2ZA-04

キータイピングに表れるユーザ個人の特徴を利用した文書改ざん防止エディタ

小林可奈子[†1]　秋岡明香[†2]
明治大学大学院[†1]　明治大学[†2]

1. はじめに

情報化社会の発展に伴い,パソコンやタブレット等の使用が必要不可欠となった.それと同時に,資料の作成等の文字入力を伴う作業においては,必然的にキーボードを利用するようになった.しかし,キーストロークにはユーザ毎に個人差があるため,他人のキータイピング傾向を真似ることは容易ではない.本稿では,ユーザ毎のキータイピング傾向を特徴量として,公文書等の重要書類編集者の識別を行うテキストエディタを提案する.キーストロークデータを生体認証に活用する取り組みには,学校や企業で実施されるオンライン授業や試験での出席確認やなりすまし防止などがある.例として,NTT コミュニケーションズはキーストロークダイナミクス認証技術を用いて,e ラーニング授業での本人確認を行う実証実験を実施している[1].しかし,より正確にユーザの識別を行うため,利用開始前に大量の文章を入力しキーストロークデータの登録を行う必要がある.

本研究では,テキストエディタの使用時,文書を編集しながらユーザのキータイピング傾向を学習する.学習したユーザ固有のタイピングデータと新しく入力された文章から抽出したデータを比較することで,正式に文書を編集する権限を持ったユーザであるかを判別する.また,必要に応じてユーザのキータイピング情報の更新を行う.ユーザの識別に使用する特徴量として,キーストロークデータに加え,複数パターンの入力方法があるローマ字の入力傾向と,ミスタイプの傾向を使用する.

2. 関連研究

渡辺らは,キーストロークパターンによる認証技術を提案している[2].入力文字の隣接 2 文字間において,それぞれのキーの押された時間と離された時間を取得し,入力した文字の時間差を総当たりに組み合わせたものを特徴点とする.ユーザが　5〜10 文字程度の文字列を数回入力して特徴点を取得したのち,これらの平均と標準偏差をテンプレートとして保存し,認証試行を行った.ユーザの文字入力時,特徴点の標準偏差が一定の値より少ない場合はなりすましが困難であるとの結果を得たため,標準偏差がその値を下回るまでテンプレートの登録を行う必要があるとしている.

3. ユーザ固有のキータイピング傾向の抽出

3.1 予備実験

ユーザ識別の特徴量を決定するにあたって,ユーザのキータイピング傾向にどのような違いがあるかを分析するため,以下のような予備実験を行った.ひらがな 15 文字程度の短文を 1 回につき 10 個入力し,ユーザ毎のキータイピング傾向を記録する.20 代男性 4 人と 20 代女性 3 人に,この実験を 5 回ずつ繰り返して収集したデータを分析し,以下の 3 つを個人識別に使用する特徴量とした.

3.2 キーストローク

関連研究や予備実験から,キーストロークはユーザ毎に異なる傾向があるため,個人識別の際に有用な特徴となる.本研究では日本語文を対象としており,日本語文を入力する上で使用頻度の少ないキーは有用でないと考える.キーストロークの特徴量抽出には渡辺らと同様の抽出方法を利用した.重要書類作成の際に使用する頻度の高いキーを得るため,想定文書と類似する単語が多く使われるビジネス記事から,頻出のひらがな 1 文字と連続した 2 文字を取得した(句読点を含む).これらをローマ字に変換し,使用頻度の高い 2 つのキーの組み合わせとした.

3.3 入力方法が複数存在するローマ字

ひらがなについては,ローマ字による入力方法が複数パターン存在するものがある.例として,「し(SI/SHI)」「ふ(FU/HU)」「ぁ(LA/XA)」等が挙げられる.これらの文字について,予備実験から,被験者全員,常に決まった単一の入力方法を使用していた.また,キーストロークと比較して,ユーザの使用環境やコンディションによる変動が少ないため,ユーザ固有の特徴として非常に有用である.

3.4 ミスタイプと修正方法

ユーザの利き手やタイピングの習熟度,タイピング時の癖などから,ユーザ毎にミスタイプしやすいキーがあると考えられる.また,特定の入力シーケンスにおいてミスタイプが多く発生する場合もある.これらを考慮し,小松らの分類[3]から,差し込み,除去,置き換えを用いて,ユーザ毎のミスタイプ傾向を取得する.

また,ミスタイプの修正方法にもユーザ毎の特徴がある.入力文字の修正方法によって,どの文字がミスタイプされたものかをリアルタイムで判別する方法が異なる.ある種のユーザはミスタイプをしたことにすぐ気が付き,漢字等への変換や Enter キーを押す前に該当の文字のみを修正する(ミスタイプ直後修正型,図 1).この型では,backspace キーが入力された直後の文字がミスタイプしたキーであると捉えることができる.そのほかに,単語や文章を入力し終え,漢字等に変換または Enter キーを押してからミスタイプに気が付き,ミスタイプした文字を含む単語や文章を最初から入力し直すタイプも存在する(単語丸ごと修正型,図 2).この型では,backspace キーが入力された前後の文字列を,Levenshtein 距離を文字列の長さを考慮して標準化したものを用いて比較し,ミスタイプした文字の特定を行う.Levenshtein 距離とは,ある文字列 S_1 を別の文字列 S_2 に変形するのに必要な操作の最小回数である.文字列の変形操作には,挿入,削除,置換の 3 つが想定されている[4].これにより,ミスタイプによる文章や単語の入力し直しではなく,改稿のために全く別の文章を入力した場合をミスタイプによるものと誤判断することを防ぐことができる.ミスタイプ文字の特定は以下のように行う.修正前の文字列の末尾を基点に,修正後の文字列の先頭文字から 1 文字ずつ増加させ,それぞれ標準化 Levenshtein 距離を求める.標準化 Levenshtein 距離は,Levenshtein 距離を比較する文字列の長い方の文字数で割ることによって求める.この値が最小になった(2 つの文字列の類似度が最大となった)とき,2 つの文字列の差分を取ることでミスタイプ文字を検知することができる(図 3).これは,一部のミスタイプ直後修正型においても,子音入力後の母音をミスタイプした場合などのミスタイプキーの特定に適用することができる.

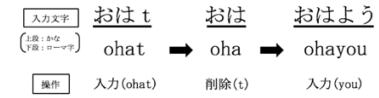

図 1: ミスタイプ直後修正型

Text editor with falsification detector based on user-specific features in key typing
†1 KANAKO KOBAYASHI, Meiji University Graduate School
†2 SAYAKA AKIOKA, Meiji University

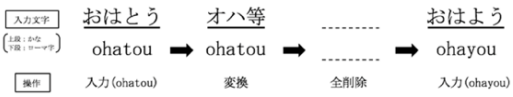

図 2: 単語丸ごと修正型

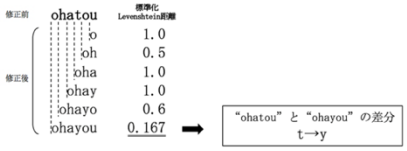

図 3: 標準化 Levenshtein 距離を用いたミスタイプキーの特定

4. 実装

キータイピング傾向の学習　ユーザが初めてテキストエディタを使用するとき，ID とパスワードによってユーザ登録を行う．各ユーザのキータイピング傾向はこの ID と連携して蓄積される．ログインしたユーザがテキストエディタを使用している間，バックグラウンドでユーザのキータイピング傾向を取得する．取得したデータは 3 節で述べた特徴量を扱うスレッドにそれぞれ渡され，リアルタイムでキータイピングの特徴量抽出を行う．

現在ログインしているユーザのキータイピング傾向が十分に得られるまで，ユーザデータの更新を行う(データ不十分状態)．ユーザが文書を上書きする度に更新は行われ，ユーザデータが十分に集まるとデータ十分状態に移行し，更新を止める．この段階でキーストロークデータを比較したときに，差分が 10%程度になる閾値を設定する．しかし，ミスタイプ傾向においては，普段はしないミスタイプをそのユーザの特徴として記録し更新を止めてしまうと，それ以降比較するときに登録されている特徴量との類似度が下がってしまう．これを防ぐために，ミスタイプ傾向には上書き保存の度に半減する重みを設定し，頻繁に行うミスタイプの傾向のみが残るようにする．

ユーザの識別　ユーザが文書の上書き保存をするとき，今回の編集した文章から抽出されたキータイピング特徴量と，あらかじめ登録されたユーザのキータイピング特徴量を比較する．キーストロークデータについては，2 つのデータの差分をとり，設定された閾値を超える割合がいくらか算出する．ローマ字入力のパターンについては入力パターンが変更されたかを，ミスタイプについてはミスタイプしやすいキーと分類，修正方法の傾向を比較する．キーストロークの閾値を上回る項目が多い，またはそのほかの傾向について異なる点が多数あった場合，ログインしたユーザとは別人であると識別する．

5. 評価

まず，被験者 A(20 代女性)が提案テキストエディタを使用して 2,000 字程度の文書を作成する．作成途中の上書き保存も可とし，被験者 A は計 8 回文書の上書き保存を行なった．ここでユーザ識別のためのキーストロークの閾値を設定する．8 回それぞれのキーストロークデータの差分を取り，閾値を超えるデータが全体の 10%におさまるように 0.12 とした．また，被験者 A のミスタイプ傾向を表 1 に示す．

次に被験者 A としてテキストエディタにログインし，被験者 A が作成した文書に対して被験者 B(20 代女性)が文書の改ざんを想定して編集を 5 回行なった．編集文字数はそれぞれひらがなで 5，25，50，75，100 程度とした．

これとは別に，被験者 B には事前に 1,000 文字ほどの文章を入力してもらい，キータイピングデータを分析したところ，キーボード左方にあるキーが全体的に被験者 A より入力速度が遅い傾向にあったが，被験者 A として設定した閾値を超えたデータは 15%とキーストローク傾向は被験者 A とやや似ていることがわかった．また，WPM(words per minute: 1 分あたりの英数キーの入力数)は，被験者 A は 243.93，被験者 B は 236.48 であった．ローマ字の入力パターンについては，両者間で異なるのは「ふ(FU/HU)」のみであった．被験者 B のミスタイプ傾向を表 2 に示す．

編集文字数が 5 のとき，被験者 B はミスタイプをせず，入力パターンが複数あるローマ字を入力しなかったため，キーストロークのみでの識別となったが，ひらがな 5 文字から得られたキーストロークデータの差分が小さく，別人と識別できなかった．

編集文字数が 25，50 のとき，被験者 B はミスタイプをしなかったが，被験者 A と異なるパターンでローマ字を入力していた．このとき，得られたキーストロークデータが少なく，差分は 16%と小さかったが，別人と識別することができた．

編集文字数が 75,100 のとき，被験者 B は複数回ミスタイプをした．ミスタイプ傾向は除去と置き換えが多かったが，被験者 A は除去型のミスタイプを全くしないため，キーストロークの差分と合わせて別人と識別することができた．

入力する文字数が多いほど，得られるキーストロークデータやローマ字の入力パターン傾向は増え，ミスタイプする確率も上昇する．よって，一度に改ざんする文章が長いほど別人と識別できる可能性が高い．また，25〜50 字程度の短い文章から得られるキーストロークデータは十分とは言えないが，その他の特徴量と組み合わせることで別人と識別できることがわかった．しかし，5〜10 文字程度の単語やそれ以下の文字数の編集について，さらに複数回試行した結果，やはりキーストロークデータだけでは本人と別人を識別できないことがあった．

表 1: 被験者 A のミスタイプ傾向

分類	キー
差し込み	I, T
除去	なし
置き換え	S→U, A→I, A→0, R→T
ミスタイプ直後修正型	

表 2: 被験者 B のミスタイプ傾向

分類	キー
差し込み	なし
除去	A,N,U
置き換え	E→U, I→Y, R→T
ミスタイプ直後修正型	

6. おわりに

本稿では，ひらがな約 20 文字以上の編集であれば，キータイピング傾向から文書の改ざんを防止できることを示した．しかし，WPM の値が近似し，キーストローク傾向が似ているユーザ間において，ひらがな 10 文字前後の改ざんを防ぐことは出来ていない．これについて，他に有効な手段を検討する必要がある．

本研究ではおおよそ完成された文書への改ざんを想定しているため，ユーザのキータイピング傾向を更新している段階で別ユーザが文書の編集をすることはないとしているが，実際にこのことが起きないと断言することはできない．

謝辞

本研究は JSPS 科研費 16KK0008 の助成を受けたものである．

参考文献

[1] NTT コミュニケーションズ，“キーボード入力の個人特性を用いた本人性確認サービス「キータッチパス」の北里大学、明治大学における実証実験の実施について”(2010)
https://www.ntt.com/about-us/press-releases/news/article/2010/s20100705.html
[2] 渡辺幸樹，福永孝，塙敏博，“キーストロークパターンによるログイン認証”，第 66 回全国大会講演論文集，pp.283-284(2004)
[3] 小松龍生，中藤良久，“キーボード入力におけるミスタイプ分類と傾向の評価”，FIT2018(第 17 回情報科学技術フォーラム)，第 3 分冊，pp.307-308(2018)
[4] Levenshtein, V. I. *BinaryCodes Capable of Correcting Deletions, Insertions, and Reversals*. Cybernetics and Control Theory, Vol.10, No.8, pp707-710(1966)