An Analysis of Relationship between Storage Usage Distribution and Per-User Quota Value

Kasahara, Yoshiaki
Research Institute for Information Technology, Kyushu University

Kawatani, Takuya
Kyosan Electric Manufacturing Co, Ltd.

Ito, Eisuke
Research Institute for Information Technology, Kyushu University

Shimozono, Koichi
Kagoshima University

出版情報：Proceedings of the 2016 ACM on SIGUCCS Annual Conference, SIGUCCS '16, pp.153-158, 2016-11-09. Association for Computing Machinery: ACM
バージョン：
権利関係：

An Analysis of Relationship between Storage Usage Distribution and Per-User Quota Value

Yoshiaki Kasahara
Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka 812-8581, Japan
+81 92 642 2297
kasahara.yoshiaki.820@m.kyushu-u.ac.jp

Takuya Kawatani
Kyosan Electric Manufacturing Co, Ltd.
2-29-1 Heiancho, Tsurumi-ku
Yokohama 230-0031, Japan
t-kawatani@kyudai.jp

Eisuke Ito
Kyushu University
6-10-1 Hakozaki, Higashi-ku
Fukuoka 812-8581, Japan
+81 92 642 4037
ito.eisuke.523@m.kyushu-u.ac.jp

Koichi Shimozono
Kagoshima University
1-21-24 Korimoto
Kagoshima 890-8580, Japan
+81 99 285 7477
simozono@cc.kagoshima-u.ac.jp

ABSTRACT
To prevent resource (especially storage) shortage, information systems such as storage services and email services usually impose an upper bound of resource consumption (quota) per user. In a conservative way, an administrator tends to set a quota value such as the storage capacity divided by the expected maximum number of users for safety and fairness, but it tends to leave large unused storage space, because the users' storage usage pattern shows a long-tailed distribution. In this paper, we analyzed storage usage distribution of some email services to approximate the distribution using a power-law distribution, and proposed a method to calculate an optimal quota value from a target size of storage consumption to increase storage utilization. We applied an optimal quota value we calculated to a real email service and analyzed the effect of quota change. Then, we analyzed actual distributions further to find a better model to approximate the distribution, and found that a log-normal distribution explained the distribution better than power-law. We also analyzed two other universities' email service to find similar distribution in these systems.

Keywords
Email System; Storage Quota Management; Long-Tail Distribution

1. INTRODUCTION
It is always important to distribute available resources efficiently and effectively. Computer systems are no exception. Computational power and storage become cheaper every year, but still it is important to utilize resources efficiently. In this paper, we focus on storage capacity and utilization of multiuser system. From users' perspectives, it is convenient if they can use as much capacity as possible, but the actual storage capacity of a system is usually limited. To prevent shortage of storage space in a service system, the administrator of the service system usually sets user quota as an upper limit of usable space for each user.

Especially for an on-premise system, the total available storage capacity is decided at the initial system installation, and usually it cannot be expanded on-demand. To avoid service failure caused by resource exhaustion, the administrator tends to set a conservative quota value such as the total storage capacity divided by the expected maximum number of users. But it is not common that all the user equally uses up to their quota value. In many information systems, users' storage utilization patterns exhibit a long-tailed distribution. Only a fraction of users (heavy users) use up their quota, and most of remaining users do not use the system much. In consequence of such a long-tailed usage pattern, applying a conservative quota value causes low utilization of the overall storage resource. On the other hand, it is hard to increase the quota value properly without a guideline.

In this research, we analyzed users' storage consumption of our email services in order to estimate the optimal quota value, which improves the utilization of the storage capacity. Specifically, we analyzed the storage usage history of a university-wide email system (called "Primary Mail Service") in Kyushu University [1][2]. In our previous work [3], we also analyzed our file sharing service (a kind of file storage service), but this time we focus on email systems. After that, we discussed how to estimate the optimal quota value setting which improve the utilization efficiency of storage resource based on per-user storage consumption distribution. By modeling the distribution, we could estimate the optimal quota value from the number of users and the whole storage capacity.

The rest of this paper is organized as follows. In section 2, we introduce our email services in Kyushu University, which are the target of our analysis. In section 3 we analyze the distribution of per-user storage usage in these systems, and how we estimate storage usage using a power-law distribution. In section 4, we analyze storage usage distribution of the current email service further. In section 5, we present two other examples from other

universities. In section 6, we describe some related studies. Finally, we present our conclusion and future works in section 7.

2. TARGET SYSTEMS

In this research, we analyzed the university-wide email system at Kyushu University. In this section, first we briefly introduce the number of staff members and students in Kyushu University, which is the number of users for these systems. Next, we describe the details of our (previous and current) mail services.

2.1 Number of users in Kyushu University

Table 1 shows the approximate number of IDs issued by the university-wide authentication service [4] in Kyushu University as of January 2015. The number also represents the number of users of the mail system.

Table 1. The number of IDs in Kyushu Univ. (Jan. 2015)

Role	Total No. of IDs (approx.)
Curricular students	19,000
Non-curricular students	500
Faculty and staff members	9,000
Non-employee workers	1,000
Total	30,500

2.2 Kyushu University Primary Mail Service

Similar to other universities, Kyushu University provides a campus-wide email service as a basis of communication infrastructure. In this paper, we analyzed two services. One is the previous email service operated from July 2009 and March 2014. The other is the current email service operated from March 2014 until now. The details of both services and migration were presented in the previous SIGUCCS [2].

2.2.1 The previous email service

The previous email service was constructed based on Mirapoint appliance servers. The users of this system were staff members (including faculty) only. There was another email service for students, but due to lack of usage history, we could not analyze the storage usage of students in the previous service.

Table 2 shows some numbers related to the resource and limitation of the system from July 2009 to March 2014. The column of "Date" denotes when the values of the row had become in effect. The system provided SMTP, POP, and webmail for free. From January 2011, a premium service class for paid users was started. The service included 10GB quota without message expiration, and IMAP support. In April 2013, the quota value was expanded to 20GB due to requests from paid users.

Furthermore, the quota value was actually a "soft" limit, which meant that a user could store messages even after the using storage size exceeded the quota value. The system sent a warning message to the mailbox of such a user, but didn't block further incoming message. It was because there was a retention period for messages of non-paid users, and expired messages were automatically removed from the system.

Table 2. Resources and limitations (previous service)

Date	Jul. 2009~	Dec. 2009~	Feb. 2011~
Total storage	1,200 GB	1,200 GB	2,600 GB
User quota	100 MB	100 MB	300 MB
Expiration	30 days	60 days	60 days
Max. message size	20 MB/message		

2.2.2 The current email service

The current email service was mainly built using open source software such as Postfix, Dovecot, Squirrelmail, and Roundcube on dozens of CentOS Linux servers. This service supports both staff members and students. Internally, staff members and students are handled separately (by using different VMs and separate NAS storage partitions), but both systems share the physical infrastructure.

Table 3 shows some numbers related to the resource and limitation of the current email service. The column of "Date" denotes when the values of the row had become in effect. The different limitation between staff members and students is the maximum size of a message only. The quota value is "hard" limit in this service, so users whose mailbox reached the quota limit cannot receive further email messages until they remove some messages. Similar to the previous email service, there is a premium service class for paid users which expands the quota value to 40GB.

Table 3. Resources and limitations (current service)

Date	Apr. 2014~	Mar. 2015~	Nov. 2015~
Total storage	10 TB (staff members) + 10 TB (students)		
User quota	1 GB	4 GB	8 GB
Expiration	Never		
Max. message size	20 MB(staff members) / 10 MB (students)		

3. STORAGE USAGE ANALYSIS AND ESTIMATION

The previous mail service had recorded storage usage per user from May 2009 to March 2014 (until the end of the service) every day. By using the record, we first analyzed the situation of storage usage in this system. Then, we approximate the distribution using a power-law distribution (for simplicity) and developed a relationship between a user quota value and estimated storage usage under the specified quota value. By using the relationship, we can determine more optimal quota value for a target size of storage.

3.1 Actual storage usage distribution of the previous email service

We sorted each user's amount of storage usage in descending order and plotted a log-log scale graph with the vertical axis of usage amount and the horizontal axis of the user rank. For example, Figure 1 shows the distribution on January 27th, 2014.

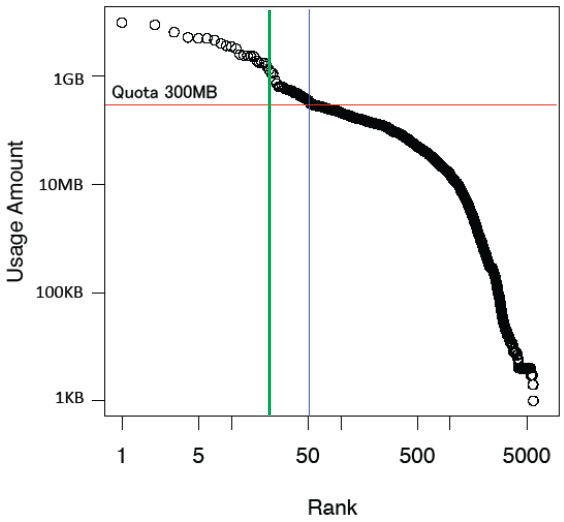

Figure 1. Previous mail service: rank-usage (log-log) (Jan. 27th, 2014).

The red horizontal line in Figure 1 indicates the quota value (300MB) of non-paid (ordinary) users. The blue vertical line indicates the rank (54th) where the user's storage usage exceeded the quota value. The rank 1st to 54th consisted of two user groups. All the users from 1st to 24th (the green vertical line) were paid users whose quota value was 10GB, and actually used more than 300MB. Users from 25th to 54th were non-paid (ordinary) users except one paid user, but they exceeded their quota value of 300MB. As described before, it was because the quota value was a "soft" limit. The system sent a warning message to the mailbox of such a user, but didn't block further incoming message or delete old messages automatically. To simplify the discussion, we will exclude users using more than 300MB and inactive users who didn't use storage space.

3.2 Storage usage estimation

In this section, we discuss the relationship between storage usage and quota value. The actual storage consumption among users wasn't homogeneous as shown in section 3.1. It was like a long-tail distribution such that only a fraction of users used up to their quota limit and most of users only used little space. Based on the distribution, we tried to estimate a quota value to increase the storage utilization.

3.2.1 Symbols for modeling

We use the following symbols for modeling.

S:	Target size of storage usage
u_i:	Storage usage amount for user i ($i = 1 \ldots n$)
U:	Actual total usage ($U = \sum u_i$)
q:	Quota value for users

S denotes the target storage space decided by an administrator, and it is not the total amount of storage in the system. For example, there is storage of capacity 1TB in a mail system, and the administrator thinks users may use up to 75% of the total capacity, then S is set to 750GB.

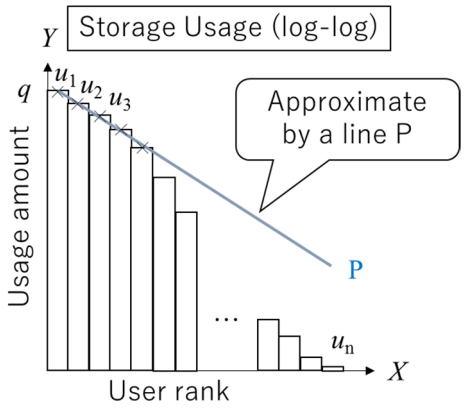

Figure 2. Concept of storage usage estimation

3.2.2 Approximation by Power-Law Distribution

First, we approximated the distribution as a power-law distribution. Figure 1 doesn't seem like a linear graph as a whole, but the partial graph with higher ranked users (~500) seems almost linear, and these are dominant users to use most of storage space. In addition, we could simplify calculation using a power-law distribution, because we can represent the distribution by two values (intercept and slope).

Figure 2 shows a concept of storage usage estimation. It shows the relationship between the real distribution and approximation in a log-log plot. The vertical axis is the usage amount of each user and the horizontal axis is the rank of the user by the usage amount. The line P denotes approximated usage values by a power-law distribution. We suppose the intercept of the line P as q (quota value), and calculate the slope (a scaling exponent a) by using the real distribution observed.

A power-law distribution is represented by (1).

$$y = f(x) = qx^a, \quad (1)$$

where a is a scaling exponent. To take logarithm of both side of (1), it is transformed as follows:

$$\log y = \log qx^a = a \log x + \log q. \quad (2)$$

Let $Y = \log y$, $X = \log x$, $Q = \log q$, then we obtain

$$Y = aX + Q. \quad (3)$$

Equation (3) is a linear function with gradient a and intercept Q, so we could simplify the estimation.

The sum of all bars equals to U, and the area under the line P is an estimated usage amount U'. The difference between each bar and the line P is the error of estimation. We could make sure that the line P was always above all the bar, then the error became an extra capacity margin. Also please note that the graph is in log-log scale, so the error is relatively small because most of error part is in the lower right part of the graph.

By using an estimation by a power-law distribution (1), the total estimated usage amount U' is as follows:

$$U' = \sum_{x=1}^{n} f(x) = q \sum_{x=1}^{n} x^a \quad (4)$$

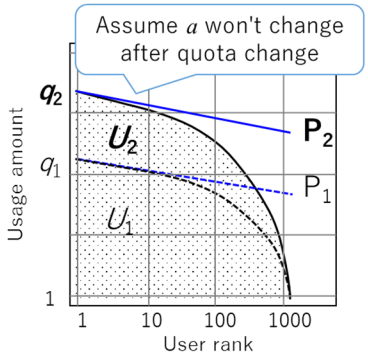

Figure 3. Distribution change after increasing the quota

3.2.3 Estimation of the optimal quota value

Consider changing the quota value from q_1 to q_2 $(q_1 < q_2)$ in order to increase the storage utilization. Let U_2 as the storage usage when the quota value was changed to q_2. We need to select q_2 where U_2 won't exceed the target capacity S while increasing storage utilization.

Let us discuss how the usage distribution would change after changing the quota value using Figure 3. After changing the quota value from q_1 to q_2, users' usage distribution would form new distribution. P_1 denotes an approximated distribution under the quota value q_1. If new distribution retains a similar gradient under the quota value q_2, the distribution will be similar to P_2. There is another possibility that it will become a distribution with a steeper gradient, but a steeper gradient means less storage consumption. To change the quota value in order to increase storage utilization, it is better to use an estimation which over-estimates users' consumption. So we assume the gradient won't change after quota change.

By using (4), we can estimate new quota value q_2 with the target size of storage usage S as follows:

$$q_2 = \frac{S}{\sum_{x=1}^{n} x^a} \qquad (5)$$

3.2.4 Example using the previous email service

In this section, we use an actual distribution from our previous email service to calculate the scaling exponent a. We used the usage distribution as of January 27th, 2014. To simplify, we excluded paid users, users whose usage was exceeded their quota value, and inactive users who stored nothing in the system. The number of remaining ordinary users was 4,139.

We used R [5] for nonlinear regression analyses to estimate the gradient of the power-law distribution. We used only the top 400 users for the estimation, because adding lower rank users will make the gradient unnecessarily steeper. At the result, we got the gradient (scaling exponent) as a = -0.404485.

On January 27th, 2014, the actual storage usage of ordinary users we took into consideration was 70GB. We calculated the value of U′ by (4) with the gradient -0.40 and we got U′ = 71.91GB.

If we want to set the target size of storage usage S as 2,000GB, we could estimate the optimal quota value as about 8.1GB by (5).

3.3 Applying to an actual service

When we had the result in section 3.2.4, our mail system had already been replaced to the current email service. So we decided to change the quota value of the current service in March 2015 (1GB at that time) to 4GB at first. The current service had 10TB storage, so even 8GB quota (~2TB target) seemed fine, but we thought we should be cautious because it is harder to reduce the expanded quota value later. Before the quota change, the entire storage consumption (sum of all the users) was increased 700MB/day in the email service for staff members. After the change, it was increased to 2GB/day. It was because top 100 heavy users had already hit 1GB quota, and they started to accumulate more messages after the quota change. Such increase rate was not a big issue, so we concluded that it should be safe to expand the quota again, and we changed the quota value to 8GB in Nov. 2015.

3.3.1 Comparison of distributions before and after quota change

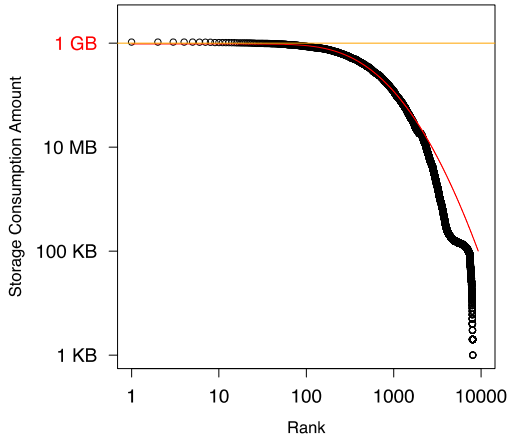

Figure 4. Distribution before quota change (Mar. 2015)

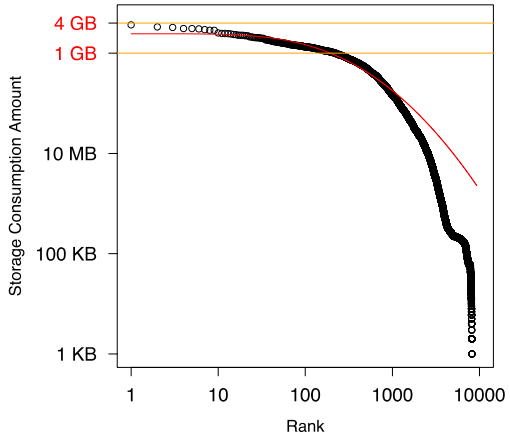

Figure 5. Distribution after quota change (Aug. 2015)

Figure 4 and Figure 5 shows distributions before and after the quota change (please disregard a curved red line for now). As mentioned before, in Figure 4 the top 100 users had already hit the quota value (1GB). That means an approximation by a power-law distribution might not be appropriate because the estimated scaling exponent becomes too shallow, which causes too much over-estimation. On the other hand, in Figure 5 most users don't hit their quota yet, and power-law estimation seems reasonable. We are still collecting data after changing the quota (twice) for our future work.

4. BETTER THAN POWER-LAW

In this section, we analyze storage usage distribution of the current email service further. In previous sections, we used a power-law distribution to approximate the actual storage usage distribution, but it is too simplified. In addition to a power-law, we selected other long-tail distributions including log-normal and exponential distributions, and performed nonlinear regression analyses to find the best fit distribution among them. Again we used R [5] for nonlinear regression analyses, and also calculated the value of AIC (Akaike's Information Criterion) to evaluate the fitness. Among candidates, the model with the smallest AIC value is the optimal model.

Table 4. AIC values of the email service for staff members

Date	Log-normal	Exponential	Power-law
2014-12	190567	181296	240136
2015-03	189127	175907	241305
2015-04	220443	213468	262697
2015-07	223660	225381	247474
2015-10	225128	228030	253537

Table 5. AIC values of the email service for students

Date	Log-normal	Exponential	Power-law
2014-12	403469	442402	485569
2015-03	399326	438545	480637
2015-04	473582	521261	571332
2015-07	405774	451361	479651
2015-10	429753	473030	499305

Table 4 and Table 5 shows AIC values calculated for each distribution with parameters obtained by nonlinear regression analyses. The double line between Mar. 2015 and Apr. 2015 denotes the quota change from 1GB to 4GB, and colored cells are considered the optimal. For the email service for staff members, both exponential and log-normal were close and better than power-law, and for students, log-normal was always the best. By these results, we considered that approximation by a log-normal distribution was generally the best. Possibly the change between log-normal and exponential was caused by hitting quota (because most students were not affected by the quota value), but we don't have a good explanation yet. Red curves in Figure 4 and Figure 5 show estimated log-normal distributions.

5. SYSTEMS IN OTHER UNIVERSITIES

In section 4, we showed that a log-normal distribution fit with the actual storage usage distribution of the email system in Kyushu University. In this section, we analyzed other universities' data to see if this feature is not local to our university. Two universities (Kagoshima Univ. and Shizuoka Univ.) provided their data.

5.1 Kagoshima University

Kagoshima University provides mail service to staff members using "DEEPMail". The total amount of storage capacity is 2TB, and the user quota is 1GB. A quota alert message will be sent to a user when the user's storage usage exceeds 800MB. We received

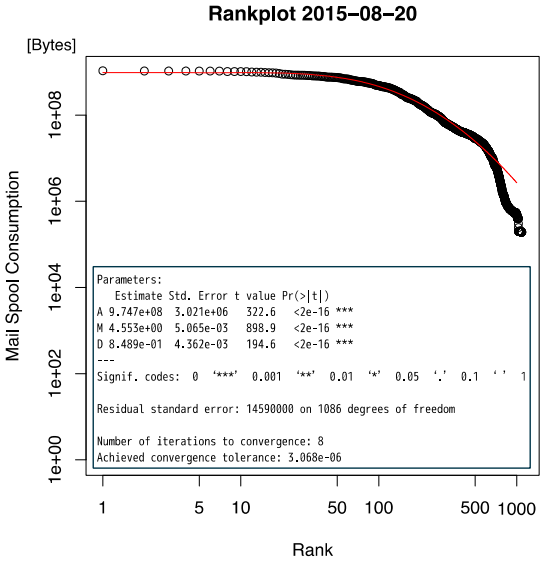

Figure 6. Kagoshima Univ.: rank-usage (log-log) (Aug. 20th, 2015).

storage usage data from Mar. 2015 to Aug. 2015. Similar to section 4, we estimated the distribution of storage usage by a nonlinear regression analysis. Figure 6 is an example of actual distribution graph. The red line in the graph denotes the estimated log-normal distribution. We concluded that log-normal distribution estimated the actual distribution the best.

5.2 Shizuoka University

We also received one-day sample data of email storage usage from Shizuoka University. The data included 1,643 students of Faculty of Informatics, Shizuoka University. The quota value was 200MB. We estimated the distribution of storage usage by a nonlinear regression analysis. Figure 7 is the actual distribution graph. The red line in the graph denotes the estimated log-normal distribution. We concluded that log-normal distribution estimated the actual distribution the best, too.

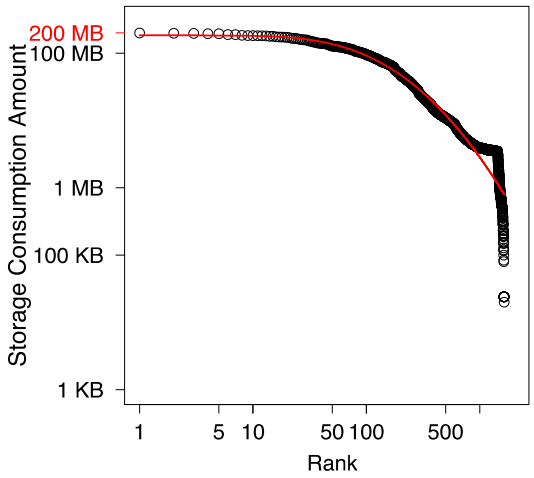

Figure 7. Shizuoka Univ.: rank-usage (log-log)

Of course, we cannot conclude only with these result that a log-normal distribution is a universal feature of email storage usage distribution, but at least we can say it is not local to our university, and it might be a good candidate as a hypothesis.

6. RELATED WORK

In article [6], Mitzenmacher mentioned that file size distributions were best modelled by a power-law distribution or a log-normal distribution. He surveyed about log-normal distribution and power-law distribution and reported that these distributions appeared frequently in various phenomena studied in economics and natural science. Also he showed that log-normal distributions had arisen as a possible alternative to power law. In our work, we also confirmed that user's storage usage distribution matched well with log-normal distribution in mail system and file sharing system of Kyushu University. As Mitzenmacher described in [6], we also didn't know an explicit model why the distribution was similar to log-normal distribution. It is our future work to establish a user behavioral model to explain how the distribution is formed.

In article [7], Kuninaka, et al. mentioned that a log-normal distribution appeared in various complex systems. They showed that many phenomena considered to be a normal distribution were actually fit better to a log-normal distribution. For example, people's height distribution was considered to be a normal distribution, but children's height distribution in a growth period fit better to a log-normal distribution. Also they showed that log-normal was more appropriate than power law or Zipf's law for representing population of cities [8]. User's storage usage could be considered to be in a growth period because it grows in time course, and might form a log-normal distribution. We want to study that more closely later.

In article [9], McKnight discussed future planning of storage preparation in information systems. He showed that "(total volume size) = (number of users) * (user quota) * 1.4" is a reasonable starting point for the initial purchase. It was based on his professional experience as a data storage administrator, and the reasoning was not discussed.

7. CONCLUSION

In this research, we analyzed the actual storage usage of email system and proposed a method to estimate overall storage usage from the user quota value by using power-law approximation. We also showed that a log-normal distribution is more suitable to approximate the storage usage distribution of various email system.

As a future work, we want to realize the relationship between three parameters of log-normal distribution (A, μ, and σ) and the quota value. Without the relationship, we cannot use a log-normal approximation to determine the optimal quota value. Actually we should be able to obtain a "bell curve" of a normal distribution using an appropriate log-scale histogram of the actual storage usage distribution if it is a log-normal distribution, and we can observe the change of parameters through the history of usage.

Another problem is the reason why a storage usage distribution exhibits a log-normal distribution. Especially for email system, a user's social network might influence the number and size of messages. Possibly such a network theory will explain the distribution.

8. ACKNOWLEDGMENTS

Our thanks to all the users using our mail services, and staff members of the Primary Mail Service working group to develop and maintain these systems in Information Infrastructure Initiative of Kyushu University. Also we appreciate Kagoshima University and Shizuoka University which kindly provided their usage data of email services.

9. REFERENCES

[1] Fujimura, N., Togawa, T., Kasahara, Y., and Ito, E. 2012. Introduction and experience with the Primary Mail Service based on their names for students. In *Proceedings of the SIGUCCS 2012* (Memphis, TN, October 17 - 19, 2012). ACM, New York, NY, 11-14. DOI= http://dx.doi.org/10.1145/2382456.2382460.

[2] Kasahara, Y., Ito, E., and Fujimura, N. 2014. Introduction of New Kyushu University Primary Mail Service for Staff Members and Students. In *Proceedings of the SIGUCCS 2014* (Salt Lake City, UT, November 2 - 7, 2014). ACM, New York, NY, 103-106. DOI= http://dx.doi.org/10.1145/2661172.2662965.

[3] Kasahara, Y., Kawatani, T., Ito, E., Simozono, K., and Fujimura, N. 2015. Optimization of Storage Quota Based on User's Usage Distribution, In *Proceedings of the 2015 IEEE 39th Annual Computer Software and Applications Conference Workshops (COMPSACW 2015)*, 149-154. DOI=http://dx.doi.org/10.1109/COMPSAC.2015.221

[4] Ito, E., Kasahara, Y., and Fujimura, N. 2013. Implementation and operation of the Kyushu university authentication system. In *Proceedings of the SIGUCCS 2013*(Chicago, IL, November 3 - 8, 2013). ACM, New York, NY, 137-142. DOI=http://dx.doi.org/10.1145/2504776.2504788.

[5] The R project for statistical computing, http://www.r-project.org/.

[6] Mitzenmacher, M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1, 2, 226-251. DOI=http://dx.doi.org/10.1080/15427951.2004.10129088

[7] Kobayashi, N., Kuninaka, H., Wakita, J., and Matsushita, M. 2011. Statistical features of complex systems–toward establishing sociological physics–. *Journal of the Physical Society of Japan*, 80, 7, 072001. DOI=http://dx.doi.org/10.1143/JPSJ.80.072001.

[8] Kuninaka, H. and Matsushita, M. 2008. Why does Zipf's law break down in rank-size distribution of cities?. *Journal of the Physical Society of Japan*, 77, 11, 114801. DOI=http://dx.doi.org/10.1143/JPSJ.77.114801.

[9] McKnight, C. J. 2006. Cost analysis and long term planning over the lifecycle of an enterprise storage solution. *Journal of Technology Management & Innovation*, 1, 5, 87-95.