Vol.2019-IOT-46 No.15
2019/6/14

ドメイン名の登録文字列に注目した機械学習による登録更新率の予測

米谷嘉朗[†1]　森健太郎[†1]　井本税[†1]

概要：ドメイン名はインターネットアクセスの基盤であり安定運用が求められる．ドメイン名の運用にはドメインレジストリ，レジストラ，DNS プロバイダ等様々な事業者が関わるが，本論文ではレジストリに集中する．レジストリのサービスは，新規登録ドメイン名数が増加することと，そのドメイン名が長期にわたり継続運用（登録更新）されることによって安定する．したがって，新規登録されたドメイン名が登録更新される確率を事前に予測し，レジストリサービス維持計画を作成することはインターネットアクセスの安定運用にとって重要である．本論文では，新規登録されたドメイン名が登録更新される確率を，ドメイン名の文字列の特徴に基づき機械学習を用いて即時に予測する手法を提案する．また，ドメイン名として登録される文字列の傾向は企業・行政等による様々な施策や社会イベントに影響を受けるため，傾向変化への対応を考察する．本手法を用いた JP ドメイン名での更新率予測の予測精度検証では，従来型の統計的な予測に比べて 1.3～1.7%の精度向上を得ることができた．

キーワード：ドメイン名，機械学習

Prediction of Domain Name Renewal Rate by Machine Learning Focusing on Registered Domain Name String

YOSHIRO YONEYA[†1]　KENTARO MORI[†1]
OSAMU INOMOTO[†1]

1. はじめに

ドメイン名はインターネットアクセスの基盤であり，ドメイン名とサービスを提供するサーバの IP アドレスを対応付ける名前解決サービス（Domain Name Service; DNS）やその登録情報管理等，運用の安定性が強く求められる．ドメイン名の運用にはドメイン名登録情報を管理するドメインレジストリ（以降レジストリ），ドメイン名登録者からの申請をレジストリに取り次ぐドメインレジストラ（以降レジストラ），利用者に DNS サーバを提供する DNS プロバイダ等様々な事業者が関わるが，本論文では，安定運用の重要性が高いトップレベルドメイン名（TLD）のレジストリに集中する．

.JP は日本の国別 TLD（Country Code TLD; ccTLD）であり，JPRS[†1] がレジストリとして第 2 レベル（汎用 JP ドメイン名）および第 3 レベル（属性型 JP ドメイン名）のドメイン名登録情報を管理している．

汎用 JP ドメイン名の登録はローカルプレゼンス（国内に組織の実体が存在すること）を要件とし，試用期間（一定期間ドメイン名登録料が猶予されること）や即時廃止（廃止手続き完了後，速やかに廃止処理が行われること）がなく，1 年毎登録更新のドメイン名であり [a]，年間の新規登録数が 12 万件程度，その 70%程度が 1 年後に登録更新されている．

.JP のような登録要件がある ccTLD は分野別 TLD（Generic TLD; gTLD）に比べて安定な（登録期間が長い）ドメイン名が多いと言われている[1]．レジストリのサービスは，新規登録ドメイン名数が増加すること，それらドメイン名が長期にわたり継続運用（登録更新）されることを原資として安定する．

一方で，企業・行政等による様々な施策や社会的イベントのために新規登録ドメイン名の傾向が変化し，期間限定で利用され登録更新されないドメイン名が一時的に急増することがある．したがって，新規登録されたドメイン名が登録更新されるか事前に予測し，レジストリサービスの維持計画を作成することはインターネットアクセスの安定運用にとって重要である．

本研究では，既存研究で行われているネームサーバの設定状況や Web サイトの開設状況の確認を行わずとも，そのドメイン名が新規登録された時点の文字列情報だけで登録更新される確率を予測する手法を考案・検証した．

本手法はドメイン名の文字列から得られる情報のみを使用するため，管理しているドメイン名の網羅的なリストを持つ他のレジストリ・レジストラや，ドメイン名登録者・研究者に対しても適用可能であると考えられる．

†1 株式会社日本レジストリサービス
Japan Registry Services Co., Ltd.

a) JP ドメイン名のライフサイクル，https://jprs.jp/about/dom-rule/lifecycle/

Vol.2019-IOT-46 No.15
2019/6/14

2. 関連研究

ドメイン名はインターネットアクセスの基盤であるため，特にインターネット利用者への影響の観点から，多くの研究が行われている．

(1) gTLD の悪性ドメイン名検知に関する研究

悪性ドメイン名とは，著名・重要ドメイン名と類似した文字列を使用したり，廃止されたドメイン名を元の登録者とは異なる第三者が登録する等によりインターネット利用者の誤認や混同を誘発し，利用者の個人情報や財産の窃用等悪用を行うドメイン名を指す．全世界から利用される検索サイトやショッピングサイト等の登録が多い gTLD がターゲットとなりやすく，悪性ドメイン名を検知するための様々な手法が研究されている．

悪性ドメイン名に関する代表的な研究には，フィッシング攻撃対象としたもの[2][3]，ホモグラフ攻撃対象としたもの[4]，廃止されたドメイン名を即時に再登録するドロップキャッチを対象としたもの[5][6]がある．また，悪性ドメイン名の自動生成（Domain Generation Algorithm; DGA）を検知のため機械学習を対象としたもの[7][8][9]がある．

(2) ccTLD の悪性ドメイン名検知に関する研究

gTLD と比較して登録要件が厳しい ccTLD では相対的に悪性ドメイン名は少ないと言われているが，ドメイン名総登録数が数百万件を超える ccTLD ではその数は無視できなくなる．歴史が長く登録数の多い欧州の ccTLD でも悪性ドメイン名を早期検知するための手法が研究されている．

代表的なものには，オランダの ccTLD レジストリである SIDN[b]によるリアルタイムトラフィック分析[10][11]がある．

3. データセット

本研究では，2016 年 10 月 1 日から 2018 年 3 月 31 日までに新規登録された JP ドメイン名[c]のうち，汎用 ASCII ドメイン名を対象とし，以下の情報を用いて分析を行った．

・登録文字列（汎用 ASCII ドメイン名の第 2 レベル）
・登録年月日
・1 年後の登録更新有無[d]

本研究での分析実施時点において，2018 年 4 月 1 日以降に新規登録された JP ドメイン名は更新有無の情報がないため対象としていない．表 1 に対象期間における月毎の更新率実績値を示す．

表 1 月毎の更新率実績値

年	月	更新率実績値
2016	10	0.7536
	11	0.7608
	12	0.7570
2017	1	0.7644
	2	0.7733
	3	0.7574
	4	0.7377
	5	0.7513
	6	0.7672
	7	0.7428
	8	0.7467
	9	0.7565
	10	0.7453
	11	0.7273
	12	0.7162
2018	1	0.6537
	2	0.7413
	3	0.7521

4. 分析方法

前章で述べたデータセットを用いて，新規登録ドメイン名の更新率を予測した．予測方法としては，統計情報を用いるものと，機械学習を用いるものを使用した．

いずれの方法も予測は月単位で行っており，予測対象月の直前 12 ヶ月間のデータを学習データとして用いた．

4.1 統計的手法による更新率予測

統計的手法による更新率予測は，対象月の直前 12 ヶ月間の更新率実績値を用いて最小二乗法（線形近似）で計算する方法と，移動平均（3 区間）で計算する方法の 2 つの方法を用いた．

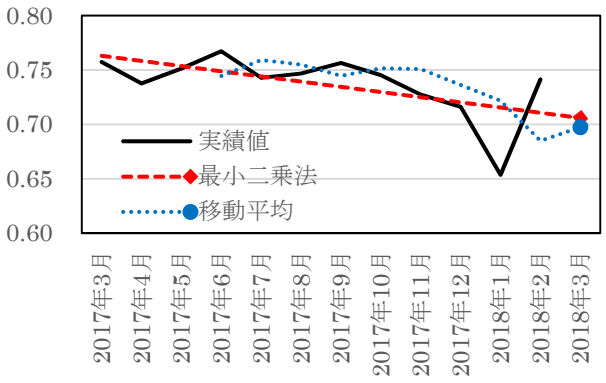

図 1 最小二乗法および移動平均による予測結果
（2018 年 3 月分）

新処理は月単位で行われる．

b) Stichting Internet Domeinregistratie Nederland, https://www.sidn.nl/en
c) JP ドメイン名の種類，https://jprs.jp/about/jp-dom/
d) JP ドメイン名の有効期限は登録年月日の 1 年後の月末であり，登録更

図 1 に最小二乗法および移動平均による予測結果（2018 年 3 月分）を例として示す.

4.2 ドメイン名の文字列のみから得られる特徴量

経験的に，ドメイン名の更新率は登録文字列（ドメイン名の第 2 レベルの文字列，以降ラベル）の特徴に応じて差異があることが分かっている．例えば，企業・行政等による施策や何らかの試験に基づくドメイン名は期間限定で使用されることが多く，そのため更新率は低いが，そのようなドメイン名の特徴として年号や通番を示す数字が含まれることが多い．また，ランダムな文字列のドメイン名も更新率は低いが,そのようなドメイン名は子音が連続したり，数字が単独で現れる等の特徴がある.

汎用 ASCII ドメイン名のラベルに使用可能な文字は英字（大文字小文字区別なしの 26 文字），数字（10 文字），ハイフン（1 文字）の計 37 文字である．本研究では，それら文字がラベル中で出現する位置や，並びの状態が更新率にどれほど寄与するかを評価可能とするため，49 の特徴量を定義した．表 2 にドメイン名の文字列のみから得られる特徴量を示す.

表 2 ドメイン名の文字列のみから得られる特徴量

名称	説　明	値
■ラベルそのものに関する特徴量 (1)		
LEN	文字数	int
■数字に関する特徴量 (13)		
DIN	数字の数	int
SDN	単独数字の数	int
NUN	数字列の数	int
NUML	最長の数字列の長さ	int
SDNUN	単独数字の数と数字列の数の和	int
DIMI	数字間の最長距離	int
NUPW	全体が 1 つの数字列か	bool
SDPB	先頭が単独数字か	bool
SDPM	途中が単独数字か	bool
SDPE	末尾が単独数字か	bool
NUPB	先頭が数字列か	bool
NUPM	途中が数字列か	bool
NUPE	末尾が数字列か	bool
■アルファベットに関する特徴量 (15)		
LTN	英字の数	int
SLN	単独英字の数	int
STN	英字列の数	int
STML	最長の英字列の長さ	int
VWN	母音の数	int
CSN	子音の数	int
VWMI	母音間の最長距離	int
VWLTR	母音数と英字の数の比	real
STPW	全体が 1 つの英字列か	bool
SLPB	先頭が単独英字か	bool
SLPM	途中が単独英字か	bool
SLPE	末尾が単独英字か	bool
STPB	先頭が英字列か	bool
STPM	途中が英字列か	bool
STPE	末尾が英字列か	bool
■ハイフンに関する特徴量 (20)		
HYN	ハイフンの数	int
SHN	単独ハイフンの数	int
HSN	ハイフン列の数	int
HSML	最長のハイフン列の長さ	int
SHPASD	単独数字の後が単独ハイフンか	bool
SHPBSD	単独数字の前が単独ハイフンか	bool
SHPASL	単独英字の後が単独ハイフンか	bool
SHPBSL	単独英字の前が単独ハイフンか	bool
SHPANU	数字列の後が単独ハイフンか	bool
SHPBNU	数字列の前が単独ハイフンか	bool
SHPAST	英字列の後が単独ハイフンか	bool
SHPBST	英字列の前が単独ハイフンか	bool
HSPASD	単独数字の後がハイフン列か	bool
HSPBSD	単独数字の前がハイフン列か	bool
HSPASL	単独英字の後がハイフン列か	bool
HSPBSL	単独英字の前がハイフン列か	bool
HSPANU	数字列の後がハイフン列か	bool
HSPBNU	数字列の前がハイフン列か	bool
HSPAST	英字列の後がハイフン列か	bool
HSPBST	英字列の前がハイフン列か	bool

4.3 機械学習による更新率予測

前章で述べたデータセット中の全ドメイン名の各ラベルに，前節で定義した 49 特徴量の値を対応付け，特徴量データとした.

更新率予測対象月の直前 12 ヶ月間に登録された全ドメイン名の各ラベルの特徴量データと更新有無情報を学習データとし，機械学習の方式としては統計解析ソフトウェア R[12]の rpart モジュール[13]を使用して決定木分析を行った[e].

図 2 に決定木分析結果（2018 年 3 月分）を例として示す.

予測対象月に新規登録された全ドメイン名の各ラベルに対し，決定木分析結果を適用してラベル個別の更新確率を計算した．機械学習による対象月の更新率予測値は，ラベル個別の更新確率の平均値とした.

e) 複雑パラメータ (complexity parameter; cp) は 0.0005 に固定した

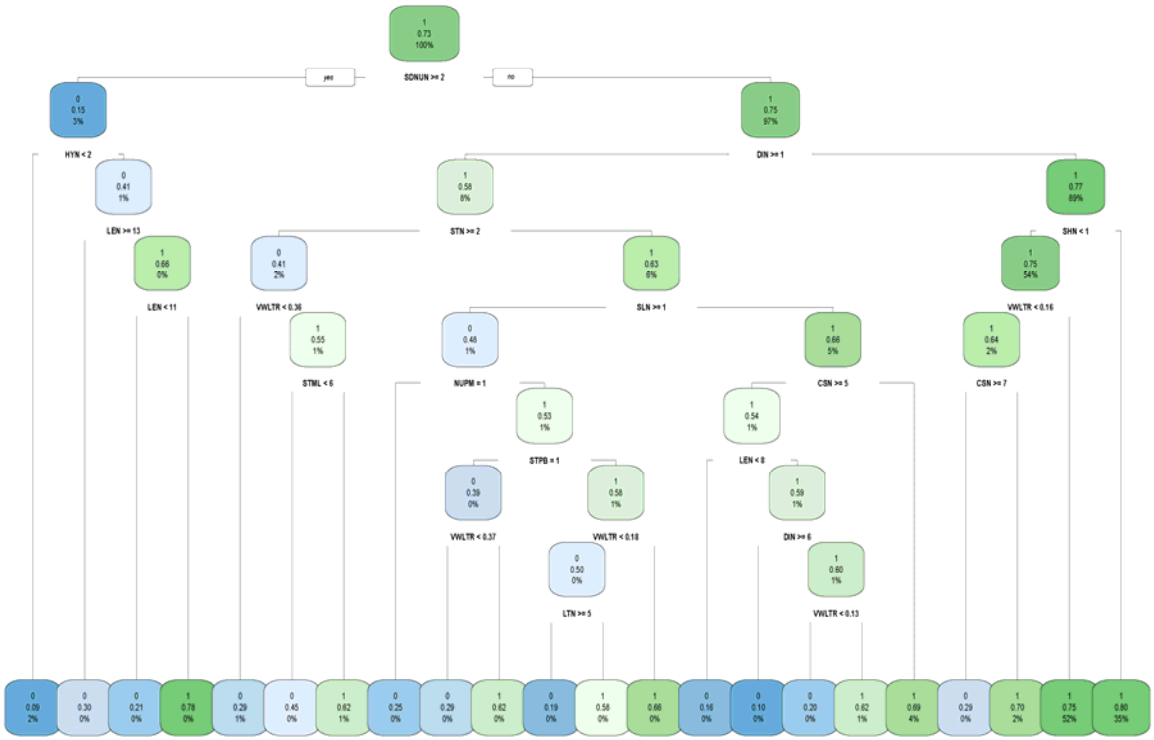

図 2 決定木分析結果（2018 年 3 月分）

5. 結果

統計的手法と提案手法（ドメイン名の文字列のみから得られる特徴量を使った機械学習）による予測の結果を比較した.

5.1 統計的手法と機械学習による更新率予測値

2017 年 10 月から 2018 年 3 月の 6 ヶ月について，手法毎に更新率予測値と，実績値からの乖離を計算した．表 3 に手法毎の更新率予測値と実績値からの乖離を示す.

5.2 結果比較

更新率予測を行った 6 ヶ月（2017 年 10 月から 2018 年 3 月の計 6 回）に関し，月毎の実績値と各予測手法による更新率予測値との乖離（絶対値）を平均すると，最小二乗法，移動平均，機械学習のそれぞれについて約 3.4%, 3.8%, 2.2%となり，機械学習による更新率予測値が最も実績値に近く，最小二乗法に比べ 1.3%，移動平均に比べ 1.7%程度予測精度が向上している.

また，月毎の更新率予測値における乖離の最大値（絶対値）に関しても，最小二乗法，移動平均，機械学習のそれぞれについて約 7.3%，6.8%，5.1%[f]であり，機械学習による更新率予測値が最も実績値に近い．6 回の予測のうち，実績値に最も近い更新率予測値となった手法は，機械学習が 4 回，最小二乗法および移動平均がそれぞれ 1 回ずつである.

表 3 手法毎の更新率予測値（上段）と
実績値からの乖離（下段）

年	月	実績値	更新率予測値		
			最小二乗法	移動平均	機械学習
2017	10	0.7453	0.7497 (0.0044)	0.7516 (0.0063)	0.7542 (0.0089)
	11	0.7273	0.7458 (0.0185)	0.7509 (0.0237)	0.7436 (0.0163)
	12	0.7162	0.7377 (0.0215)	0.7363 (0.0201)	0.7399 (0.0237)
2018	1	0.6537	0.7267 (0.0730)	0.7217 (0.0680)	0.7045 (0.0508)
	2	0.7413	0.6983 (0.0430)	0.6850 (0.0563)	0.7316 (0.0097)
	3	0.7521	0.7060 (0.0461)	0.6975 (0.0546)	0.7317 (0.0204)
平均		0.7227	0.7274 (0.0344)	0.7238 (0.0382)	0.7343 (0.0216)

※実績値に最も近い更新率予測値を色付けした

f) 各手法ともに，2018 年 1 月の更新率予測値と実績値の乖離

6. 議論

提案手法（ドメイン名の文字列のみから得られる特徴量を使った機械学習）による更新率予測精度向上の理由，および今後の課題について議論する．

6.1 更新率予測精度向上の理由考察

ドメイン名は利用者に認知されることが重要であるため，ブランド名や組織名等の固有名詞，広く一般に使われている単語や用語の組み合わせで構成されることが多く，また，継続して使用されるため一定の更新率が見込まれる．

従来その傾向は安定しており，更新率は統計的に十分予測が可能であった．しかし，企業・行政等による様々な施策や社会的イベント等により期間限定で利用されるドメイン名の新規登録数が一時的に急増すると，統計的手法は予測精度が低下する．

2017 年 11 月から 2018 年 1 月に見られるような月毎の更新率に大きな変動が生じた場合，統計的手法では更新率予測値の実績値からの乖離が変動後も一定期間継続するが，提案手法では各ドメイン名の更新率を個別に予測するため，変動後の更新率予測値に影響がない．このため，提案手法による更新率予測は統計的手法に比べ予測精度を維持でき，全体としての予測精度が向上したと考えられる．

6.2 今後の課題

本論文は，機械学習によるドメイン名の登録更新率予測の端緒に過ぎず，以下に示すような課題がある．

(1) 予測時期の早期化

本論文では，対象月の更新率予測を実際の更新前月に実施しているが，更新率予測をレジストリ維持計画に反映させる上ではより早期の実施が求められる．予測実施の早期化のためには，予測精度を維持しつつ，機械学習データの選択期間，分析アルゴリズム，複雑パラメータ設定値等を調整する必要がある．

(2) 傾向変化への対応

企業・行政等による施策の実施やイベントが発生する機会や，それらが継続する期間は多様である．一方，それらは新規登録ドメイン名の文字列的傾向に大きな変動をもたらすことがある．そのような変動を検知する新たな特徴量を分析する必要がある．

予備的な調査では，文字の出現頻度を求める N-Gram 手法を用いた分析によって，傾向の変動を捉えられる可能性を見出している．

(3) 他事業者等への適用性

ドメイン名の文字列に注目して特徴量を抽出するという方法は汎用的であり，管理しているドメイン名の網羅的なリストを持つ他のドメイン事業者（レジストリ・レジストラ）やドメイン名登録者・研究者にも適用可能であると考えられる[g].

(4) ドメイン名の文字列以外の情報による影響

ドメイン名登録者情報や申請を取り次いだレジストラは，ドメイン名新規登録時に即時に得られるドメイン名文字列以外の情報であり，更新率予測の精度を向上させる特徴量である可能性がある．

また，DNS サーバ設定状況，Web サイト開設状況，サーバ証明書利用状況等の情報は，ドメイン名新規登録後に観測される情報であり経時変化するものの，更新率予測の精度を向上させる特徴量である可能性がある．

(5) 2 年目以降の更新率予測

JP ドメイン名全体では登録更新の回数が多くなるにつれ，次回の更新率が高くなる傾向がある．2 年目以降の更新率予測についても，本論文で定義した特徴量に更新回数等を新たな特徴量として追加することで，予測精度が向上する可能性がある．

(6) ドメイン名文字列の語彙分析

本論文では，ドメイン名の文字列に注目しているが，主に文字が文字列中で出現する位置や並びの状態を分析しており，どのような単語や用語から構成されているかについては分析していない．ドメイン名文字列に単語・用語が含まれているかを，辞書を用いて語彙分析し，新たな特徴量として追加することで，予測精度が向上する可能性がある．

語彙分析に使用する辞書としては，英語由来のもののみでなく，日本語由来（ローマ字表記）のものも考慮する必要がある．

(7) 国際化ドメイン名への対応

本論文では ASCII ドメイン名を対象にしているが，非 ASCII 文字を使用する国際化ドメイン名（Internationalized Domain Name; IDN）も普及しており，考慮する必要がある．IDN ではより言語的な特性が現れるため，前項の語彙分析を含む IDN に特化した特徴量の分析が必要である．

g) JPRS は gTLD レジストラでもあるため，レジストラとして本手法を適用し，更新率予測の精度を検証できる

7. まとめ

ドメイン名の登録文字列のみから得られる情報を機械学習することで，登録更新率を予測する手法を考案した．また，提案手法による更新率予測は統計的手法による更新率予測と比べ, 予測精度が向上することを確認した. さらに，月毎の更新率に大きな変動が生じた場合，統計的手法では更新率予測値の実績値からの乖離が変動後も一定期間継続するが，提案手法では各ドメイン名の更新率を個別に予測するため，変動後の更新率予測値に影響がないことを確認した．

今後，議論で述べた課題について研究を継続する．

謝辞 本研究の着想を与えてくれた株式会社日本レジストリサービス東田幸樹社長に感謝の意を表する．また，本研究にアドバイスをいただいた早稲田大学基幹理工学部情報通信学科森達哉教授に感謝の意を表する．

参考文献

[1] CENTRstats Global TLD Report 2019/1. https://centr.org/library/library/statistics-report/centrstats-global-tld-report-2019-1.html
[2] Hossein Shirazi et al., “Kn0w Thy Doma1n Name”: Unbiased Phishing Detection Using Domain Name Based Features. 23nd ACM on Symposium on Access Control Models and Technologies, 2018
[3] Rashid Tahir et al., It's All in the Name: Why Some URLs are More Vulnerable to Typosquatting, 2018, IEEE Conference on Computer Communications
[4] Jonathan Woodbridge et al., Detecting Homoglyph Attacks with a Siamese Neural Network, 2018, 1st Deep Learning and Security Workshop, co-located with the 39th IEEE Symposium on Security and Privacy
[5] Chaz Lever et al., Domain-Z: 28 Registrations Later Measuring the Exploitation of Residual Trust in Domains, 2016, IEEE Symposium on Security and Privacy
[6] Najmeh Miramirkhani, Panning for gold.com: Understanding the Dynamics of Domain Dropcatching, 2018, World Wide Web Conference 2018
[7] Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique, 2018, Cybernetics and Information Technologies, Volume 18: Issue 1
[8] Michael Weber et al., Unsupervised Clustering for Identification of Malicious Domain Campaigns, 2018, The 13th ACM ASIA Conference on Computer and Communications Security
[9] Hong Zhao et al., Malicious Domain Names Detection Algorithm Based on N-Gram, 2019, Journal of Computer Networks and Communications Volume 2019
[10] Giovane C. M. Moura et al., nDEWS: a New Domains Early Warning System for TLDs, IEEE/IFIP International Workshop on Analytics for Network and Service Management, 2016
[11] Giovane C. M. Moura et al., Domain names abuse and TLDs: from monetization towards mitigation, 3rd IEEE/IFIP Workshop on Security for Emerging Distributed Network Technologies, 2017
[12] The R Project for Statistical Computing. https://www.r-project.org/
[13] rpart: Recursive Partitioning and Regression Trees. https://cran.r-project.org/web/packages/rpart/index.html