpose of project. The project description is mostly short sentence / paragraph with the peak at about 36 words. The project description is classified into three categories which are ‘short’ (< 26 words – 43.78%), ‘middle’ (26 - 36 words – 32.58%) and ‘long’ (> 36 words – 23.64%).

Development Status: There are 128,215 OSS Projects (95.29%) of recorded projects that list the development status of their projects. The classification is based on the development status of the project which are ‘1 – Planning’ (18.84%), ‘2 – Pre-Alpha’ (15.15%), ‘3 – Alpha’ (17.15%), ‘4 – Beta’ (24.05%), ‘5 – Production / Stable’ (20.56%), ‘6 – Mature’ (1.83%), and ‘7 – Inactive’ (2.42%).

Number of Download: Table 2 shows the statistics of the number of downloads of OSS projects. The number of download is 0 may means that there are no download or the project does not have any downloadable file .

Table 2. Statistics about Number of Download

Download	Population	Percentage
0 / NA	55,986	41.61%
1 - 99	10,080	7.49%
100 - 999	33,539	24.93%
1000 - 9999	24,438	18.16%
10,000 - 99,999	8302	6.17%
100,000 - 999,999	1831	1.36%
1,000,000 - 9,999,999	322	0.24%
≥ 10,000,000	51	0.04%

Note: NA - not available (downloadable file is not yet available).

The number of download of OSS projects in categorized as ‘none’ (41.61%), ‘hundred or less’ (32.42%), and 'thousands or more' (25.97%). In this research, the number of download is assumed as the indication of success in OSS Project. If an OSS Project is successful, it will be accepted by many users that is indicated by the large number of downloads for the project. Therefore, the number of download is classified into three categories which are 'none' (0 / NA download), 'hundreds or less' (1 up to 999 downloads), and 'thousands or more' (more than 1000 downloads). The Association Rule that has 'Download – Thousands or more' as Consequent with any possible combinations of two other Antecedents are the interested rules.

Filename: This experiment only record the filename and its size listed on the projects' site on the first page (http://sourceforge.net/project/project_name). This filename is not necessary the only available filename, and there is also no guarantee that the filename is always the source code of the project. The filename is then classified based on its extension which are 'zip' (47.27%), 'tar.gz' (29.72%), 'jar' (7.42%), 'tar.bz2' (5.82%), 'tgz' (5.02%), 'rar' (2.46%), and 'other format' (2.28%).

File Size: The size of the downloadable filename was also recorded and then categorized based on its order of magnitude (BYTES, KB, MB, or GB). The classifications are 'BYTES' (0.26%), 'KB' (71.90%), 'MB' (27.79%), and 'GB' (0.06%).

License: There are 131,777 OSS Projects (97.94%) of recorded projects that list the applicable license for the project. There are 75 distinct values for the license for OSS projects, and they are classified into 'GPL' (61.57%), 'LGPL' (10.64%), 'BSD License' (6.75%), 'Apache License' (3.78%), 'Public License' (3.40%), 'MIT License' (2.62%), 'AFL' (2.62%), 'Mozilla License' (1.40%), and 'Other' (8.28%).

Operating System: There are 111,760 OSS Projects (83.06%) of recorded projects that list the applicable Operating System. There are 85 distinct Operating System for the OSS projects which are then classified as 'Linux-like' (35.63%), 'Windows' (34.25%), 'Cross-Platform' (23.19%), or 'Other' (6.93%).

Programming Language: There are 127,247 OSS Projects (94.57%) of recorded projects that list the applicable programming language. There are 97 distinct programming languages for the OSS projects which are then classified into 'Java' (20.10%), 'C++' (16.27%), 'Other OOP' (7.45%), 'C' (14.91%), 'PHP' (13.14%), 'Other Script-based' (18.57%), or 'Other' (9.56%).

Thumb (Up and Down): There are only 16,829 OSS Projects (12.50%) of recorded projects that being thumb-reviewed (users give either thumb up or thumb down). The classifications of thumb are 'single' (48.58%), 'two or three' (24.34%), 'four to ten' (15.86%), and 'eleven or more' (11.21%).

Topic: There are 440 distinct topics for the OSS Projects which are then classified into 6 categories which are 'Software Development' (19.97%), 'Internet/Networking' (17.41%), 'Data Management' (17.37%), 'Games/Entertainment' (14.49%), 'Scientific/Engineering' (11.65%), 'Other topic' (19.10%).

Translation: There are 77,269 OSS Projects (57.43%) of recorded projects that list the available language translation. There are 67 distinct values for available language translation that is then classified into three classes which are 'English' (59.18%), 'European' (33.85%), and 'Other' (6.98%).

User Interface: There are 97,302 OSS Projects (72.32%) of recorded projects that list the available user interface for the project. There are 60 distinct values which is then classified into 4 classes which are 'Desktop-based' (46.91%), 'Web-based' (25.57%), 'Text-based' (17.13%) and 'Other' (10.39%).

Parameter's Count: The count of parameters are also recorded and classified. The classification is categorized into only three classes which are 'one', 'two', and 'three or more'. The parameters that are classified in this scheme are audience count, database count, developer count, development status count, license count, operating system count, programming language count, review, and user interface count.

4.3 Result of Datamining 3-Itemset Association Rule

The process of Datamining Association Rule was conducted with 3-Itemset. There are 277 possible combinations of 3-Itemset that have been processed using Weka resulting in 111 interesting rules that surpass the minimum values of Support and Confidence. The result of Datamining Association rule that have 'Download - Thousands or more' as Consequent with Confidence more than 50% and Support more than 10% are the interested rule. Due to the limited table space, the value of Antecedent1 and Consequent ('Download – Thousands or more') are not stated in the tables.

Audience: Table 3 shows the result with 'Audience – Common Users' as Antecedent1.

Table 3 Result for Antecedent1: Audience-Common Users

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
License	GPL	132554	10.86%	51.84%
User Interface	Desktop-based	129654	11.01%	58.37%

Audience Count: Table 4 shows the result with ‘Audience Count – One’ as Antecedent1.

Table 4. Result for Antecedent1: Audience Count – One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	three or more	7241	14.71%	93.75%
Review Count	one	7241	19.64%	69.33%
Total Thumb	single	13982	16.01%	62.50%
User Interface	Desktop-based	70439	14.33%	51.10%

Database: Table 5 shows the result with ‘Database – SQL-based’ as Antecedent 1.

Table 5. Result for Antecedent1: Database – SQL-based

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	one	2630	11.94%	74.06%
Review Count	three or more	2630	10.87%	94.70%

Database Count: Table 6 shows the result with ‘Database Count – one’ as Antecedent1.

Table 6. Result for Antecedent1: Database Count – One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status	5 – Production / Stable	13858	12.56%	55.57%
Review Count	one	1593	23.85%	63.65%
Review Count	three or more	1593	18.77%	90.61%
Total Thumb	single	2941	18.02%	54.53%
Translation	European	15985	20.51%	63.29%

Developer Count: Table 7 shows the result with ‘Developer Count – one’ as Antecedent1.

Table 7. Result for Antecedent1: Developer Count – One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	three or more	7804	16.57%	92.29%
Review Count	one	7804	25.03%	66.50%
Total Thumb	single	15088	21.10%	60.41%
User Interface	Desktop-based	75306	19.20%	50.32%

Development Status: Table 8 shows the result with ‘Development Status – 5 - Production/Stable’ as Antecedent1.

Table 8. Result for Antecedent1: Development Status – 5 – Production/Stable

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Database Count	one	13858	12.56%	55.57%
License Count	one	81997	16.34%	59.68%
Programming Language Count	one	80463	12.95%	57.45%
Review Count	one	8407	17.06%	80.29%
Review Count	three or more	8407	16.87%	97.26%
Translation	European	82545	12.02%	80.31%
Translation	English	82545	11.94%	67.24%

Development Status Count: Table 9 shows the result with ‘Development Status Count – one’ as Antecedent1.

Table 9. Result for Antecedent1: Development Status Count – One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Operating System	Linux-like	127057	17.45%	51.54%
Operating System	Windows	127057	15.99%	54.38%
Review Count	one	7517	32.98%	70.81%
Total Thumb	single	14521	26.54%	63.91%
Translation	English	72218	27.31%	51.00%
User Interface	Desktop-based	72999	23.49%	53.52%

Filename: Table 10 shows the result with ‘Filename – zip’ as Antecent1.

Table 10. Result for Antecedent1: Filename – Zip

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	three or more	7234	15.05%	94.12%
Review Count	one	7234	19.75%	73.28%

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Total Thumb	single	13970	15.16%	65.33%
Translation	English	66468	15.80%	55.87%
Translation Count	one	44286	18.59%	51.84%
User Interface	Desktop-based	67284	14.08%	56.79%

File Size: There are two groups of result which are either ‘Size – KB’ or ‘Size – MB’ as Antecedent1. Table 11 shows the result with ‘Size – KB’ as Antecedent1.

Table 11 Result for Antecedent1: Size – KB

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Operating System	Linux-like	116876	14.51%	50.12%
Operating System	Windows	116876	11.59%	54.04%
Review Count	one	7234	21.69%	100.00%
Total Thumb	single	13970	19.02%	64.15%
User Interface	Desktop-based	67284	16.25%	51.92%

Table 12 shows the result with ‘Size – MB’ as Antecedent1

Table 12. Result for Antecedent1: Size – MB

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	one	7234	15.61%	77.06%
Review Count	three or more	7234	17.11%	96.04%

License: Table 13 shows the result with ‘License – GPL’ as Antecedent1.

Table 13. Result for Antecedent1: License – GPL

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Audience	Common Users	132554	10.86%	51.84%
Review Count	three or more	8516	18.91%	94.26%
Review Count	one	8516	22.19%	70.76%
Total Thumb	single	16302	17.27%	63.35%

License Count: Table 14 shows the result with ‘License Count – One’ as Antecedent1.

Table 14. Result for Antecedent1: License Count - One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status	5 – Production / Stable	81997	16.34%	59.68%
Operating System	Windows	127733	16.97%	54.74%
Operating System	Linux-like	127733	18.52%	51.81%
Review Count	three or more	7698	26.42%	94.17%
Review Count	one	7698	34.18%	71.51%
Total Thumb	two to three	14858	17.46%	75.87%
Total Thumb	single	14858	27.57%	64.26%
Translation	English	72265	28.95%	51.31%
User Interface	Desktop-based	73709	24.96%	53.93%

Operating System: There are two groups of result which are either ‘Operating System – Linux-like’ or ‘Operating System – windows’ as Antecedent1. Table 15 shows the result with ‘Operating System – Linux-like’ as Antecedent1.

Table 15. Result for Antecedent1: Operating System – Linux-like

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status Count	one	127057	17.45%	51.54%
License Count	one	127733	18.52%	51.81%
Size	KB	116876	14.51%	50.12%
Programming Language Count	one	125992	13.89%	50.31%
Total Thumb	single	27811	10.60%	74.75%
Translation Count	one	93190	15.52%	52.77%

Table 16 shows the result with ‘Operating System – Windows’ as Antecedent1.

Table 16. Result for Antecedent1: Operating System – Windows

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status Count	one	127057	15.99%	54.38%
License Count	one	127733	16.97%	54.74%
Size	KB	116876	11.59%	54.04%
Programming Language Count	one	125992	12.78%	52.57%
Review Count	three or more	15185	15.06%	96.01%
Total Thumb	single	27811	10.88%	70.49%
User Interface Count	one	101187	13.65%	54.22%

Operating System Count: Table 17 shows the result with ‘Operating System Count – one’ as Antecedent1.

Table 17. Result for Antecedent1: Operating System Count: One

Antecedent2		Analysis		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Review Count	one	6966	19.57%	69.36%
Review Count	three or more	6966	14.10%	93.26%
Total Thumb	single	13404	16.70%	63.34%
Total Thumb	two to three	13404	10.65%	75.38%

Programming Language Count: Table 18 shows the result with ‘Programming Language Count – One’ as Antecedent1.

Table 18. Result for Antecedent1: Programming Language Count - One

Antecedent2		Analysis		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status	5 – Production / Stable	80463	12.95%	57.45%
Operating System	Windows	125992	12.78%	52.57%
Operating System	Linux-like	125992	13.89%	50.31%
Review Count	one	7508	27.33%	69.42%
Review Count	three or more	7508	19.90%	93.43%
Total Thumb	single	14478	22.56%	63.27%
Translation	European	71510	14.68%	61.22%
User Interface	Desktop-based	72879	19.75%	51.67%

Review Count: There are two groups of result which are either ‘Review Count – one’ or ‘Review Count – three or more’ as Antecedent1. Table 19 shows the result with ‘Review Count – One’ as Antecedent1.

Table 19. Result for Antecedent1 : Review Count – One

Antecedent2		Analysis		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Audience Count	one	7241	19.64%	69.33%
Database	SQL-based	2630	11.94%	74.06%
Database Count	one	1593	23.85%	63.65%
Developer Count	one	7804	25.03%	66.50%
Development Status	5 – Production / Stable	8407	17.06%	80.29%
Development Status Count	one	7517	32.98%	70.81%
Filename	zip	7234	19.75%	73.28%
License	GPL	8516	22.19%	70.76%
License Count	one	7698	34.18%	71.51%
Operating System Count	one	6966	19.57%	69.36%
Programming Language Count	one	7508	27.33%	69.42%
Size	MB	7234	15.61%	77.06%
Size	KB	7234	21.69%	100.00%
Translation	European	13745	11.45%	82.49%
Translation	English	13745	12.73%	79.44%
Translation Count	one	4873	26.45%	75.47%
User Interface	Desktop-based	8947	19.84%	79.42%
User Interface Count	one	6186	26.06%	70.52%

Table 20 shows the result with ‘Review Count – three or more’ as Antecedent1.

Table 20. Result for Antecedent1: Review Count – three or more

Antecedent2		Analysis		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Audience Count	one	7241	14.71%	93.75%
Database	SQL-based	2630	10.87%	94.70%
Database Count	one	1593	18.77%	90.61%
Developer Count	one	7804	16.57%	92.29%
Development Status	5 – Production / Stable	8407	16.87%	97.26%
Filename	zip	7234	15.05%	94.12%
License	GPL	8516	18.91%	94.26%
License Count	one	7698	26.42%	94.17%
Operating System	Windows	15185	15.06%	96.01%
Operating System Count	one	6966	14.10%	93.26%
Programming Language Count	one	7508	19.90%	93.43%
Size	MB	7234	17.11%	96.04%
Translation	European	13745	32.99%	98.67%
Translation	English	13745	10.47%	95.61%
Translation Count	one	4873	17.26%	94.49%
User Interface	Desktop-based	8947	20.35%	96.30%
User Interface Count	one	6186	20.43%	93.70%

Thumb: There are four groups of result which are 'Thumb – single', 'Thumb – two or more', 'Thumb – four to ten', or 'Thumb – eleven or more' as Antecedent1. Table 21 shows the result with 'Thumb – single' as Antecedent1.

Table 21. Result for Antecedent1: Thumb - Single

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Topic Count	one	15117	14.28%	59.96%
Audience Count	one	13982	16.01%	62.50%
Database Count	one	2941	18.02%	54.53%
Developer Count	one	15088	21.10%	60.41%
Development Status Count	one	14521	26.54%	63.91%
Filename	zip	13970	15.16%	65.33%
License	GPL	16302	17.27%	63.35%
License Count	one	14858	27.57%	64.26%
Operating System	Linux-like	27811	10.60%	74.75%
Operating System	Windows	27811	10.88%	70.49%
Operating System Count	one	13404	16.70%	63.34%
Programming Language Count	one	14478	22.56%	63.27%
Size	KB	13970	19.02%	64.15%
Translation Count	one	9330	23.00%	69.88%
User Interface	Desktop-based	16463	15.37%	71.59%
User Interface Count	one	11770	21.89%	63.71%

Table 22 shows the result with 'Thumb – two to three' as Antecedent1.

Table 22. Result for Antecedent1: Thumb – Two or Three

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
License Count	one	14858	17.46%	75.87%
Operating System Count	one	13404	10.65%	75.38%
Translation Count	one	9330	13.27%	80.86%
User Interface	Desktop-based	16463	10.37%	80.99%

Table 23 shows the result with 'Thumb – four to ten' as Antecedent1.

Table 23. Result for Antecedent1: Thumb – Four to ten

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Translation Count	one	9330	10.32%	91.71%

Table 24 shows the result with 'Thumb – eleven or more' as Antecedent1.

Table 24. Result for Antecedent1: Thumb – Eleven or More

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Translation	European	21170	20.35%	99.75%
User Interface	Desktop-based	16463	10.14%	99.52%

Translation: There are two groups of result which are either 'Translation – English' or 'Translation – European' as Antecedent1. Table 25 shows the result with 'Translation – English' as Antecedent1.

Table 25. Result for Antecedent1: Translation - English

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Development Status	5 – Production / Stable	82545	11.94%	67.24%
Development Status Count	one	72218	27.31%	51.00%
Filename	zip	66468	15.80%	55.87%
License Count	one	72265	28.95%	51.31%
Review Count	three or more	13745	10.47%	95.61%
Review Count	one	13745	12.73%	79.44%
Translation Count	one	73412	22.24%	50.19%
User Interface Count	one	60703	21.69%	50.66%

Table 26 shows the result with 'Translation – European' as Antecedent1.

Table 26 Result for Antecedent1: Translation - European

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Database Count	one	15985	20.51%	63.29%
Development Status	5 – Production / Stable	82545	12.02%	80.31%
Programming Language Count	one	71510	14.68%	61.22%
Review Count	three or more	13745	32.99%	98.67%
Review Count	one	13745	11.45%	82.49%
Total Thumb	eleven or more	21170	20.35%	99.75%

Translation Count: Table 27 shows the result with 'Translation Count – one' as Antecedent1.

Table 27. Result for Antecedent1: Translation Count - One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Filename	zip	44286	18.59%	51.84%
Operating System	Linux-like	93190	15.52%	52.77%
Review Count	one	4873	26.45%	75.47%
Review Count	three or more	4873	17.26%	94.49%
Total Thumb	single	9330	23.00%	69.88%
Total Thumb	two to three	9330	13.27%	80.86%
Total Thumb	four to ten	9330	10.32%	91.71%
Translation	English	73412	22.24%	50.19%
User Interface	Desktop-based	51586	20.58%	54.78%

User Interface: Table 28 shows the result with 'User Interface – Desktop-based' as Antecedent1.

Table 28. Result for Antecedent1: User Interface – Desktop-based

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Audience	Common Users	129654	11.01%	58.37%
Audience Count	one	70439	14.33%	51.10%
Developer Count	one	75306	19.20%	50.32%
Development Status Count	one	72999	23.49%	53.52%
Filename	zip	67284	14.08%	56.79%
License Count	one	73709	24.96%	53.93%
Programming Language Count	one	72879	19.75%	51.67%
Review Count	three or more	8947	20.35%	96.30%
Review Count	one	8947	19.84%	79.42%
Size	KB	67284	16.25%	51.92%
Total Thumb	single	16463	15.37%	71.59%
Total Thumb	two to three	16463	10.37%	80.99%
Total Thumb	eleven or more	16463	10.14%	99.52%
Translation Count	one	51586	20.58%	54.78%

User Interface Count: Table 29 shows the result with 'User Interface Count – One' as Antecedent1.

Table 29. Result for Antecedent1: User Interface Count – One

Antecedent2		**Analysis**		
Parameter	*Class*	*Pop.*	*Sup.*	*Conf.*
Operating System	Windows	101187	13.65%	54.22%
Review Count	one	6186	26.06%	70.52%
Review Count	three or more	6186	20.43%	93.70%
Total Thumb	single	11770	21.89%	63.71%
Translation	English	60703	21.69%	50.66%

5. Interpreting the Result

Combining the interpretation from Datamining 3-Itemset Association Rules are the success factors that should be followed by the project initiators and other developers to increase the probability of success of their OSS Projects. The interpretation is done qualitatively by noticing the frequency of appearance of Antecedent1 and Antecedent2 of table 3 through table 29. These success rules are:

1. Project should target for common users as audience.
2. Project source code should already in 5 – Production / Stable development status.
3. Project should work on either Linux-like or Windows operating system.
4. Project should be reviewed and thumb-reviewed by at least one users. Project with windows operating system should have at least three reviews.
5. Project should have Desktop-based User Interface.
6. Project should select a single type of license, preferable GPL license.
7. Project has filename in zip format with size in either KB or MB in magnitude. For project with file size MB, it needs three or more reviews.
8. If the project is using database environment, select SQL-based database, and it should be reviewed by at least one user.
9. Project should have either English or European language translation.

Rule number 1, 2, 3 and 9 are similar to the previous findings using 2-Itemset Association Rule [5], rule number 4 and 7 are more specific compared to the previous findings, and rule number 5, 6, and 8 are new rules. It is also observed that some freedom is still available for project initiator to decide such as the topic, programming language and description of his/her project without affecting the number of download.

Some caution should be considered regarding to these rules. The subject being researched is small to medium OSS Projects from Sourceforge that may not reflect the whole population of OSS Projects that are small, medium and large scale using OSS development portals or hosting in their own website. These result should also verified using OSS Project data from other portal such as launchpad.net, Google code, etc. to verify their validity.

6. Conclusions

We present the Datamining 3-Itemset Association Rule of 134,549 OSS Projects crawled from Sourceforge portal. This covers about 84% of the total of 160,141 OSS Projects registered at the portal in the month of January 2010. There are more than 27 parameter being recorded which are project's name, audience, audience count, database environment, database environment count, developer count, development status, development status count, number of download, filename and file size, license, license count, operating system, operating system count, programming language, programming language count, review count, topic, topic count, translation, translation count, user interface, and user interface count.

The result of this datamining process are 9 success rules that may be applied by initiators and contributors of OSS Project in order to increase the probability of success of their projects. The details of the guidelines is shown in Section 4. Future work of this research include expanding the experiment to cover other portal such as launchpad.net, Google code and Freshmeat. Other possible exploration is by using more advanced learning rule other than the association rule.

Acknowledgments

The authors of this paper would like to thank Maranatha Christian University (http://www.maranatha.edu) that provide the funding for this research and Department of Computer Science and Electronics at Gadjah Mada University (http://mkom.ugm.ac.id) which provide guideline and technical assistance for the research.

References

[1]. R.Agrawal, R. Srikant, "Fast Algorithm for Mining Association Rule", Proceeding of 20th International Conference Very Large Database, 1994, pp 1 - 32.

[2]. A. Capiluppi, J. F. Ramil, "Studying the Evolution of Open Source Systems at Different Levels of Granularity: Two Case Studies", Proceeding of the 7th International Workshop of Principles of Software Evolution, 2004, 113 - 118.

[3]. S. Christley, G. Madey, "Analysis of Activity in the Open Source Software Development Community", Proceeding of the 40th IEEE Annual Hawaii International Conference on System Sciences, 2007, 166b.

[4]. T.T. Dinh-Trong, J.M. Bieman, "The FreeBSD Project: A Replication Case Study of Open Source Development", IEEE Transaction on Software Engineering Vol. 31 No. 6, June 2005, 481 – 494.

[5]. A.W.R. Emanuel, R.Wardoyo, J.E. Istiyanto, K. Mustofa, Success Factors of OSS Projects from Sourceforge using Datamining Association Rule, Proceeding of the 2nd International Conference on Distributed Frameworks and Applications (DFmA), 2010, 141 - 148

[6]. V.K. Gurbani, A. Garvert, J.D. Herbsleb, "A Case Study of Open Source Tools and Practices in Commercial Setting", Proceeding of the fifth Workshop on Open Source Software Engineering, 2006, 1 - 6.

[7]. P.L. Li, J. Herbsleb, M. Shaw, " Finding Predictors of Field Defects for Open Source Software Systems in Commonly Available Data Sources: a Case Study of OpenBSD", Proceeding of 11th IEEE International Software Metrics Symposium, 2005, 32.

[8]. G. von Krogh, S. Spaeth, S. Haefliger, "Knowledge Reuse in Open Source Software: An Exploratory Study of 15 Open Source Projects", Proceeding of 38th Hawaii International Conference on System Sciences, 2005, 198b

[9]. A. Mockus, R.T. Fielding, J. Herbsleb, "Two Case Studies of Open Source Software Development: Apache and Mozilla", ACM Transaction on Software Engineering and Methodology Vol. II No. 3, Juli 2002, 309 – 346

[10]. A. Mockus, R.T. Fielding, J. Herbsleb, "A Case Study of Open Source Software Development: The Apache Server", ACM ICSE, 2000, 263 – 272.

[11]. E.S. Raymond, "The Cathedral and the Bazaar". Knowledge, Technology & Policy, vol. 12 no. 3 pp. 23 – 49, 1999.

[12]. S. Spaeth, M. Stuermer, "Sampling in Open Source Development: The Case for Using the Debian GNU/Linux Distribution", Proceedings of the 40th IEEE Hawaii International Conference on System Sciences, 2007, 166a.