

適切なクォータ値設定のためのディスク使用量分析

川谷卓哉[†] 伊東栄典[‡] 笠原義晃[‡] 藤村直美^{*}

ストレージの容量不足を防ぐため、システム管理者は各利用者の使用上限値（クォータ値）を設定する。システム管理者はサービスの安定提供のために、少なめのクォータ値を設定しがちである。また、クォータ値に単純に容量を利用者数で均等割りした値を設定することも多い。本研究では、筆者らが所属する九州大学の電子メールとファイル共有（蓄積）システムを対象に、長期間のディスク使用量を分析した。ほとんどの時点で、利用者のディスク使用量はロングテール型の分布をしており、その分布は対数正規分布に近いことが分かった。長期間のディスク使用量の推移を調べた所、全体の使用量が微増していたことも分かった。ストレージの活用度を定義し、それに基づく評価方法を考案した。さらに、ディスク使用量分布の冪分布近似により、活用度が高くなるクォータ値の算定方法を考案した。実際のデータに基づき、ストレージの活用度を最高にするクォータ値も算出した。

A study of disk usage distribution analysis to set optimal quota value

Takuya Kawatani[†] Eisuke Ito[‡] Yoshiaki Kasahara[‡]
Naomi Fujimura^{*}

To prevent shortage of storage space in a service system, an administrator usually set per-user quota as an upper limit of usable space for each user. To prevent service failure caused by resource exhaustion, an administrator tends to set a conservative quota value such as the storage capacity divided by the expected maximum number of users. In this research, authors analyzed long-term storage usage history of our email system and file sharing (archiving) system in Kyushu University. Mostly through the analyzed period, the usage pattern showed a long-tailed distribution similar to lognormal distribution. Also the overall storage consumption slowly increased during the analyzed period. Based on these analysis, the authors defined "storage utilization ratio" to evaluate how the storage is effectively used. By approximating a disk utilization pattern as a power-law distribution, we proposed a method to calculate an optimal quota value to maximize the utilization ratio. We applied the method to the actual data to calculate the best quota value to maximize the utilization ratio.

1. はじめに

組織の情報システム担当者が組織内用の情報サービスを構築する場合、必要な資源量（CPU、ネットワーク帯域、記憶装置の容量）の適切な見積りが必要である。利用者が多い場合、またその情報システムを用いる期間が長い場合、最適な必要資源量見積りは困難になる。例えば、自組織内にメールシステムを構築する場合、メールで用いるディスク容量について数年後は推測できても、10年後のディスク容量見積りは困難である。あるサービスに必要なディスク資源量の長期的傾向を予想できれば、システム維持費用の減少を期待できる。

通常システムのストレージ容量は有限である。容量不足はサービス停止の原因となることから、システム管理者は各利用者に使用量の上限值（クォータ値）を設定し、容量不足を未然に回避することが多い。この時、システム管理者は全利用者が上限値に近い値まで使用することを恐れ、クォータ値として小さめの値を設定する傾向がある。しかしながら、本稿で詳細を述べるように、全利用者のディス

ク使用量分布はロングテール型の分布になる。極少数の利用者だけが上限値一杯までを使い、少数の利用者が上限値の60%程度まで使う。一方、残るほとんどの利用者は上限値の20%以下の量しか使わない。このようなロングテール型の使用量分布の場合、単純な人数割でのクォータ値設定では結果的にストレージ全体の使用割合が低くなり、ディスク資源の活用度が下がる。

本研究では、ストレージの活用度を向上するクォータ値を設定するために、利用者のディスク使用量を分析する。分析対象として、筆者らが所属する九州大学で運用している電子メールとファイル共有（蓄積）システムの状況を分析する。分析にはシステムが定期的に出力していた利用状況ログを用いる。

次に、ディスク使用量分布に基づき、ディスク資源の活用効率を上げるクォータ値設定を考える。分布の傾向が分かれば、利用者数と全ストレージ容量から、利用者に課すクォータ値を設定できる。

本稿の構成を述べる。2節では、本研究が対象とする九州大学の全学基本メールと、ファイル共有システムについて説明する。3節では利用者数とディスク使用量の推移分析を述べる。4節では、3節の分析に基づく適切なクォータ値設定について述べる。最後に5節でまとめと今後の課題を述べる。

[†] 九州大学大学院 システム情報科学府
Faculty of Information Science and Electrical Engineering, Kyushu University.
[‡] 九州大学 情報基盤研究開発センター
Research Institute for Information Technology, Kyushu University.
^{*} 九州大学 芸術工学研究院
Faculty of Design, Kyushu University.

2. 対象システム

本研究で分析対象にした情報システムは、九州大学の電子メールシステムと、ファイル共有システムである。まずシステムの利用者数となる九州大学の構成員数を示す。その後、本稿で分析対象としたメールシステムとファイル共有システムを説明する。

2.1 学内構成員数

大学の主たる構成員は学生と教職員である。学生は、正課生と非正課生からなる。正課生は学部学生および大学院生である。非正課生は研究生や科目等履修生などの在籍者である。学生と教職員以外にも、特別研究員や、外部組織で雇用されている派遣社員、訪問研究者など、様々な身分の職員が在職している。九州大学における2014年12月現在の発行ID概数を表1に示す。表1に示すIDの数が、メールサーバの利用者数である。

表1 九州大学のID数(2014年1月現在)

種類	ID総数(概数)
正課生	19,000
非正課生	500
職員	9,000
派遣等	800
合計	29,300

2.2 全学基本メール

「全学基本メールサービス」は、大学の活動に公的に従事する全構成員に対して、基本的なメール環境を提供することを目的とするものである。本研究で利用状況を分析した全学基本メールシステムは職員向けで、2009年4月に導入したMirapoint社の製品で構成されていた。導入当初はメールゲートウェイにRazorGate600を1台、メールサーバにM6000を2台使用した。M6000のディスクサイズは1台600GBで、2台で1,200GBのディスク容量があった。これらの機器を用いて2009年7月よりサービスを提供した。

利用者アカウント管理と利用者認証はLDAPサーバとの連携で実現した。サービス開始当初は全学共通認証システムが提供するLDAPサーバを直接参照していた。その後、負荷分散や有料サービスクラスの導入のために、別途専用のLDAPサーバを導入した。

2010年12月にゲートウェイとしてRazorGate700を1台、2011年2月にメールサーバを2台追加した。追加した機器は同じくM6000だがマイナーアップデートされており、1台のディスク容量は700GBであった。システム全体のディスク容量は、600GBx2+700GBx2で、2,600GBになった。図1に2011年2月から2014年2月までの全学基本メールシステムの機器構成を示す。

表2に、2009年7月から2014年2月までの運用期間に

おける全学基本メールの資源量の制限値などを示す。なお、表2の時期は、その制限値に設定した時期を表している。

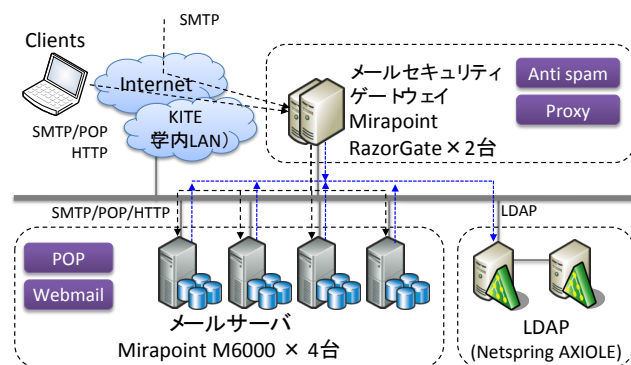


図1 九州大学全学基本メールの構成
 (2011年2月~2014年2月)

表2 全学基本メール性能諸元

時期(年.月)	2009.07	2009.12	2011.2
全ディスク容量	1,200GB	1,200GB	2,600GB
利用者の容量	100MB	300MB	300MB
保存期間	30日	60日	60日
受信メール上限	20MB/1通		
サービス	POP, Web mail		

2011年1月から、保存容量が10GBで保存期間を制限しないサービスクラスを有償で提供した。このサービスでは上記に加えてIMAPを利用できるようにした。有償サービス利用者から容量拡大の要望が出たため、2013年4月に保存容量を20GBに拡大した。

なお、全学基本メールにおける、利用者のメールスプール容量の制限はソフトリミットで設定していた。(有償でない)一般利用者については保存期間を過ぎた古いメールは自動的に削除される設定のため、メールの容量が上限値に近づくと警告メッセージが出るものの、新たなメール受信を停止したり容量を超えた分を削除したりする処理は行わなかった。

2.3 ファイル共有システム

九州大学では2010年8月から全職員を対象に「ファイル共有システム」と名付けたサービスを提供した。このシステムは、ノースグリッドが開発販売しているオンラインストレージ構築パッケージ「Proself」を導入したものである。

当初、メールの添付ファイル削減を目的にサービスを開始した。そのため、容量やファイルの保持期間の制限を設けていた。表3に容量等の制限値などを示す。表3の時期は、その制限値に設定した時期を表している。

表 3 ファイル共有システム資源量と上限値

時期 (年.月)	2010.8	2011.1	2011.6	2011.9
全ディスク容量	700GB			
利用者の容量	1GB			
保存期間	14 日	28 日	90 日	
利用者	教職員		教職員と学生	

このサービスは学内構成員からの要望が多かったため、容量制限の緩和、保存期間の延長、学生を利用可能にするなどの改善を継続した。

まず、全学基本メールの有償サービス利用者から、ファイル共有システムの容量拡大の要望が出た。そのため、有償サービス利用者の保存容量を2011年9月に10GBに拡大した。さらに容量拡大の要望が出たため、2013年4月には20GBに拡大した。

ファイル共有システムにおける利用者の保存容量制限は、ハードリミットで設定していた。すなわち、容量制限を超えるファイルをシステム上に保持することは出来ないように設定されていた。

3. ディスク使用量分析

全学基本メールについては、2009年5月から2014年3月のサービス終了までディスク使用量等が毎日ログとして出力・保持されていた。またファイル共有システムについては定期的な記録は取得されておらず、2014年1月27日時点におけるディスク使用量を手動で収集した。これらのデータを用いて利用者数とディスク使用量の状況を分析した。

3.1 ファイル共有システムのディスク使用量分布

各利用者のディスク使用量を降順に並べ、その順位を取る。図2は縦軸を使用量、横軸に使用量での順位をとる散布図を描いたものである。図2は2014年1月27日時点での使用量で、ディスク使用量の単位はMBである。

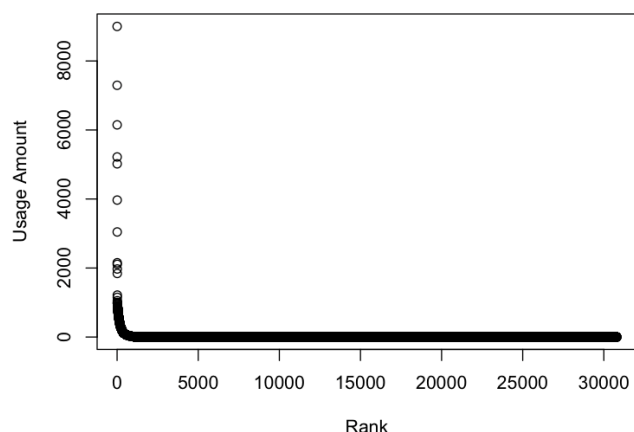


図 2 ファイル共有システム: 順位-使用量
 (2014年1月27日)

図2に示すように、使用量の分布はロングテールの形をしている。極少数の利用者が大容量を使う一方、ほとんどの利用者は小容量しか使わないような分布である。図2では下位の状況が分かりにくいので、図2のグラフの横軸と縦軸を対数尺度で描いた散布図を図3に示す。

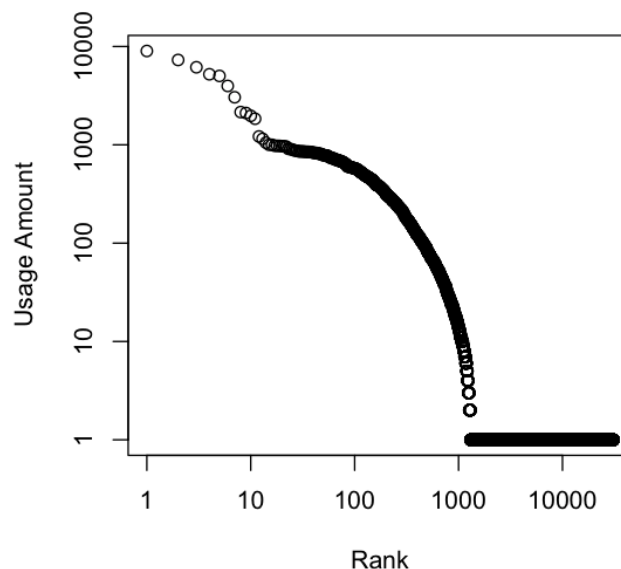


図 3 ファイル共有システム: 順位-使用量 (両対数)
 (2014年1月27日)

次に、図2のデータを用いて非線形回帰分析でディスク使用量の分布を推定する。推定では上位と下位の利用者を除外した。下位の使用量「1」の利用者は、設定ファイルだけが存在するアカウントで、利用者自身のファイルは存在しない。グラフ上位に位置している有償サービス利用者も計算対象外とした。一般利用者の上限値が1GBであるのに対し、有償サービス利用者の上限値は20GBである。このため、使用量分布推定にはノイズとなるためである。以上より、使用量の分布推定計算には、利用量が1GB以下かつ2MB以上となる1,287人の使用量データを用いた。図3のグラフでは、15位から1,302位に位置し、その部分のグラフは弓なりの曲線を描いている。

使用量分布の候補として、ロングテール型になる分布である冪分布、指数分布、対数正規分布の3種類を選び、各分布を仮定して非線形回帰分析を行った。非線形回帰分析にはRを用いた。また、各分布のモデルと実際の使用量分布との適合度合いを調べるため、赤池情報量規準 (AIC) を算出した。AIC値の計算結果を表4に示す。

表 4 ファイル共有システム: 各分布の AIC 値

分布	AIC 値
冪分布	15966.28
指数分布	10559.04
対数正規分布	9244.65

AIC 値が最小のモデルが最適なモデルである。表 4 より、対数正規分布が最適である。対数正規分布モデルによる計算値を、実際の使用量に重ねてグラフを描いたものを図 4 に示す。図 4 を見ると、最上位の利用者以外は、概ね対数正規分布の推定値が実データと合致していることが分かる。

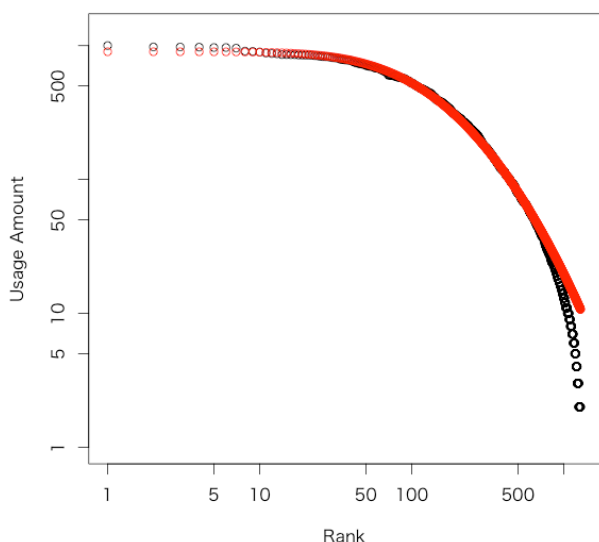


図 4 ファイル共有システム: 順位-使用量 (黒) と、非線形回帰分析で求めた対数正規分布の値 (赤)

3.2 全学基本メールのディスク使用量分布

全学基本メールでは、毎日使用状況データを取得している。各利用者のディスク使用量を降順にソートし、順位を横軸、使用量を縦軸として両対数グラフを描いた。例として 2014 年 1 月 27 日のディスク使用量分布を図 5 に示す。

図 4 の赤色の水平線は、一般利用者の上限値である 300MB を示している。1 位から青色の縦線 (54 位) までの利用者は、使用量が 300MB を超えている。この範囲の利用者は 2 つのグループに分けることができる。まず、1 位から緑色の縦線 (24 位) は、すべて有償利用者である。有償利用者は上限値が 10GB であるため、300MB より大きい値を示している。また、緑色の縦線 (24 位) から青色の縦線 (54 位) までは、有償利用者である 1 アカウントを除いて一般利用者であったが、上限値 300MB を超えて利用している。これは、全学基本メールシステムの利用者の容量制限がソフトリミットであり、上限値を超えても警告メールを送信するだけで、超過したメールを自動的に削除する処理をなかったためである。

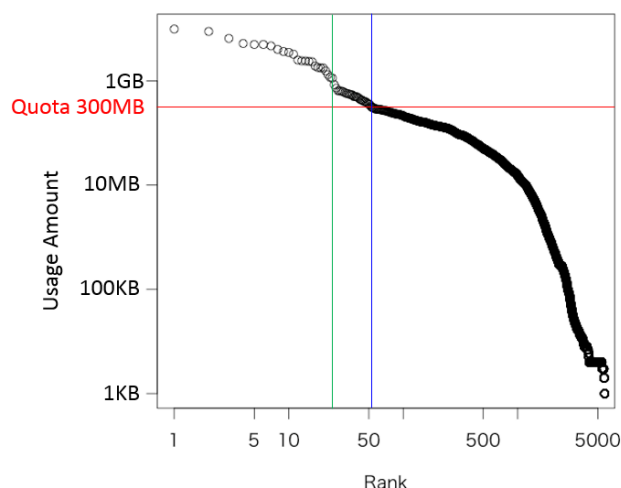


図 5 全学基本メール: 順位-ディスク量 (両対数)
(2014 年 1 月 27 日)

3.1 節と同様に、非線形回帰分析でディスク使用量の分布を調査した。その際、上位の有償利用者および、一般利用者のうち、使用量が上限値 300MB を超えている利用者、さらに下位のアカウントのみが存在する利用者を外して分析した。すなわち、使用量の分布推定計算には、使用量が 300MB 以下かつ 5KB 以上となる 4,139 人のデータを用いた。これは、図 5 のグラフでは青線から右に位置する利用者となる。回帰分析実施後の AIC 値の計算結果を表 5 に示す。

表 5 全学基本メール: 各分布の AIC 値

分布	AIC 値
冪分布	95892.74
指数分布	81720.74
正規分布	98252.47
対数正規分布	79958.36

表 5 より、全学基本メールの使用量分布近似においても、対数正規分布モデルが最適であるという結果を得た。対数正規分布モデルによる計算値を、実際の使用量に重ねてグラフを描いたものを図 6 に示す。図 6 を見ると、使用量の少ない層において、予測値が実際の使用量よりも多くなっているものの、上位 2,000 人については予測値と実際の使用量がよく合致していることがわかる。

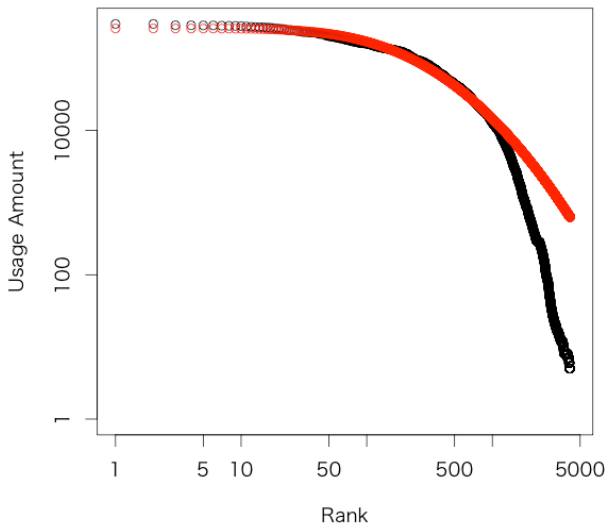


図6 全学基本メール: 順位-使用量 (黒) と、非線形回帰分析で求めた対数正規分布の値 (赤)

3.3 全学基本メールのディスク使用量推移

次に、全学基本メールにおける、ディスク使用量の時間推移を調査した。期間は2009年7月から2014年2月までの間で、毎月15日のデータを用いて推移を調査した。図7に2011年2月～2014年2月の推移状況を示す。

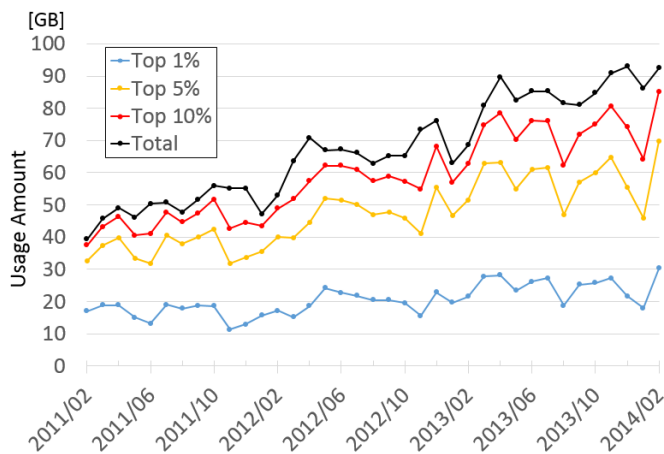


図7 ディスク使用量の時間的推移 (2011年2月～2014年2月)

図7最上位の黒線は、全ユーザのディスク使用量である。以降の線は、上位1%、5%、10%の利用者のディスク使用量を示している。図7の分析では、上限値が10GBの有償利用者は外している。また、人数の割合算出には、アカウント登録だけの利用者は含めていない。

図7より、上位10%の利用者だけで、90%の使用量になることが分かる。運用した全期間において、多くの利用者はディスクを使わず、上位の少数の利用者がディスクを使っていることが分かる。

4. 最適クォータ値設定のための分析

ここでは、最適なクォータ値についての考察を行う。利用者はなるべく大きな容量を使いたい。できれば制限値のないことが望ましいと考えるだろう。一方、現実システムのストレージ容量は有限であるため、システム管理者はストレージ容量の範囲に利用を抑えたい。容量ぎりぎりまで使われると、突発的なメール増加、時間とともに増える使用量増加に対応できない。そのため、ある程度の空きを常に確保しておくことを管理者は望む。

4.1 使用する数式記号

以降の数式で用いる記号を以下に示す。

- S : ディスクの使用上限値,
- u_i : 利用者 i のディスク使用量,
- U : 全使用量 ($U = \sum u_i$),
- q : 上限値 (クォータ値), $\forall i, u_i < q$.

上記の S は、ストレージの実容量ではなく、システム管理者が使われて良いと考える値とする。例えば、メールシステムに容量1TBのストレージが付いており、このうち75%まで使われて良いと管理者が考えるのであれば、 S を750GBとする。

U が S に近いほど、ディスクを十分に活用していると考える。そこで、 S に対する U の割合、 U/S の値をディスクの活用度として評価する。

4.2 九州大学全学基本メールのディスク活用度

全学基本メールにおけるディスク活用度を時系列で描く。使用上限値 S はストレージの全容量の約75%にあたる2000GBとした。期間は2011年2月～2014年2月としている。

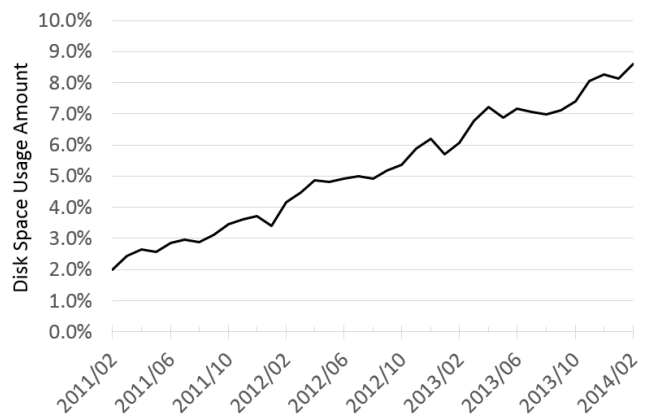


図8 ディスク活用度の推移 (2011年2月～2014年2月)

図8を見ると、九州大学で運用していた全学基本メールは2011年2月から2014年2月までの間、活用度が低いまままで続いていることが分かる。当時、管理者（著者ら）は、利用者が均等にディスクを使用すると単純に想定をしていた。そのためクォータ値を低い値に設定していた。

4.3 ディスク使用量の冪分布近似

第3節で述べたように、実際のディスク使用量は均質ではなく、ロングテールの形をした分布で、上限値近くまで使う利用者は極少数で、大多数の利用者は少ない容量しか使用しない。この使用量分布に基づいて、より活用度合いが高くなるクォータ値を設定する。

適切なクォータ値を推定するために冪分布での近似を考える。最も近い分布は対数正規分布であるものの、冪分布の方が計算を単純化できるため最適クォータ値の推定に良い。対数正規分布のデータを上位値だけ評価すると冪分布で近似できることが知られているため、冪分布での近似で良いと判断している。

冪分布は

$$f(x) = qx^a,$$

で表される。ここで a はスケーリング指数 (scaling exponent) と呼ばれる。この分布を両対数尺度でプロットすると、

$$y = qx^a,$$

$$\log y = a \log x + \log q,$$

$$Y = \log y, X = \log x, Q = \log q \text{ とすると}$$

$$Y = aX + Q,$$

となるため切片が Q で傾きが a の直線になる。

直線であるため単純化できる。図9に両対数グラフでの近似についての関係を示す。縦軸は使用量、横軸は順位である。曲線Aが実際のディスク使用量分布で、直線Bが冪分布での近似値とする。

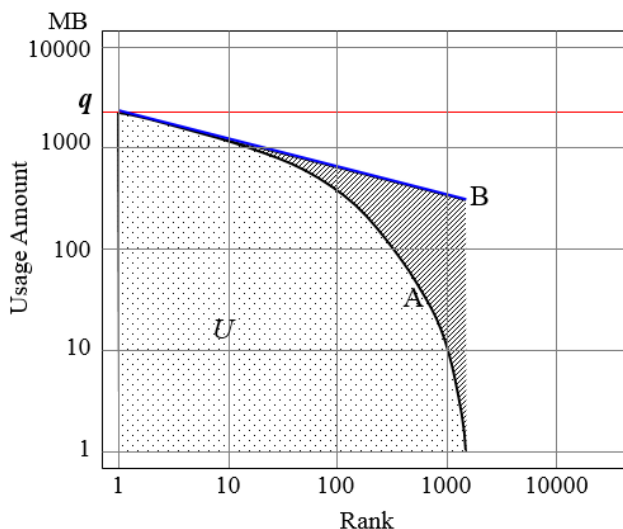


図9 ディスク使用量の実分布と冪分布近似

曲線Aで囲まれた部分は、実際のディスク使用量 U になり、直線Bで囲まれた領域と U との差分(図9の斜線部分)は誤差になる。実際の使用量分布Aから、冪分布の近似値Bを算出する場合、BがAより上になるようにする。こうすれば、誤差の分だけ余るため容量不足にならない。図9のグラフは対数尺度である事を留意されたい。

4.4 冪分布近似に基づくクォータ値推定

ディスクの活用度が低い場合、使用量の上限値(クォータ値)を、 q_1 から q_2 に変更することを考える。ただし $q_1 < q_2$ である。クォータ値を q_2 に変更した際の、ディスク使用量を U_2 とする。ディスクの活用度を高めつつ、 U_2 が全体の使用上限値 S を超えないように q_2 を設定しなければならない。

図10を用いてクォータ値変更に伴う使用量分布の変化を考察する。クォータ値を q_1 から q_2 に変更した際、利用者のディスク使用量は新たな対数正規分布になるであろう。

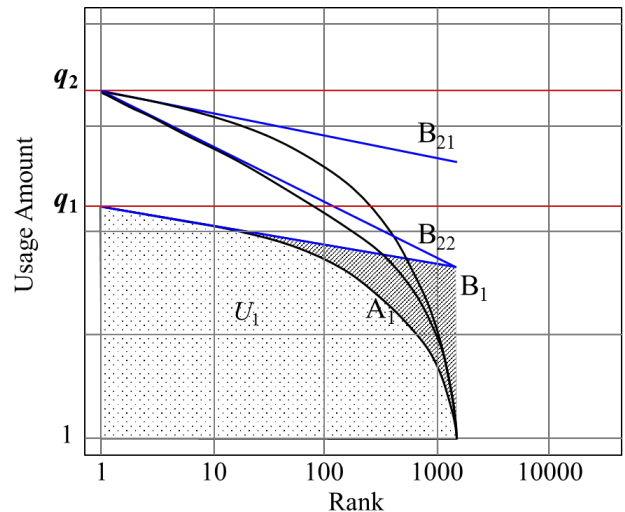


図10 クォータ値変更に伴う使用量分布の変化

クォータ値が q_1 の場合における冪分布近似 B_1 と同様の傾きを、クォータ値が q_2 の場合も維持するならば、図10の B_{21} に近い分布になる。図10の B_{22} のように、傾きが急になる可能性もある。どちらになるのか、著者らには今のところ分かっていない。クォータ値変更後のディスク使用量分布を観測すれば、変更後の分布がどのようになるか判明する。変更後の分布観測は今後の課題である。

活用度向上のためにクォータ値を変更する場合、より安全な推測をする方が良い。図10の B_{21} は B_{22} よりディスク使用量が大きいので、 B_{21} の分布であると想定してクォータ値を設定すべきであろう。

冪分布近似に基づいてストレージ活用度が高くなるクォータ値を推定する。冪分布を $f(x) = qx^a$ とする。このとき、全使用量 U は、

$$U = \sum f(x) = q \sum x^a, (x = 1 \dots n),$$

である。ここでシステムの n は利用者数である。ディスクの活用度 U/S を 100%にするには $U=S$ であるため、この場合のクォータ値 q は、

$$q = S / (\sum x^a), (x = 1 \dots n),$$

であればよい。

4.5 九州大学全学基本メールのクォータ値

先に 4.2 節で示したように、当時の九州大学全学基本メールのディスク活用度は低かった。ここでは当時のシステムで、ディスク活用度を 100%にするためのクォータ値について算出する。算出には 2014 年 1 月 27 日時点の使用量を用いる。単純化のため、大容量を使う有償利用者、および上限値を超えて使用している利用者は省き、また、使っていない利用者は無視した。残る一般利用者は 4,139 人であった。

ディスク使用量分布を冪分布を用いて近似し、図 9 の B に当たる直線の傾きを算出した。算出は分析ソフト R を用いて、線形回帰分析により行った。なお、傾きを算出する際、ディスク使用量が上限値を下回っている一般利用者のうち上位 400 人のデータだけを用いた。これは、図 9 に示すように、上限値が単一の値で固定され、かつ使用量が上限値を超えることがないハードリミットモデルに基づいているため、また、下位利用者のデータを用いると傾きが予測と異なるためである。表 6 に算出した値を示す。

表 6 冪分布パラメータの計算結果

傾き (スケーリング指数)	-0.404485
---------------	-----------

2014 年 1 月 27 日時点の一般利用者の全使用量 U は約 70GB であったが、算出した傾き -0.40 では

$$U = q \sum x^a (x = 1 \dots n),$$

の値が 71.91GB となった。

当時の九州大学全学基本メールのストレージ容量は 2,600GB で、実利用者数 n は 4,139 人であった。ディスクの使用上限値 S を全体の約 75%である 2,000GB とすると、当時のシステムでディスク活用度を 100%にするクォータ値は約 8.1GB であると算出できた。

5. おわりに

本研究では、筆者らが所属する九州大学で運用している、電子メールとファイル共有（蓄積）システムを対象に、長期間のディスク資源の使用量を分析した。システムが定期出力するログを用いて、全ユーザの資源利用の分布を把握した。ディスク使用量の分布を調べた結果、利用者のディスク使用量はロングテール型の分布をしていた。使用量の分布を非線形回帰分析と比較したところ、対数正規分布に近いことが分かった。また、長期的なディスク使用量の推移を調べることで、全体の使用量が微増していく状況も分かった。

また、ディスクの利用効率を向上するため、活用度を定義し、それに基づく評価方法を考案した。さらに、冪分布による近似で、より活用度の高いクォータ値を設定するための計算方法を考案した。実際のデータに基づき、当時のシステムでストレージの活用度を向上するクォータ値も算出した。

今後は、本研究で考案した方法を実運用システムに適用する予定である。2014 年度に運用開始している新メールシステムと、ファイル共有システムのディスク利用状況ログは定期取得している。このログを分析し、本研究と同様の分析を行う。現在のディスク活用度を算出して評価するとともに、実システムのクォータ値を拡大することで活用度の向上と時間経過による推移の検証を行う予定である。

参考文献

- 九州大学全学メールサービス, <http://www.m.kyushu-u.ac.jp/> (accessed at Dec. 10, 2014).
- 藤村直美, 戸川忠嗣, 笠原義晃, 伊東 栄典: 姓名をベースにしたアドレスによる学生基本メールの運用について, 情処研報 Vol.2011-IOT-14, No.10, pp.1-6 (2011).
- Naomi Fujimura, Tadatsugu Togawa, Yoshiaki Kasahara, and Eisuke Ito: Introduction and Experience with the Primary Mail Service based on their Names for Students, Proc. of ACM SIGUCCS'12, pp.11-14 (2012).
- Eisuke Ito, Yoshiaki Kasahara, and Naomi Fujimura: Implementation and operation of the Kyushu university authentication system, Proc. of ACM SIGUCCS2013, ACM, pp.137-142 (2013).
- Yoshiaki Kasahara, Eisuke Ito and Naomi Fujimura: Introduction of New Kyushu University Primary Mail Service for Staff Members and Students, Proc. of ACM SIGUCCS2014, pp.103-106 (2014).
- 國仲寛人, 複雑系にひそむ規則性:対数正規分布を軸にして, 日本物理学会誌, vol.66, no.9, pp.658-665 (2011).
- Michael Mitzenmacher, "A Brief History of Generative Models for Power Law and Lognormal Distributions", Internet Mathematics Vol. 1, No. 2: 226-251, 2004.
- The R Project for Statistical Computing, <http://www.r-project.org/> (accessed at Jan. 30, 2015).