Vol.2015-IOT-30 No.4
2015/7/3

送信された電子メールサイズの頻度モデルの改良

松原 義継[1,2,a)]　武藏 泰雄[3,b)]

概要：とある大学の電子メール送信サーバで処理された電子メールサイズの頻度分布を説明するためのモデルに改良を施した．既存モデルでは頻度分布に表れるべき則の説明を重視したのに対して，本モデルではそのべき則に加え頻度分布のより正確な形を説明可能にした．改良モデルは，ラプラス分布の形を参考に，テキスト/HTML 形式メール用および画像等の添付ファイル付きメール用の分布式を用意し，それらを組み合わせた．今回のモデルと観測データとの相関係数を算出したところ，その値は 1 に近いことを確認した．

キーワード：電子メールサイズ，べき則分布

Improvements of frequency distribution model of e-mail sizes processed in e-mail sending servers

Matsubara Yoshitsugu[1,2,a)]　Musashi Yasuo[3,b)]

Abstract: Previously, we proposed a power-law distribution model to explain the frequency of e-mail sizes obtained from two e-mail servers in the university campus network. The model focuses on the power-law distribution and it can explain the principal fluctuations. We propose an improved model to explain the frequency of that. The improved model can moreover explain the details of the frequency, and the correlation coefficient between the observed data is close to 1.

Keywords: E-mail sizes, Power-law distribution

1. はじめに

現代社会において重要な通信媒体の１つであるインターネットについて，その構造やデータの流量の動的性質を研究対象とする動きがある [1–6]．インターネット上のコミュニケーションツールの 1 つである電子メールの動的性質に関しても，例えば送信間隔に見出されるべき則に従う相関等の報告がある [7–11]．

[1] 熊本大学 大学院
Graduate School of Science and Technology, Kumamoto University, 2-40-1 Kurokami Chuo-ku, Kumamoto-shi, Kumamoto, 860-8555 Japan
[2] 佐賀大学
Saga University, 1 Honjo-machi, Saga-shi, Saga, 840-8502, Japan
[3] 熊本大学
Kumamoto University, 2-40-1 Kurokami Chuo-ku, Kumamoto-shi, Kumamoto, 860-8555 Japan
[a)] 146d9301@st.kumamoto-u.ac.jp
[b)] musashi@cc.kumamoto-u.ac.jp

電子メールの動的性質に関する報告の中に，電子メール送信サーバにリクエストされた電子メールのサイズの頻度分布を分析したものがある [12]．その頻度分布形からは電子メールのサイズおよびそのサイズを有する電子メールの頻度との間にべき則の関係があることを示している．

文献 [13] では，その頻度分布を説明するためのモデルの提案が行われている．そのモデルは，言語学における一文の長さの頻度分布に関する研究成果を電子メールへ拡張したものである．その拡張の内容は，(1) 画像や音声等の文章ではない情報の長さを扱っていること，(2) 頻度分布の領域をサイズの小さい領域 (平文・HTML 形式メール) と大きい領域 (画像・音声等の添付ファイル付きメール) の 2 領域に分割して考えたこと，である．そのモデルでは，電子メールのサイズが大きくなるに従い頻度分布形がべき則に従うことが説明可能である．電子メールのサイズの頻度分布に関する先行研究は，我々の知りうる限り，これら２

つの文献のみである．

本論では，文献 [13] で提案された電子メールサイズの頻度分布モデルの改良型を提案する．そのモデルの特徴の 1 つである頻度分布の領域の 2 分割は，平文・HTML 形式メールのサイズと添付ファイル付きメールのサイズが重なる可能性を考慮されていない．さらに，そのモデルはべき則の説明を重視していることから，電子メールサイズの小さい領域での頻度分布形を説明できているとは言い難い．本論で提案するモデルでは，べき則を説明可能な上に，これら 2 点についても説明可能である．併せて，本論で提案するモデルに基づく頻度分布形と実際の頻度分布形との間の相関係数値が 1 に近いことも示す．

2. 観測データに基づく頻度分布の作成

本論で扱う電子メールサイズの頻度分布形は，文献 [13] で扱われた，佐賀大学の学内ネットワーク内端末群から同大学総合情報基盤センターにてサービス中である電子メール送信サーバに送信され処理された電子メールを基にしている．その送信サーバで処理可能な電子メール 1 通のサイズの上限値は 10 メガバイト [MB] である．佐賀大学で運用されている電子メール送信サーバは，教職員用および主に学生が利用する教育用の 2 種類である．それらのログファイルの期間は，教職員用は 2009 年 4 月 1 日から 2013 年 3 月 31 日までの 5 年間分，教育用は 2008 年 4 月 1 日から 2013 年 3 月 31 日までの 6 年間分である．分析に用いる頻度分布は，ログファイルを教職員用および教育用にそれぞれ分け，分けたログファイルを年度単位でさらに小分けし，小分けされたログファイルを基に 1 キロバイト単位での頻度で集計することにより作成された．

作成した各頻度分布を図 1(教職員用) および図 2(教育用) にそれぞれ示す．両図内での表記 ‘AY’ は 学年度 (academic year) の略記である．各図の縦軸は電子メールサイズの頻度を該当年度の電子メール総数で割った値，横軸は電子メールサイズ (キロバイト単位) である．両軸とも対数化されており，$s = 0$ での値が表されていないことに注意である．図中の各青点は実際の値を表す．赤線は図全体の傾きを表している．傾きは青点全体を最小二乗法でフィッティングすることで得られている．全体の傾きが分かるように赤線の引かれている位置は全体的に上にずらしている．

図 1 および図 2 に描かれている各年度の分布は，一部に形の歪みを読み取れるが，その大きな形は直線的であるように読み取れる．両図中の各頻度分布は両対数グラフであり，両対数グラフ上で直線的であることは，その頻度分布の形はべき則に従っていることを意味している．

3. 改良モデルの内容

図 1 および図 2 内の各頻度分布に対して，以下に説明する改良モデルは良好なフィッティングを示すことが分かった．改良モデルは これら頻度分布図では表されていない $s = 0$ での頻度もフィッティングしてる．

$$p(s) = \frac{1}{a_0 + a_1}\Big\{\frac{a_0}{C_0}\exp\{-\frac{|\ln(s+1) - \ln\mu_0|}{\sigma_0}\} + \frac{a_1}{C_1}\exp\{-\frac{|\ln(s+1) - \ln\mu_1|}{\sigma_1}\}\Big\} \quad (1)$$

ここで, s は電子メールサイズ (≥ 0 step 1[KB]) , $p(s)$ $[0, 1]$ は サイズ s の頻度の割合である．式 1 内の各変数である $\mu_0, \sigma_0, \mu_1, \sigma_1$ の値は $\mu_0 > 0, \sigma_0 > 0, \mu_1 > 0, \sigma_1 > 0, a_0 > 0, a_1 > 0$ の実数である．C_0, C_1 はそれぞれ

$$C_0 = \sum_{s=0}^{s_{max}} \exp\{-\frac{|\ln(s+1) - \ln\mu_0|}{\sigma_0}\} \quad (2)$$

$$C_1 = \sum_{s=0}^{s_{max}} \exp\{-\frac{|\ln(s+1) - \ln\mu_1|}{\sigma_1}\}. \quad (3)$$

である．式 2 および式 3 内の s_{max} は，電子メールサーバが処理可能な電子メールサイズの最大値である．もし s_{max} の値が無制限ならば，両式中の s_{max} の表記は ∞ に置き換わる．

式 1 の第一項目は 平文・HTML 形式メールの場合の頻度分布，第二項目は PDF ファイル・画像ファイル等のファイル添付がある場合の頻度分布に対応する．電子メールサイズ s に対して，その電子メールは第一項目もしくは第二項目のいずれかに対応することから，両項は加算演算子での結合になる．

式 2 および式 3 は，電子メールサイズの上限値無制限において，$0 < \sigma_0 < 1$ および $0 < \sigma_1 < 1$ の条件下でそれぞれ収束する．例えば，式 2 は式中にある絶対値の表記に注意して，

$$C_0 = \{\mu_0^{-\frac{1}{\sigma_0}} \sum_{s=0}^{\lceil\mu_0\rceil - 1} (s+1)^{\frac{1}{\sigma_0}}\} - \{\mu_0^{\frac{1}{\sigma_0}} \sum_{s=0}^{\lceil\mu_0\rceil - 1} (s+1)^{-\frac{1}{\sigma_0}}\} + \mu_0^{\frac{1}{\sigma_0}}\zeta(\frac{1}{\sigma_0})$$

になる．式中にある関数 $\zeta()$ は，リーマンゼータ関数である．リーマンゼータ関数の性質より，式 2 は $1/\sigma_0 > 1$，つまり $0 < \sigma_0 < 1$ の条件下で収束する．式 3 についても同様に示される．

4. 改良モデルと実際の頻度分布との相関

改良モデルにより生成される電子メールサイズの頻度分布および観測データに基づく実際の頻度分布とのフィッティングを行い，併せて相関係数も算出した．本論では，教職員用での 2009 年度および 2013 年度，教育用での 2009 年度および 2013 年度の 4 つの場合をそれぞれ図 3，4，5，6 に示す．対数グラフでは表記不可である $s = 0$ での両値を示すために，各図の横軸は $s + 1$ であることに注意であ

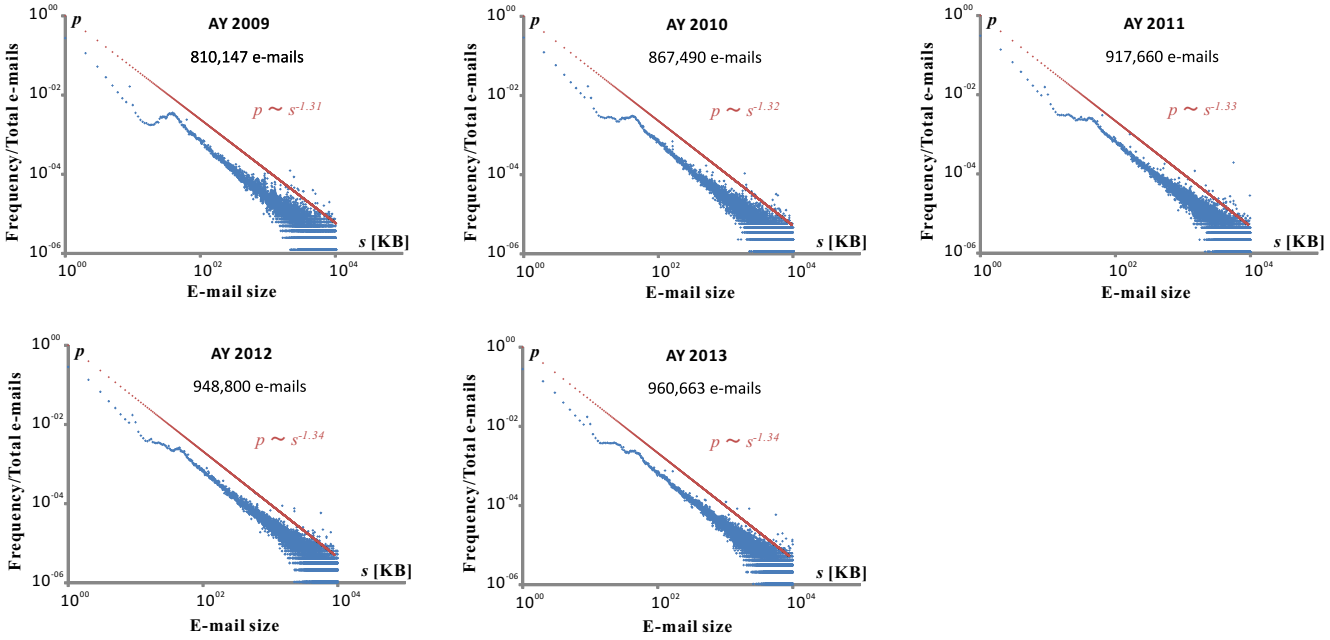

図 1 教職員用電子メール送信サーバにおける電子メールサイズ頻度の年度別変化.

Fig. 1 Frequency in e-mail sizes for staffs per academic year.

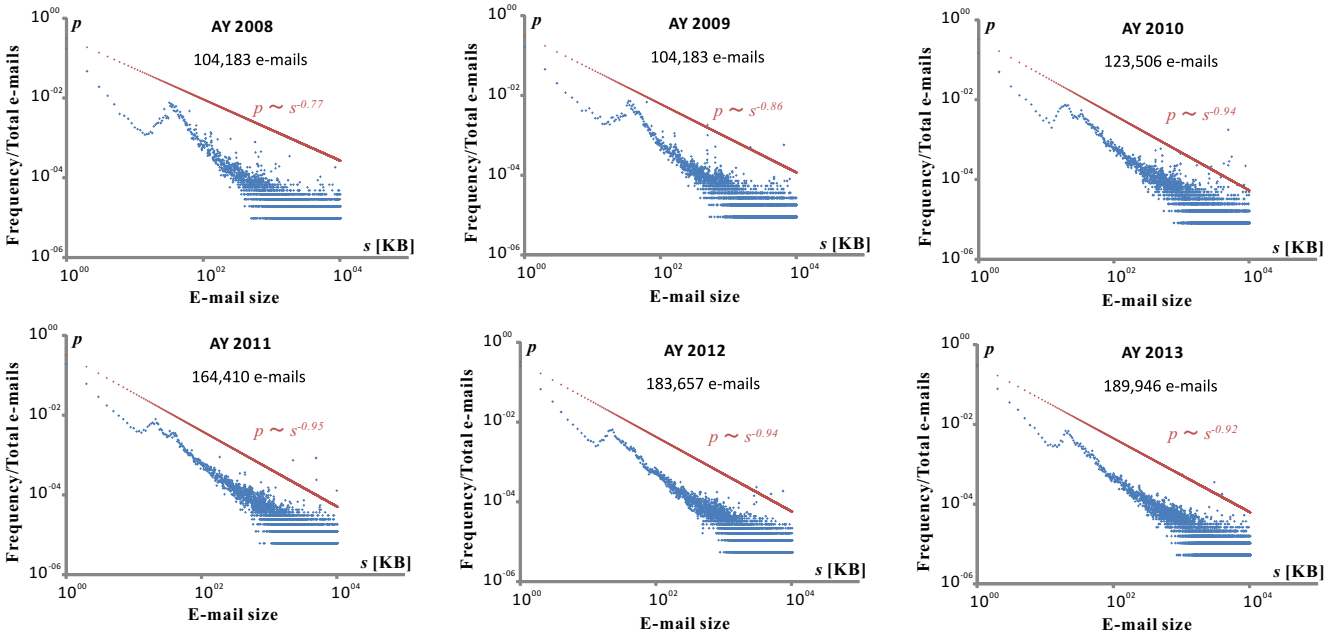

図 2 教育用電子メール送信サーバにおける電子メールサイズ頻度の年度別変化.

Fig. 2 Frequency in e-mail sizes for students per academic year.

る．相関係数値および算出に用いたパラメータ値を表 1 に示す．表 1 の各相関係数値およびそれぞれに対応する図 3,4,5,6 より，改良モデルは実際の頻度分布全体を説明可能であることが分かる．

5. まとめ

電子メール送信サーバで処理された電子メールのサイズの頻度について，それをよりよく説明可能なモデルを提案した．分析に用いたデータは，とある大学で運用されている教職員用および主に学生が利用する教育用の 2 種類の電子メール送信サーバから得られた．それぞれの送信数は，450 万 4760 通 (5 年間分) および 87 万 7271 通 (6 年間分) である．頻度分布は，得られたデータを基に 1 キロバイト単位で集計し，教職員用での各年度および教育用での各年度で作成された．得られた各頻度分布形は，その大きな形としてべき則を読み取れる．

文献 [13] で提案された電子メールサイズの頻度分布モデルは，言語学で研究されている 1 文の長さの頻度分布を電子メールへ拡張している．その式には，ガンマ分布に画像やグラフのような非文情報の影響が組み込まれており，平

表 1 相関係数値および算出に用いたパラメータ値.
Table 1 Correlation coefficients and the parameters.

年度	対応図	相関係数値	パラメータ値
教職員用 2009 年度	図 3	0.99907	$a_0 = 2.62547, \mu_0 = 1.57605, \sigma_0 = 0.42484, a_1 = 0.947001, \mu_1 = 38.8706, \sigma_1 = 0.68236$
教職員用 2013 年度	図 4	0.99924	$a_0 = 1.70517, \mu_0 = 2.12383, \sigma_0 = 0.40962, a_1 = 0.75421, \mu_1 = 23.1444, \sigma_1 = 0.88246$
教育用 2009 年度	図 5	0.99758	$a_0 = 0.9, \mu_0 = 1.2, \sigma_0 = 0.38, a_1 = 0.35, \mu_1 = 37.5, \sigma_1 = 0.6$
教育用 2013 年度	図 6	0.99909	$a_0 = 1.0, \mu_0 = 1.46883, \sigma_0 = 0.36093, a_1 = 0.4, \mu_1 = 23.2086, \sigma_1 = 0.63113$

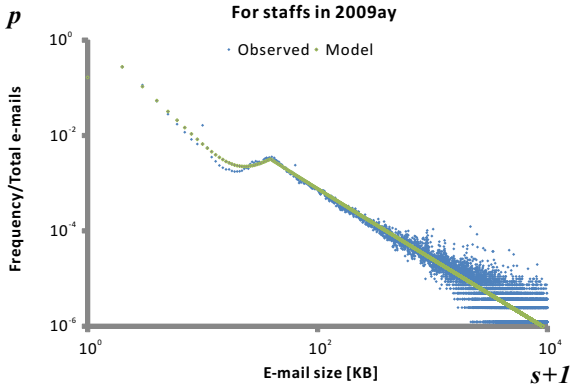

図 3 教職員用 2009 年度の頻度分布と改良モデルとのフィッティング結果．相関係数値は 0.99907．比較対象の頻度分布は図 1 の教職員用 2009 年度の頻度分布.

Fig. 3 Fitting the model to observed data of academic year 2009 for staffs.

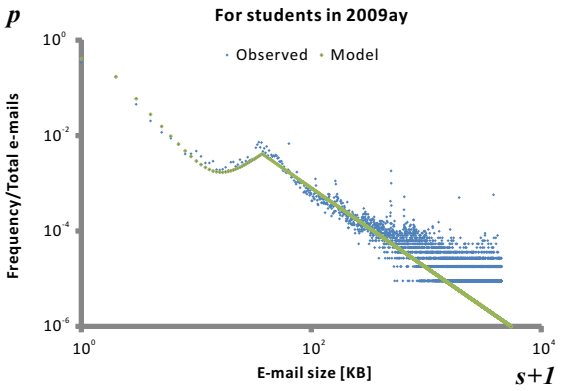

図 5 教育用 2009 年度の頻度分布と改良モデルとのフィッティング結果．相関係数値は 0.99758．比較対象の頻度分布は図 2 の教育用 2009 年度の頻度分布.

Fig. 5 Fitting the model to observed data of academic year 2009 for students.

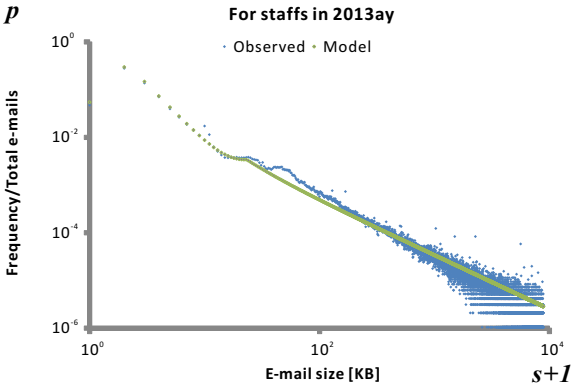

図 4 教職員用 2013 年度の頻度分布と改良モデルとのフィッティング結果．相関係数値は 0.99924．比較対象の頻度分布は図 1 の教職員用 2013 年度の頻度分布.

Fig. 4 Fitting the model to observed data of academic year 2013 for staffs.

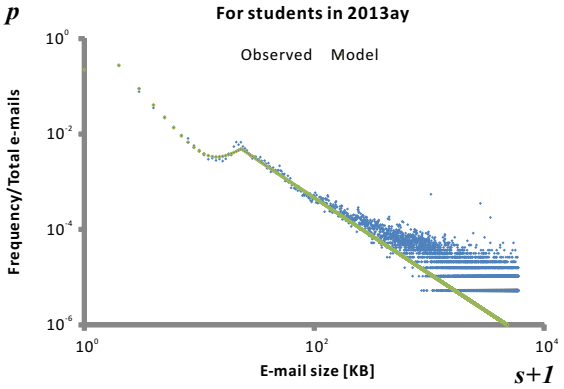

図 6 教育用 2013 年度の頻度分布と改良モデルとのフィッティング結果．相関係数値は 0.99909．比較対象の頻度分布は図 2 の教育用 2013 年度の頻度分布.

Fig. 6 Fitting the model to observed data of academic year 2013 for students.

文・HTML 形式メールの分布領域および非文情報が添付されている電子メールの分布領域で構成されている．そのモデルでは，べき則を説明可能であるが，電子メールサイズの小さい領域での頻度分布形の説明に難がある上に，平文・HTML 形式メールの分布領域および非文情報が添付されている電子メールの分布領域との繋がりに難がある．

そこで，本論ではこれらを改良したモデルを提案した．改良モデルおよび観測データに基づく実際の頻度分布との間で行ったフィッティングの結果を見る限り，今回提案した改良モデルは実際の頻度分布全体を説明可能である．もし，電子メールの本文の分析もしくは電子メールヘッダの 1 つであるコンテンツタイプ [14–19] の分析が可能になるならば，モデルは更に改良可能になる．これは今後の課題である．

参考文献

[1] Faloutsos, M., Faloutsos, M. P. and Faloutsos, C.: On Power–Law Relationship of the Internet Topology, *Pro-*

ceedings of the ACM SIGCOMM, Vol. 29, pp. 251–262 (1999).

[2] Paxson, V. and Floyd, S.: Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. Networking*, Vol. 3, pp. 226–244 (1995).

[3] Csabai, I.: 1/f noise in computer network traffic, *Journal of Physics A: Mathematical and General*, Vol. 27, No. 12, p. L417 (online), available from ⟨http://stacks.iop.org/0305-4470/27/i=12/a=004⟩ (1994).

[4] Takayasu, M., Takayasu, H. and Sato, T.: Critical behaviors and $1/f$ noise in information traffic, *Physica A*, Vol. 233, pp. 824–834 (1996).

[5] Tadaki, S.: Power-Law Fluctuation in Internet Traffic, *Journal of the Physical Society of Japan*, Vol. 76, No. 3, pp. 044001–044001–5 (2007).

[6] Karsai, M., Kaski, K., Barabási, A. L. and Kertész, J.: Universal features of correlated bursty behaviour, *Scientific Reports*, Vol. 2 (online), available from ⟨http://dx.doi.org/10.1038/srep00397⟩ (2012).

[7] Eckmann, J. P., Moses, E. and Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic, *Proceedings of the National Academy of Sciences*, Vol. 101, No. 40, pp. 14333–14337 (online), available from ⟨http://www.pnas.org/content/101/40/14333⟩ (2004).

[8] Barabási, A. L.: The origin of bursts and heavy tails in human dynamics, *Nature*, Vol. 435, pp. 207–211 (2005).

[9] Goh, K. I. and Barabási, A. L.: Burstiness and memory in complex systems, *EPL (Europhysics Letters)*, Vol. 81, No. 4, p. 48002 (online), available from ⟨http://stacks.iop.org/0295-5075/81/i=4/a=48002⟩ (2008).

[10] Malmgren, R. D., Stouffera, D. B., Motter, A. E. and Amaral, L. A. N.: A Poissonian explanation for heavy tails in e-mail communication, *Proceedings of the National Academy of Sciences*, Vol. 105, No. 47, pp. 18153–18158 (online), available from ⟨http://www.pnas.org/content/105/47/18153⟩ (2008).

[11] Anteneodo, C., Malmgren, R. D. and Chialvo, D. R.: Poissonian bursts in e-mail correspondence, *The European Physical Journal B*, Vol. 75, pp. 389–394 (online), available from ⟨http://www.springerlink.com/content/t1321475062jm273/⟩ (2010).

[12] Matsubara, Y., Hieida, Y. and Tadaki, S.: Fluctuation in e-mail sizes weakens power-law correlations in e-mail flow, *The European Physical Journal B*, Vol. 86 (online), available from ⟨http://dx.doi.org/10.1140/epjb/e2013-40209-x⟩ (2013).

[13] 松原義継，武藏泰雄：送信された電子メールサイズの頻度に現れるべき則の分析，インターネットと運用技術シンポジウム（IOT），Vol. 7, pp. 71–77（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00107198/⟩ (2014).

[14] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies, RFC 2045 (Draft Standard) (1996). Updated by RFCs 2184, 2231, 5335, 6532.

[15] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types, RFC 2046 (Draft Standard) (1996). Updated by RFCs 2646, 3798, 5147, 6657.

[16] Moore, K.: MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text, RFC 2047 (Draft Standard) (1996). Updated by RFCs 2184, 2231.

[17] Freed, N. and Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples, RFC 2049 (Draft Standard) (1996).

[18] Freed, N. and Klensin, J.: Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures, RFC 4289 (Best Current Practice) (2005).

[19] Freed, N., Klensin, J. and Hansen, T.: Media Type Specifications and Registration Procedures, RFC 6838 (Best Current Practice) (2013).