Vol.2018-CSEC-82 No.34
Vol.2018-SPT-29 No.34
2018/7/26

日本人が付けるパスワードの特性調査と他国データとの比較

愛 野乃子[1]　金岡 晃[1]

概要：多くのパスワードを含むデータセットを対象にした分析により、効率的なパスワード推測や、それらの結果をもとにしたパスワード設定の強化策などが考えられるようになった。しかしこれらのもとになったデータセットは英語圏ユーザが付けたパスワードデータセットであり、他の言語圏のユーザが付けたパスワードデータセットでは同様の結果にならない可能性がある。Li らは中国語データセットでこのことを明らかにし、さらなる調査や分析が必要なことを示した。本研究では日本人が付けたパスワードデータセットに焦点を当て、同様に調査を行うことで日本語パスワードの特徴を明らかにすることを目的とし分析を行った。その結果、中国語とも英語とも異なった特徴を持つことがわかり、さらに詳細な分析を行うことで単語の利用などに特有な要素があることがわかった。

1. はじめに

RockYou.com のパスワード流出事件があった 2009 年以降、パスワード推測の研究は大きな発展を遂げた。RockYou データセットに含まれる 3100 万もの平文パスワードを分析することにより利用者のパスワード構成の統計的特徴が判明し、その特徴を利用した推測手法はランダムな選択よりも効果的に推測を成功させた [1]。

パスワード推測の効率化はさまざまなパスワード強化方策に対する評価指標としても利用され、パスワードに関わる研究の発展にも寄与した。一方で、RockYou データセットや他のデータセットで提供されるデータがパスワードの真の情報量を反映しているかという問題についてはあまり深く議論されてこなかった。Li らは中国人ユーザが付けるパスワードは RockYou や他のデータセットとは異なる統計的特徴があることを示し [2]、ユーザの属性により統計的な特徴が異なることがあることを明らかにした。森らはデータセット中の日本人ユーザと思われるパスワードから日本人ユーザが付けるパスワードの特徴解析を行った [3]。

本研究では、森らの研究と同じく、日本人ユーザが付けるパスワードの特徴分析に焦点をあて、異なるアプローチと異なるデータセットで分析を行う。

2. データセット

2.1 14 億パスワードデータセット

2017 年、Casal により 14 億個のメールアドレス付きの平文パスワードデータがインターネット上に公開されていることが伝えられた [4]。それらは特殊な技術を用いることなく取得可能であったため、本研究ではそのデータセットを利用して調査をした。日本人が付けたパスワードの特徴調査に用いるために、そのデータセットのうち、メールアドレスが”.jp”で終了するメールアドレスのパスワードだけを抽出して分析に用いた。ここで、”.jp”のメールアドレスを持つパスワードは日本人が付けたパスワードである可能性が非常に高い、という前提を置いた。

抽出されたパスワードは 6,195,726 個であった。本論文ではこのデータを”1.4B-jp”と呼ぶことにする。

2.2 クラウドソーシングサービスを利用した収集

1.4B-jp はある前提を置いたデータセットであるため、実際に日本人が付けたパスワードではないパスワードが混じっている可能性もある。そこで、クラウドソーシングサービスを利用してパスワード収集を行い、そのデータセットも分析に用いることとした。

2.2.1 利用したサービスと期間、報酬

2017 年 11 月 22 日（水）〜12 月 4 日（月）、ランサーズにて被験者を募集し、アンケートに回答してもらうとともに、研究室で準備したサーバ上での作業をしてもらうことでパスワードの収集を行った。タスク終了の目安時間は 10 分で報酬は 150 円である。目安時間の 10 分については、実際に研究室で他の学生に模擬実験を行ってもらい、計測した作業時間から妥当であると考えた値である。報酬については、日本の 1 時間あたりの最低賃金を基準とした額である。本学がある千葉県の最低賃金が、2017 年 10 月 1 日現在時給 868 円であり、目安時間が 10 分であることから報

[1] 東邦大学
Toho University, Miyama 2-2-1, Funabashi, Chiba, Japan

酬は 150 円 (時給 900 円）が妥当であると考えた。

2.2.2 実験目的

本来の「特徴分析のためのパスワード収集」と言う目的を伝えると、被験者がパスワード作成に影響を及ぼしてしまう恐れがあったため、なるべく自然にパスワードを作成してもらえるよう、別の目的を伝え実験を行った。別の目的とは、ユーザがウェブサイトやアプリなど複数のサービスにアカウントを持つ現代において、どれだけ複数のアカウントを管理できているかを調査したいと言う、「利用者の記憶の調査」である。そのため被験者には最初、「利用者による記憶を測るために、今回 10 種類の仮ウェブサイトにユーザ登録してアカウントを作ってもらう。日程をあけてもう 1 度実験を行い、どれだけ前回登録したパスワードを覚えていられるか調査する。」と言う説明をし、パスワードを作成してもらった。そして最後に真の目的を伝え、同意を得た人のデータのみサーバに取得した。

2.2.3 実験の詳細

実験の流れは以下の通りである。

(1) ランサーズのページから今回実験のために作成したウェブサイトにアクセスする
(2) 仮の目的を提示し、同意を得る
(3) 実験のために作成したウェブサイトの仮サービス説明画面に遷移する
(4) ユーザ登録画面に遷移し、ユーザ登録を行う
※ 3〜4 を 10 回繰り返す
(5) 真の目的を提示し、同意を得る（同意を得た人のみデータを取得）
※実験で作成したウェブサイトから入力してもらったデータは研究室のサーバに取得
(6) アンケートに回答する

以上の流れにより、1 人 10 個のパスワードを作成してもらう。最終的には 500 人に実験を行ってもらったため、全部で 5000 個のパスワードを収集した。

データは、ランサーズ上にて被験者の性別、年齢、利用端末 (パソコン、スマートフォン、その他)、職業、最終学歴を取得し、研究室のサーバに送信元 IP アドレス、利用ブラウザ情報、送信日時、10 個のパスワード、アンケート回答を取得した。そしてランサーズ上で取得した情報とサーバで取得した情報のデータ照合を行った結果、途中で入力が途切れているものや、4 文字未満のパスワードが入力されているものなど、実験には使えないデータがいくつか存在した。そのため最終的には、正常に入力されていた 448 人のパスワード 4384 個を分析に使用した。

2.2.4 生命倫理審査委員会の承認

本実験は学内の生命倫理審査委員会に承認を得て行った。

3. 分析方法

3.1 基本分析

基本的な分析として、以下の分析を行う。

- 出現回数が最も多い上位 5 つのパスワードの調査
- パスワード構成の率
  - アルファベットだけで構成されているパスワード
  - 数字だけで構成されているパスワード
  - 記号だけで構成されているパスワード
  - アルファベットと数字で構成されているパスワード
  - アルファベットと記号で構成されているパスワード
  - 数字と記号で構成されているパスワード
  - アルファベットと数字と記号で構成されているパスワード
- パスワード構造

パスワード構造とは、アルファベットを L、数字を D、記号を S としたときに、パスワードがどういった組み合わせて構成されることが多いかを調べるものである。たとえばパスワードが aa123+B のとき、その構造は LLDDDSL となる。

3.2 指標による分析

パスワードに関するデータセットを分析する指標は複数提案されている。本研究で利用した指標を解説する。

Bonneau は、Renyi エントロピー H_n の中で H_0 と H_∞ に着目した。H_0 はパスワード種類数の対数であり、種類数の大きさを示す値となっている。また H_∞ は最も出現率の高いパスワードの対数を取り符号反転させたものである。Boztas により提案された λ_n はデータセット中のパスワード i 出現の確率を p_i とし、それを出現率の降順で並び変えたときの上位 n パスワードの出現率の和を示したものであり、$\lambda_n = \Sigma_{i=1}^{n} p_i$ で示される [5]。Bonneau はさらに λ_n を発展させた $\tilde{\lambda}_n = lg(\frac{n}{\lambda_n})$ を提案した [6]。この指標は Bonneau の論文や Li の論文 [2]、森らの論文 [3] で採用されている。

Pliam により提案された μ_α は、0-1 の値を α について、α の割合だけ推測攻撃を成功させるために必要な推測数を評価する指標であり、パスワードの出現確率を上位から足していき、それが α を超える段階になったときの順位を指す [7]。Bonneu はこの μ_α と λ_n を組み合わせて、新たな指標 G_α を提案した。この値は成功率 α で推測攻撃が成功することが期待できるような推測数を示すものである。さらに G_α を拡張した $\tilde{G}_\alpha$ も提案した。

3.3 キーボードパターン分析

Li らは、パスワードの分析に際し、キーボードパターンに着目し、隣接するキーを使ったパスワードのうち、同一

列に属するものを *Same Row*、上下の異なる列に移動するものを *Zig Zag*、それ以外のものを *Snake* とし、その出現回数を調べた。本研究でもその回数を分析する。

3.4 単語利用分析

Li らは、パスワード分析に際し、PinYin の辞書と英語の辞書を用いて、パスワードに含まれる単語の特徴を分析した。本研究でも日本語の単語と英語の単語がどう使われているかを分析する。

辞書として、日本語の平仮名綴りの辞書を利用した。本研究で利用した辞書は私立 PDD 図書館の百科事典 [8] である。2 文字以上の単語 78,974 個を訓令式ローマ字に変換して利用した。英語の辞書は Github で公開されている辞書 [9] を利用し、3 文字以上の単語 369,646 個を利用した。

訓令式ローマ字に規定のない特殊文字に関しては以下のように独自で設定した。

- アポストロフィーの定義に関しては無視
- ヴぁ、ヴぃ、ヴ、ヴぇ、ヴぉ → va、vi、vu、ve、vo
- ヴゃ、ヴゅ、ヴょ → vya、vyu、vyo
- ぁ、ぃ、ぅ、ぇ、ぉ → xa、xi、xu、xe、xo
- 伸ばし棒 ー → 無、- (ハイフン)、^ (サーカムフレックス) の 3 パターンで調査
- っ の後に母音が来る場合→子音ならば重ねるが母音の場合は無視

4. 分析結果

4.1 基礎分析結果

1.4B-jp データセットとランサーズデータセットを用いた分析結果を示す。

4.1.1 最もよく使われるパスワード

最もよく使われるパスワードは、1.4B-jp では”123456”、ランサーズでは”1234”であった。その他、上位 5 位までの結果を他のデータセットでの結果と合わせて表 1 に示す。近い結果が表れている一方で、ランサーズデータセットは同じ日本語データセットの他の 2 者には相関関係がみられない様子がわかる。

4.1.2 パスワード構成

パスワード構成の分析結果を表 2 に示す。Li らの研究では、英語パスワードデータでは文字のみで構成されるパスワードが多い一方で中国語パスワードデータは数値のみで構成されるパスワードが多数を占めていることが示されていたが、日本語パスワードデータではそのいずれの特徴も持たず、最も多い構成は文字と数字で構成されるパスワードであった。これらは 1.4B-jp とランサーズのデータセットのいずれも 50%を超える値になっており、他の言語のデータセットにはない特徴が表れていることがわかる。

4.1.3 パスワード構造

1.4B-jp とランサーズデータセットのパスワード構造の分析結果を表 3 に示す。ランサーズデータセットの最頻出は DDDD であり 1.4B-jp には上位に位置していないものがあるが、その他の構造は 1.4B-jp との類似性が見える。

最頻出パスワード構造の違いを各データセットごとに見たものを表 4 に示す。Li らの研究では、英語パスワードは文字のみの構成が多いことに対して、中国語パスワードは数値のみの構成が多いことが示されたが、表 3 にも示した通り、日本語パスワードはこれらとは違う特徴を持っていることがわかる。

4.2 指標による分析

各種指標による分析結果を表 5 に示す。H_∞ と λ_{10} は日本語データセット間で同様の傾向を示す一方、$\tilde{G}_{0.25}\tilde{G}_{0.5}$ では頻出パスワードや頻出パスワード構造と同じく、ランサーズデータだけが離れた特徴を持っていることがわかる。

他の言語との指標値の違いを見ると、各言語のデータセット間で強く特徴が分かれていることが見える。たとえば中国語パスワードでは λ_{10} が大きな値になる一方、英語と日本語のデータセットはそういった特徴は持っていないことがわかる。

4.3 キーボードパターン分析結果

キーボードパターンの分析結果を表 6 に示す。*Zig Zag* については大差がないものの、*Same Row* と *Snake* の率が他の言語パスワードに比べて高いことがわかる。

4.4 単語利用分析結果

単語利用分析の結果得られた頻出単語を表 7 に示す。

日本語辞書で分析した結果、最も頻度が高いパスワードは”sakura”であった。また上位 10 個のパスワードを見ると、人名、とくに男性の名前に利用される言葉が上位にあることがわかる。上位 50 個のパスワードのうち、人名と思われるものは 15 個があった。これらの特徴は他の研究結果からは得られておらず、日本語パスワードに特有の情報である可能性もある。

さらに、上位 10 個のパスワードにある”nekoneko”のように、同じ表現を 2 回以上繰り返すパスワードも上位に多くある。同じく上位 50 個のパスワードのうちに同じ表現を 2 回以上繰り返すパスワードを数えると、11 個あった。”sasasa”が 11 位、”nikoniko”が 16 位、”kenken”が 18 位など、こういった特徴も日本語パスワード特有の可能性もある。

5. 考察と今後の課題

5.1 クラウドソーシングサービスの利用

ランサーズで収集したデータは、1.4B-jp データセットでの分析結果や森らの結果とは異なる特徴を持つことがわかった。データセットごとに特徴が変わることがあること

表 1　最もよく使われるパスワード上位 5 つ

順位	1.4B-jp	ランサーズ	森ら [3]	RockYou	Li ら [2]
1	123456 (0.22%)	1234 (0.27%)	! (0.27%)	123456 (0.88%)	123456 (2.17%)
2	password1 (0.14%)	5678 (0.23%)	123456 (0.15%)	12345 (0.24%)	123456789 (0.65%)
3	123456789 (0.09%)	1111 (0.21%)	12345 (0.07%)	123456789 (0.23%)	111111 (0.59%)
4	asdfghjk (0.08%)	7777 (0.21%)	password (0.07%)	password (0.18%)	12345678 (0.39%)
5	12345678 0.08%)	5555 (0.16%)	123456789 (0.06%)	iloveyou (0.15%)	000000 (0.34%)

表 2　各データセットにおけるパスワード構成の割合

データセット	数値のみ	文字のみ（小文字のみ）	文字＋数字（小文字＋数字）	文字＋記号（小文字＋記号）	記号＋数字	文字＋数字＋記号（小文字＋数字＋記号）
1.4B-jp	17.11%	28.91% (28.15%)	50.88% (50.37%)	1.23% (1.22%)	0.03%	1.81 (1.72%)
ランサーズ	18.45%	25.66% (25.14%)	51.80% (50.98%)	0.64% (0.62%)	0.00%	3.44% (3.22%)
RockYou[2]	15.93%	44.04% (41.68%)	36.22% (33.17%)	1.91% (1.64%)	0.16%	1.71% (1.44%)
CSDN[2]	45.06%	12.39% (11.68%)	39.02% (35.60%)	0.50% (0.42%)	0.61%	2.39% (2.04%)
Tianya[2]	64.56%	10.20% (9.89%)	23.12% (21.27%)	0.25% (0.22%)	0.71%	1.14% (1.01%)

表 3　頻出パスワード構造

順位	1.4B-jp	ランサーズ
1	LLLLLLLL (11.92%)	DDDD (10.24%)
2	DDDDDDDD (6.65%)	LLLLLL (5.11%)
3	LLLLDDDD (5.87%)	LLLLLLLL (4.43%)
4	LLLLLL (5.81%)	LLLLDDDD (3.81%)
5	DDDDDD (4.84%)	LLLLLLL (3.03%)

表 4　最頻出パスワード構造のデータセットによる違い

データセット	構造	出現率
1.4B-jp	LLLLLLLL	11.92%
ランサーズ	DDDD	10.24%
CSDN[2]	DDDDDDDD	21.50%
Tianya[2]	DDDDDD	30.10%
RockYou[2]	LLLLLL	5.40%
Yahoo[2]	LLLLLL	9.19%

表 5　各種指標値のデータセットによる違い

データセット	H_∞	λ_{10}	$\tilde{G}_{0.25}$	$\tilde{G}_{0.5}$
1.4B-jp	8.80	1.08%	18.49	20.63
ランサーズ	8.51	1.80%	11.73	11.89
森ら [3]	8.53	0.86%	16.7	17.7
CSDN[2]	4.77	10.44%	15.60	20.30
Tianya[2]	4.55	8.11%	14.67	19.11
RockYou[2]	6.81	2.05%	15.88	19.80
Yahoo[2]	8.05	1.01%	16.31	17.68

表 6　キーボードパターン

	日本語 (1.4B-jp)	日本語 (ランサーズ)	中国語 [2]	英語 [2]
Same Row	1.28%	3.72%	0.55%	0.25%
Zig Zag	0.29%	0.23%	0.26%	0.06%
Snake	1.73%	4.43%	0.27%	0.08%

は Bonneau の研究や Li らの研究でも示されているが、ランサーズのデータセットに関してはその違いが顕著である。「評価として利用」という収集の前提が、被験者が付けるパスワードの特徴に影響を及ぼしてしまったことが考えられる。

表 7　1.4B-jp における頻出単語

	日本語	英語
1	sakura (0.044%)	password (0.078%)
2	takahiro (0.019%)	sakura (0.044%)
3	hiroki (0.015%)	aaaaaa (0.027%)
4	masahiro (0.014%)	excite (0.020%)
5	makoto (0.014%)	takahiro (0.019%)
6	doraemon (0.014%)	hiroki (0.015%)
7	kazuki (0.014%)	marahiro (0.014%)
8	takuya (0.014%)	soccer (0.014%)
9	masaki (0.013%)	takayuki (0.013%)
10	nekoneko (0.013%)	baseball (0.013%)

パスワードデータ収集に関する生態学的妥当性（Ecological Validity）については Fahl らが指摘しており [10]、あらためてその難しさが見えた結果となった。

5.2　推測攻撃に与える影響

もし攻撃対象のパスワードが日本人によりつけられたことが判明している場合、推測攻撃の学習に用いるデータセットを日本語データセットにすることでより効率的な攻撃が可能になる可能性がある。一方で、さまざまな指標で比較をした場合、日本人が付けたパスワードの特徴は英語パスワードの特徴からは大きく異なっておらず、英文データセットで学習したデータでの推測攻撃でも、良好な結果が得られることも考え得る。もちろん日本語パスワードのデータセットで学習させることが最も良い結果が得られることが考えられるが、一方で学習自体に係るコストを考えることも重要になってくるものと思われる。

5.3　日本語特有のパスワード

人名や繰り返し表現など、これまでの研究では焦点が当てられてこなかった要素が本研究で抽出された。今後はこれらに注目して分析することでより日本語特有の要素が明

らかになる可能性がある。繰り返し表現は日本語の人名や単語をローマ字に変換したときに文字数が多くなることから、シンプルな表現を繰り返すことで覚えやすさと文字の数のバランスを取っていることも考えらえる。その前提に立つと、たとえば日本語では 5 音節以上の単語を 4 音節に略すことが多く行われている（例：ナショナルジオグラフィック→ナショジオ、パーソナルコンピューター→パソコン）ため、そういった略語の点に着目した分析も考えられる。

6. まとめ

本研究では、日本人が付けるパスワードの特徴を知るために、大量のパスワードデータセットから複数の視点で分析を行い、他の日本語パスワード分析研究との比較や、英語や中国語パスワードの分析研究との比較を行った。また、漏えいパスワードだけではなく、本研究のためにクラウドソーシングを利用して収集したパスワードデータセットも分析に利用し、その違いを考察した。

その結果、日本語パスワードは英語や中国語のパスワードとは異なる特徴を持っていることがあらためて明らかになり、その中でも人名を利用したパスワードや単純な言葉の繰り返し表現が使われることが多いことが明らかになった。また、クラウドソーシングにより収集されたパスワードは他の日本語パスワードデータセットとな異なる特徴を持つことがわかり、実験環境がデータセットの特徴に影響を及ぼしていることが考えられることがわかった。

ユーザビリティとセキュリティの研究では、対象となるデータセットを変えて分析をするレプリケーション研究が推進されているが、本研究もその 1 つである。データセットが異なれば結果が変わることが示され、レプリケーション研究の重要性を示す 1 つの例になったと考えられる。

参考文献

[1] Weir, M., Aggarwal, S., Collins, M. and Stern, H.: Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords, *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, New York, NY, USA, ACM, pp. 162–175 (online), DOI: 10.1145/1866307.1866327 (2010).

[2] Li, Z., Han, W. and Xu, W.: A Large-Scale Empirical Analysis of Chinese Web Passwords, *23rd USENIX Security Symposium (USENIX Security 14)*, San Diego, CA, USENIX Association, pp. 559–574 (online), available from ⟨https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/li_zhigong⟩ (2014).

[3] 森　啓華，森　達哉：文化の差異がパスワード生成に与える影響の考察，2018 年暗号と情報セキュリティシンポジウム (SCIS 2018) (2018).

[4] Casal, J.: 1.4 Billion Clear Text Credentials Discovered in a Single Database.

[5] Boztas, S.: Entropies, guessing, and cryptography (1999).

[6] Bonneau, J.: The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords, *2012 IEEE Symposium on Security and Privacy*, pp. 538–552 (online), DOI: 10.1109/SP.2012.49 (2012).

[7] Pliam, J. O.: On the incomparability of entropy and marginal guesswork in brute-force attacks, *International Conference on Cryptology in India*, Springer, pp. 67–79 (2000).

[8] : 私立 PDD 図書館.

[9] dwyl: english-words.

[10] Fahl, S., Harbach, M., Acar, Y. and Smith, M.: On the ecological validity of a password study, *Proceedings of the Ninth Symposium on Usable Privacy and Security*, ACM, p. 13 (2013).