DNSアクセスの uniq率に基づく外れ値検知の試み

松原 義継[1,2,a)]　武藏 泰雄[1,b)]

概要：DNS アクセスのエントロピーに基づく外れ値検知よりも簡易な方法として uniq 率に基づく外れ値検知を考案した．本論文における uniq 率とは，単位時間当りの DNS サーバへのアクセス数に対する，クライアント数もしくはクエリの種類数の割合である．もし uniq 率がエントロピーと比較して実用上の支障がないならば，uniq 率はエントロピーの前処理として期待できる．エントロピーと uniq 率の関係を検討するため，実際の DNS アクセスのデータを基に クライアントおよびクエリに基づく uniq 率とエントロピーの時系列を作成した．作成された時系列を比較検討したところ，uniq 率には実用性の可能性はあることが分かった．

キーワード：DNS ログデータ，エントロピー，uniq 率

An attempt of outlier detections by non-parametric method

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

Abstract: We developed “uniq rate” as an outlier detection method. The uniq rate in this paper is the ratio of the kind of client or query for the number of queries to the DNS server per unit time (the number of access). If the uniq rate is not practical problem compared with entropies, the uniq rate can be expected as a pretreatment for entropy. We collected actual DNS access data, and made timeseries of uniq rate and entropy to compare them. As a result of the comparison, we found that the uniq rate is possible for usefulness to detect outliers.

Keywords: DNS log data, entropy, uniq rate.

1. はじめに

現代社会においてインターネットに代表されるコンピュータネットワークは，その社会的重要性を増している．その一方で，DoS (Deny of Services) のようなサービス不能攻撃やシステム上の脆弱性を悪用した攻撃があり，これらは社会にとって憂慮すべき事態である．これらの攻撃を速やかに検知することは，攻撃によるサービスへの悪影響を最小限に食い止める上で欠かせない技術であり，様々な報告がある [1–7].

本論文では DNS アクセスのエントロピーに基づく外れ値検知よりも計算量の少ない方法として uniq 率に基づく外れ値検知を考案した．DNS は，電子メールやウェブページのようなネットワークサービスにおける基本的なサービスであり，多くのネットワークサービスから利用されている．そのため，DNS アクセスの分析からは，複数のネットワークサービスのアクセス動向を効率良く得ることが期待できる．

エントロピーを用いる方法は，DoS 攻撃や DDoS 攻撃のような通常とは異なる偏りのあるアクセス動向を検知できる [8–11]．エントロピーはその定義上，DNS アクセスの内訳に基づく計算を行うため計算量は多くなりやすい．そこで，我々はエントロピーよりも計算量の少ない方法として uniq 率を考案した．もし，uniq 率がエントロピーの簡易

[1] 熊本大学
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
[a)] 146d9301@st.kumamoto-u.ac.jp
[b)] musashi@cc.kumamoto-u.ac.jp

版としての実用性を認められるのであれば，uniq 率はエントロピーに基づく外れ値検知を行うか否かを判断する前処理として外れ値検知全体の計算量削減を期待できる．

本論文では，uniq 率の概要 および uniq 率の実用性を検討するための同一の DNS データに基づくエントロピーとの比較を述べる．

2. エントロピー

本論文で用いるエントロピーの定義および今回考案した uniq 率の定義を示す．

DNS アクセスのエントロピーとは，DNS アクセスの内訳のばらつきの程度に対応する．エントロピーは元々は情報量として考案されたものであるが，外れ値検知手法としても用いられている．

ある観測時間の区間 $[t, t+\Delta t)$ での確率変数を X_t，DNS アクセス数を n_t とする．これ以後，観測時刻 t とは $[t, t+\Delta t)$ を意味する．X_t の値が i となる確率密度を p_i で表す．この時，X_t のエントロピー $H(X_t)$ は

$$H(X_t) = -\sum_{i \in X_t} p_i \log_2 p_i \tag{1}$$

である．

$H(X_t)$ の値の取り得る範囲は，

$$0 \leq H(X_t) \leq \log_2 n_t \tag{2}$$

である．最小値である $H(X_t) = 0$ は，n_t 回全てのアクセス内容が同一であること ($p_i = 1\ (= n_t/n_t)$) を意味する．最大値である $H(X_t) = \log_2 n_t$ は，各アクセス内容に重複のないこと ($p_i = 1/n_t$) を意味する．

ここで本論文では，エントロピー $H(X_t)$ の値の取り得る範囲を $[0, 1]$ に正規化する．その理由は，$H(X_t)$ の取り得る値の上限は，観測時刻で異なる値を取り得る n_t に依存するためである．異なる観測時刻で $H(X_t)$ の値の取り得る範囲が異なると，異なる時刻でのエントロピー値同士を比較することに支障をきたす懸念がある．そこで本論文では，取り得る範囲を正規化することにより，エントロピー値の取り得る範囲に対する観測時刻依存性をなくす．

本論文で用いるエントロピーは

$$H'_t = \frac{H(X_t)}{\log_2 n_t} \tag{3}$$

で表す ($0 \leq H'_t \leq 1$)．

本論文では，正規化された 2 種類のエントロピーを用いる．それぞれを以下のように表す．

- $H'_{t,\mathrm{c}}$: クライアント数の内訳に基づく正規化されたエントロピー
- $H'_{t,\mathrm{q}}$: クエリの内訳に基づく正規化されたエントロピー

3. uniq 率

uniq 率は，2 節にて定義した DNS アクセス回数 n_t におけるクライアント数もしくはクエリの種類数との割合である．t におけるクライアント数もしくはクエリの種類数を $uniq_t$ として，uniq 率 U_t を以下のように定義する．

$$U_t = \frac{uniq_t}{n_t}. \tag{4}$$

例えば，ある時刻の DNS アクセス回数が 1 万回で，DNS アクセスしたクライアント数が 100 台の場合，その uniq 率は 0.01 になる．

uniq 率は その定義上，DNS アクセスの内訳はエントロピー程には考慮されない．例えば クライアント数に基づく uniq 率の場合，各クライアントの DNS アクセス回数は uniq 率には反映されない．uniq 率に反映されるのは，クライアント数のみとなる．そのため，uniq 率の計算量はエントロピーよりは少なくなる．

uniq 率の値の取り得る範囲は，その定義より

$$\frac{1}{n_t} \leq U_t \leq 1\ (= \frac{n_t}{n_t}) \tag{5}$$

である．

最小値である $1/n_t$ は，n_t 回のアクセス内容が全て同じこと (1 種類) を意味する．一方，最大値である 1 は，n_t 回の各アクセス内容に重複のないこと (n_t 種類) を意味する．

本論文では，uniq 率 U_t の値の取り得る値を $[0, 1]$ に正規化する．その理由は，U_t の取り得る値の下限は，観測時刻で異なる値を取り得る n_t に依存するためである．異なる観測時刻で U_t の値の取り得る範囲が異なると，異なる時刻での uniq 率を比較することに支障をきたす懸念がある．

そこで，本論文で用いる uniq 率を

$$U'_t = -\log_{n_t} U_t \tag{6}$$

で表す．U'_t の値の取り得る範囲は，式 5 全体を $-\log_{n_t}$ で対数化し整理することにより，

$$1 \geq U'_t \geq 0 \tag{7}$$

になる．

U'_t の値の取り得る範囲の不等号は， 式 5 および前節で述べたエントロピー H'_t の値の取り得る範囲と逆になる．U'_t の値が 0 に近づくほど，H'_t の値は 1 に近づく．逆に，U'_t の値が 1 に近づくほど，H'_t の値は 0 に近づく．そのため，次節以降で行う実際の DNS アクセスデータに基づく uniq 率とエントロピーとの比較では，互いの値の傾向は逆方向に現れる．

本論文では，2 種類の uniq 率を用いる．それぞれを以下のように表す．

- $U'_{t,\mathrm{c}}$: クライアント数の内訳に基づく正規化された uniq 率
- $U'_{t,\mathrm{q}}$: クエリの種類数に基づく正規化された uniq 率

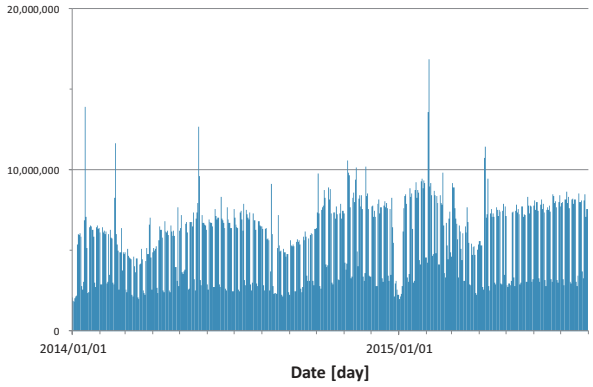

図 1 学内からの DNS アクセス回数の時系列データ.
Fig. 1 Time series of DNS access counts from the campus network.

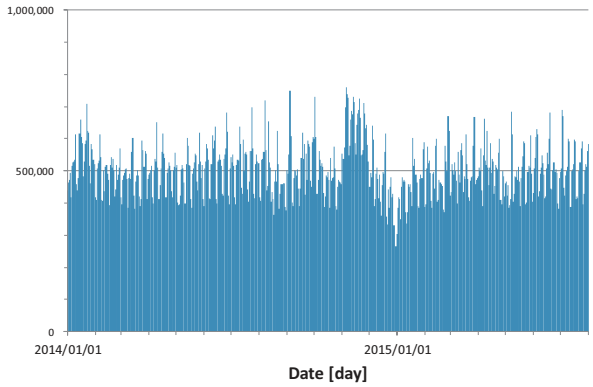

図 2 学外からの DNS アクセス回数の時系列データ.
Fig. 2 Time series of DNS access counts from the Internet.

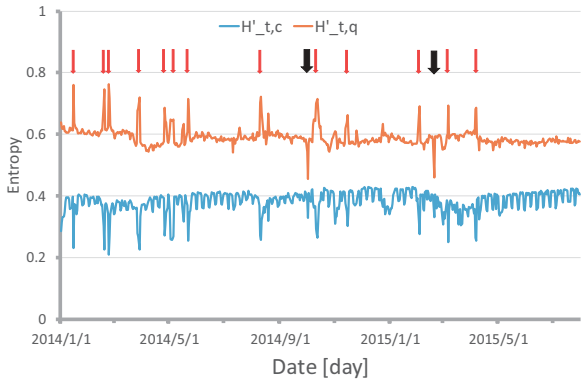

図 3 学内からのエントロピーの時系列データ.
Fig. 3 Time series of entropies from inside.

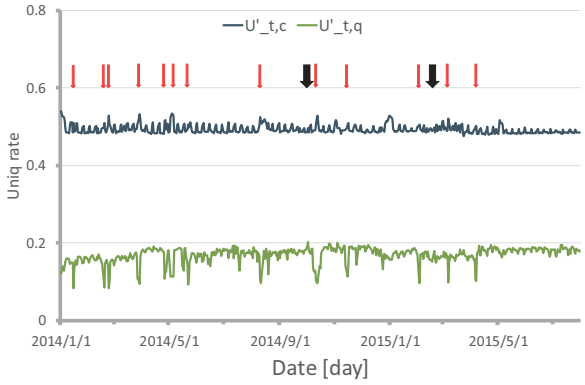

図 4 学内からの uniq 率の時系列データ.
Fig. 4 Time series of uniq rates from inside.

4. DNS アクセスデータ収集

uniq 率とエントロピーとの比較を行うため，佐賀大学の DNS サーバで用いられている BIND 9 [12] のログファイルから DNS アクセスデータを集計した．集計期間は，2014 年 1 月 1 日から 2015 年 7 月 31 日である．その集計期間内のログファイルを基に，1 日単位で DNS アクセスデータを学内からのアクセスおよび学外からのアクセスに分けて集計した．集計単位である 1 日は，2 節にて述べた Δt に対応する．

学内からの DNS アクセス回数および学外からの DNS アクセス回数それぞれの時系列データを図 1 および図 2 に示す．図 1 からは，一時的に DNS アクセス数の増加した日を読み取れ，全体としては学内の年度行事および 1 週間周期に基づく変動を読み取れる．図 2 からは，図 1 程ではないが，全体としては年末年始の DNS アクセスの減少および 1 週間周期変動を読み取れる．

5. uniq 率とエントロピーの比較

前節で収集した DNS アクセスデータを基に，uniq 率 $U'_{t,\mathrm{c}}, U'_{t,\mathrm{q}}$ およびエントロピー $H'_{t,\mathrm{c}}, H'_{t,\mathrm{q}}$ の時系列データを作成した．

始めに，学内からの DNS アクセス回数に基づくエントロピーの時系列データを図 3 に示す．横軸は観測日，縦軸はエントロピー $H'_{t,\mathrm{c}}$ および $H'_{t,\mathrm{q}}$ の値である．図 3 からは，15 個の注視するデータを読み取れる．その 15 個の内，赤い矢印で示している 13 個のデータの意味は，$H'_{t,\mathrm{c}}$ および $H'_{t,\mathrm{q}}$ の変動方向より，一部端末からの DNS アクセスの増加およびクエリの内容の分散化の傾向である．残り 2 個の黒い矢印のデータの意味は，一部端末からの DNS アクセスの増加およびクエリの内容の集中化の傾向である．

学内からの DNS アクセス回数に基づく uniq 率の時系列データを図 4 に示す．横軸は観測日，縦軸は uniq 率 $U'_{t,\mathrm{c}}$ および $U'_{t,\mathrm{q}}$ の値である．エントロピーの時系列データである図 3 に記している矢印を図 4 にも記している．図 4 のデータを図 3 と比較したところ，クエリの種類数に基づく $U'_{t,\mathrm{q}}$ については図 3 で記した矢印のところの傾向を読み取ることは可能である．クライアント数に基づく $U'_{t,\mathrm{c}}$ については，$U'_{t,\mathrm{c}}$ のスケールを拡大することで図 3 で記した矢印のところの傾向を読み取ることは可能である．これらのことから，学内からの DNS アクセスについては，uniq 率はエントロピーの代わりとして実用になる可能性はある．

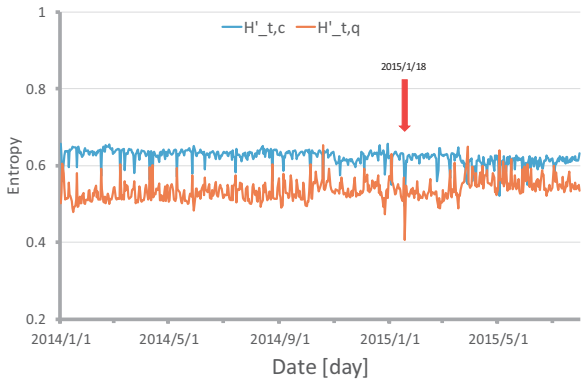

図 5 学外からのエントロピーの時系列データ.
Fig. 5 Time series of entropies from outside.

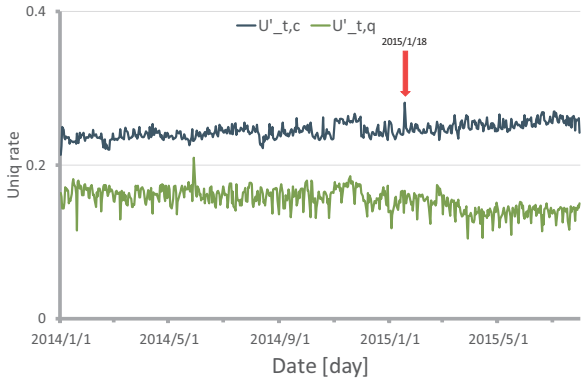

図 6 学外からの uniq 率の時系列データ.
Fig. 6 Time series of uniq rates from outside.

次に，学外からの DNS アクセスに基づくエントロピーの時系列データを図 5 に示す．横軸は観測日，縦軸はエントロピー $H'_{t,c}$ および $H'_{t,q}$ の値である．図 5 からは，1 個の注視するデータを読み取れる．赤い矢印で示しているその 1 個のデータの意味は，$H'_{t,c}$ および $H'_{t,q}$ の変動方向より，一部端末からの DNS アクセスの増加およびクエリの内容の集中化の傾向である．

学外からの DNS アクセスに基づく uniq 率の時系列データを図 6 に示す．横軸は観測日，縦軸は uniq 率 $U'_{t,c}$ および $U'_{t,q}$ の値である．エントロピーの時系列データである図 5 に記している矢印を図 6 にも記している．図 6 のデータを図 5 と比較したところ，クライアント数に基づく $U'_{t,c}$ については図 5 で記した矢印のところの傾向を読み取ることは可能である．一方，クエリの種類数に基づく $U'_{t,q}$ については，図 5 に記した矢印のところの傾向は図 6 で読み取り難い．これらのことから，学外からの DNS アクセスについては，更なるデータに基づく考察が必要と思われる．

6. まとめと議論

DNS サーバのアクセスログを基に uniq 率を用いた外れ値検知を考案した．uniq 率は，単位時間当りの DNS サーバへのアクセス数に対する，クライアント数もしくはクエリの種類数の割合である．uniq 率を考案した目的は，エントロピーよりも少ない計算量で外れ値検知を行うことである．uniq 率はエントロピーに基づく外れ値検知の前処理として，エントロピーのみによる検出よりも外れ値検知のための全体計算量の削減を期待できる．uniq 率およびエントロピーは，それぞれの定義上，値の取り得る範囲が各観測時刻の DNS アクセス数に依存することから，正規化されたものを用いる．uniq 率を用いた外れ値検知をエントロピーの場合と比較検討するため，佐賀大学における 2014 年 1 月 1 日から 2015 年 7 月 31 日までの DNS アクセスデータから uniq 率およびエントロピーの時系列データを作成した．作成された時系列データを比較したところ，学内からの DNS アクセスについては，uniq 率はエントロピーの代わりとなれる可能性が分かった．

本論文では，uniq 率の値とエントロピーの値との関係に対する理論考察を述べるまでには至らなかったことから，これは今後の課題である．

参考文献

[1] Feinstein, L., Schnackenberg, D., Balupari, R. and Kindred, D.: Statistical approaches to DDoS attack detection and response, *DARPA Information Survivability Conference and Exposition, 2003. Proceedings*, IEEE, pp. 303–314 (2003).
[2] Hodge, V. J. and Austin, J.: A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, pp. 85–126 (2004).
[3] Celenk, M., Conley, T., Willis, J. and Graham, J.: Anomaly detection and visualization using Fisher Discriminant clustering of network entropy, *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, IEEE, pp. 13–16 (2008).
[4] Lee, K., Kim, J., Kwon, K. H., Han, Y. and Kim, S.: DDoS attack detection method using cluster analysis, *Expert Systems with Applications*, Vol. 34, pp. 1659–1665 (online), DOI: 10.1016/j.eswa.2007.01.040 (2008).
[5] Lu, K., Wu, D., Fan, J., Todorovic, S. and Nucci, A.: Robust and efficient detection of DDoS attacks for large-scale internet, *Computer Networks*, Vol. 51, pp. 5036–5056 (online), DOI: 10.1016/j.comnet.2007.08.008 (2007).
[6] Xiao, B., Chen, W. and He, Y.: An autonomous defense against SYN flooding attacks: Detect and throttle attacks at the victim side independently, *Journal of Parallel and Distributed Computing*, Vol. 68, pp. 456–470 (online), DOI: 10.1016/j.jpdc.2007.06.013 (2008).
[7] 山西健司：データマイニングによる異常検知，共立出版 (2009).
[8] Takeda, Y., Musashi, Y. and Moriyama, K. S. T.: DNS ANY Request Cannon Activity in DNS Query Packet Traffic, *International Journal of Intelligent Engineering and Systems*, Vol. 7, No. 1, pp. 8–16 (2014).
[9] Musashi, Y., Takeda, Y., Shibata, N., Kubota, S. and Sugitani, K.: A Statistical Study of ANY Resource Record Based DNS Query Request Packet Traffic, *Information*, Vol. 16, No. 12(B), pp. 8901–8908 (2013).
[10] Takemori, K., Kong, W. J., na Romaña, D. A. L., Kub-

ota, S., Sugitani, K. and Musashi, Y.: Entropy Study on A Resource Record Query Traffic from the Campus Network, *IPSJ SIG Technical Reports, Internet Operation and Technology 4th (IOT4)*, Vol. 2009, No. 21, pp. 101–106 (2009).

[11] na Romaña, D. A. L. and Musashi, Y.: Entropy Based Analysis of DNS Query Traffic in the Campus Network, *Proceedings of The 4th International Conference on Cybernetics and Information Technologies, System and Applications (CITSA2007)*, Vol. 6, No. 5, pp. 162–164 (2007).

[12] Internet Systems Consortium: BIND home page. `http://www.isc.org/downloads/BIND/`.