ソーシャルメディアにおける非構造化テキストデータの k-匿名化によるプライバシー保護

前田 若菜[1,a)]　鈴木 優[1]　吉野 幸一郎[1]　Graham Neubig[1]　中村 哲[1]

概要：ソーシャルメディアなどに存在する大量のテキストデータは，新たな価値を創出するビッグデータの一つとして注目されている．しかし，プライバシー保護のために，利用に際して個人情報に相当する文字列を匿名化する必要がある．従来手法では，秘匿属性の設定や固有名詞抽出などを行う必要があった．しかし，属性設定漏れによる匿名化漏れや固有名詞以外の文字列から匿名化済み文字列を特定できるおそれがある．さらに，話し言葉のようなくだけた文には対応できない．文の差異に基づく手法では，有効範囲が定型句で構成されたテキストに限定される．そこで本研究では，文字 n-gram フレーズの出現回数に基づいた k-匿名化を提案した．これは，事前の秘匿属性設定が不要，匿名化文字列を固有名詞に限定しないことで文脈からの特定を防止可能，定型句を含まないような多様なテキストに有効，構文解析や単語分析が不要なため話し言葉や未知語にも対応可能という利点がある．実験結果より，k の値の変化と匿名化された文字数の割合が n の値によって異なることを明らかにした．さらに，適切な匿名化状態を定義し，文字特定率から特定不可能性を，人手評価で理解可能性を測った．

1. はじめに

ソーシャルメディアとは，個人が自由に情報を送受信できるメディアである．従来の一方的な送受信とは異なり，双方向のコミュニケーションに開かれている特徴がある．そのため，人々がコミュニティや社会を構築するメディアとして機能している．代表的なものとして，SNS(Social Networking Service) がある．これは，インターネットを介して情報交換を行い，ソーシャルネットワークを構築するサービスである．例えば，Facebook や Twitter，ブログなどがある．これらの普及に伴い，テキストデータは高い頻度で作成されている．さらに，コールセンターの顧客とのやりとり音声のテキスト化や市場調査における自由記述文の電子化などがなされ，多種多様なテキストが蓄積されている．これらのデータから有益な情報を取り出すことが期待されている．しかし，個人情報およびプライバシー保護の観点で問題がある．ここでは個人情報を個人に関するデータとする．プライバシーは私生活情報，非公知情報，個人が公知を望まない情報とする．個人を特定・識別できるに関わらず，プライバシーは保護を期待される領域である．蓄積されたテキストデータ中には，大量の個人情報が含まれている．そのため，収集データの公開やその解析結果の公開は，予期せぬ個人情報流出をまねくおそれがある．例えば，コールセンターのテキスト化データや自由記述のアンケートなど個人情報を含んだこれらのデータは契約の範囲内での利用に限定されている．しかし，データの公開範囲が広がれば，多くの分析・解析が行われ，新たな価値創出につながる．これを可能にするためには，匿名化の処理が必要である．一方，SNS に投稿されたメッセージは，投稿者による自発的な公開情報である．そのため，SNS のメッセージに対して，プライバシーの保護は期待されておらず，自由にデータを利用し公開してよいと考えられている．しかし，SNS は積極的な情報公開と共有を目的としたサービスであり，投稿者の多くは特定の聴衆を念頭において投稿している．そのため，自発的な公開情報であっても，プライバシー領域に属する場合がある（社会ネットワークにおける情報の伝播可能性 [6]）．これらの問題を解決するためには，匿名化処理が必要である．

ところが，匿名化研究の主流は構造化データを対象としたものである．テキストデータは構造定義がなされておらず，情報や属性，並びにその単位は自明ではない．そのため従来は，あらかじめ抽出したい固有名詞の属性を設定し，テキスト中の固有名詞を除去 [1] あるいは一般化 [2]，[3] する手法がとられていた．しかし，属性設定に不備があれば個人情報流出防止効果が低下する．また，匿名化する文字列を固有名詞に限定することによって，それ以外の文字列

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] maeda.wakana.mo7@is.naist.jp

から匿名化された文字列を復元できる可能性がある．さらに，これらは構文解析や単語分析を必要とするため，話し言葉のようなくだけた文や未知語には対応できない．これらの問題を解決する手法として，文の差異に基づく匿名化手法がある [5]．しかし，この手法の有効範囲は定型句を多く含むようなテキストに限定される．

そこで，本研究では，文字 n-gram の出現回数に基づいた匿名化手法を提案する．すべての文字 n-gram が秘匿情報である可能性があると仮定し，出現回数の低い文字 n-gram を秘匿情報と仮定する．そして，同じ文字 n-gram フレーズが k 回以上出現する状態になるように部分文字列を代替文字に置換する．この手法は，従来法と比較して次の四つの利点がある．はじめに，事前の秘匿属性設定が不要なため，属性設定漏れによる匿名化漏れを生じない．次に，文字 n-gram フレーズに基づいて匿名化するため，固有名詞以外の文字列からの特定を防止できる．三つ目に，文字 n-gram フレーズの出現回数に基づいて匿名化しているので，定型句を含まないような多種多様なテキストに有効である．最後に，文字 n-gram フレーズに基づいて匿名化するため，構文解析や単語分析が困難な文である話し言葉や未知語を含む文を匿名化対象とすることができる．

実験では，適切な匿名化を定義し，匿名化率と文字匿名化率を測った．その結果，k の値の変化と匿名化された文字数の割合が n の値によって異なることを明らかにした．さらに，適切な匿名化状態を定義し，文字特定率から特定不可能性を，人手評価で理解可能性を測った．文字特定率から特定不可能性は評価できなかったが，人手評価により理解可能性を算出することができた．

2. 関連研究

Sweeney[4] は．識別の困難さを示す k-匿名性（k-anonymity）という指標を提案している．これは，個人を識別することができる情報を k 個未満に絞り込むことができないことを表す，匿名性の尺度である．同じ情報や属性をもつ個人が k 個以上になる状態を k-匿名性を満たすと呼ぶ．そして，k-匿名性を満たすようにデータを加工することを k-匿名化と呼ぶ．

しかし，Sweeney の指標が想定としているデータは，リレーショナルモデルをベースとした構造化データである．一方，テキストデータは通常構造定義がなされていない．Sweeney の指標をテキストデータにあてはめるためには，情報や属性，並びにその単位を独自に設定する必要がある．しかし，これらは自明ではない．さらに，どのように設定するのが妥当なのかは明らかではない．そのため，従来の匿名化手法の多くは構造化データを対象としていた．

そのなかで，テキストデータを対象とした研究では，匿名化したい単語の属性を事前に設定したうえで，名前や病名，症状などの固有名詞の除去 (サニタイズ) [1] や，職業や地名などを曖昧化する汎化 (一般化)[2]，[3] などの手法を提案している．しかし，問題点として次の二つがある．一つは，あらかじめ消したい単語のカテゴリ，属性を設定する必要がある．Chakaravarth ら [1] は，カルテのテキストデータを対象とし，システム設計の段階で秘匿属性を定義している．Nguyen ら [3] や，Kataoka ら [2] の SNS のテキストメッセージを対象とした手法は，プロフィールに設定した属性に基づいて匿名化している．つまり，秘匿情報としたい属性を利用者が手動で設定できる．しかし，何を秘匿情報とすべきかを決定するのは難しい．なぜなら，秘匿としたい属性を網羅的に把握するのは困難であるからである．事前設定に不備があれば，匿名化漏れが発生するため，個人情報流出の防止の効果は低下する．もう一つの問題として，除去や一般化した固有名詞・固有表現以外の文中の表現から匿名化された文字列を特定できる可能性がある．例えば，「滋賀県を観光．日本最大の湖を見た．」という文について考える．秘匿情報として属性・地名を匿名化するならば，固有名詞「滋賀県」は除去される．あるいは，「関西」や「西日本」など，より上位の単語・概念に一般化される．しかし，知識があれば，あるいは，検索や調査を行えば，「日本最大の湖」は「滋賀県にある琵琶湖」だと推測できる．結果，匿名化された固有名詞「滋賀県」は特定される．このような事象を，本研究では「文脈から特定できる」状態と定義する．

これらの問題を解決するため，荒牧ら [5] は文を単位とした k-匿名化を提案している．これは，どのような文字列で検索してもヒットする文が k 個以上になるようにテキストの文字列を除去する手法である．任意の一文 S_x に対し，順序も一致する最長共通部分列の長さ（「ABC」と「CAB」では，「AB」のみマッチ）を算出し，値が大きい順に $(k-1)$ 個の類似文 S_i を抽出する．そして，抽出された S_i との最長共通部分列以外の文字列を S_x から削除，あるいは代替文字列（$*$など）に置換する．この手法は，固有名詞に限定せずに文字列を削除することによって，文脈からの匿名化済み文字列の特定を防ぐことができる．さらに，比較する文同士の差異に基づいて匿名化を行うので，あらかじめ消したい属性を設定する必要がない．ただし，荒牧らが用いた実験データは日本語のカルテテキスト EHR(Electronic Health Record) である．カルテのようなテキストデータは定型句が多く含まれており，文の多くが似通う可能性が高い．そのため，k 個の文の差異を除去したとき，もとの文を一定量保持することができる．しかし，この手法は多様性のあるテキストには適合しない．例えば，SNS のテキストデータに手法を適用した場合，多くの文中の部分文字列が除去される．SNS には共通の定型表現はなく，個々人が自由に多種多様なメッセージを投稿しているからである．したがって，この手法は，有効なテキストの種類が定型句を多く含むようなテキストに限定されるという問題がある．

3. 文字 n-gram に基づく k-匿名化

本研究では，文字 n-gram を情報とする．秘匿すべき属性は設定しない．そのため，すべての文字 n-gram が秘匿情報である可能性があると仮定する．そして，n-gram の出現回数を単位として数え上げる．出現回数の低い文字 n-gram は，秘匿情報であると仮定する．そして，秘匿情報である文字 n-gram フレーズを少なくとも k 個未満に絞り込めないことを匿名性の尺度とする．したがって，本研究における匿名化とは，同じ文字 n-gram フレーズが k 回以上出現する状態になることである．この状態を本研究では k-匿名性を満たすと呼ぶ．そして，k-匿名性を満たすようにデータを加工することを k-匿名化と呼ぶ．

提案手法の利点は次の四つである．

事前の秘匿属性設定が不要

従来法では，消したい単語のカテゴリ，属性を事前に設定する必要があった [1]，[2]，[3]．しかし，何を秘匿情報とすべきかを決定するのは難しく，秘匿属性を網羅するのは困難である．さらに，属性設定に不備があった場合，匿名化漏れが生じる．提案手法では，秘匿属性を設定せず，すべての文字 n-gram を秘匿情報である可能性があると仮定している．そのため，属性設定不備による匿名化漏れを生じない．

文脈からの特定を防止可能

従来法では，固有名詞に基づいて秘匿情報を匿名化している [1]，[2]，[3]．しかし，固有名詞以外の文字列から，匿名化された文字列を特定できる可能性がある．提案手法では，全ての文字 n-gram の出現回数に基づいて匿名化を行う．出現回数の低い秘匿情報となる可能性のある文字 n-gram は，固有名詞に限らず匿名化される．そのため，固有名詞に限定された匿名化よりも，文脈から特定できる状態を回避できる．

広範囲のテキストに有効

従来法では，比較する k 個の文同士の差異に基づいて匿名化している [5]．しかし，定型句を含まないような多様性のあるテキストに対しては，ほとんどの文字列が匿名化される．そのため，適合するテキストの種類が限定される．提案手法では，文字 n-gram の出現回数に基づいて匿名化するため，文型の影響をうけない．そのため，文の差異に基づいた匿名化よりも，広範囲のテキストに対して有効である．

話し言葉や未知語に対応可能

従来法では，単語分析や構文解析などを用いて匿名化する単語を抽出している [1]，[2]，[3]．しかし，SNS のような，話し言葉や文法に従わない文では，解析が困難である．加えて，未知語に対応することができない．提案手法では，文字 n-gram に基づいて匿名化するため，テキストを構文解析や単語分析をする必要がない．そのため，文のくずれに由来する構文解析誤りや未知語の影響を受けない．

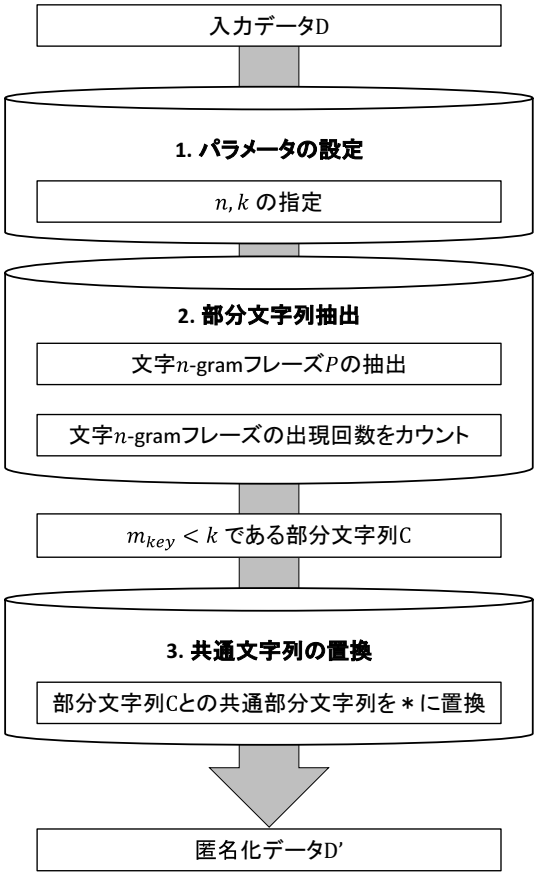

図 1 システムの枠組み

3.1 システム構成

提案手法は図 1 のような枠組みで成り立っている．入力データとして，あらかじめ空白文字を除去した匿名化したい N 個のデータを要素とする群 $D = \{d_0, ..., d_{N-1}\}$ を用いる．そして，次の処理に従って入力データを匿名化する．

(1) パラメータの設定
(2) 部分文字列抽出
(3) 共通部分文字列の置換

以下，詳細を記述する．

3.2 パラメータの設定

パラメータとして文字 n-gram の n, k-匿名化の $k(\geq 2)$ の値を設定する．n の値によって，何文字の部分文字列を抽出するかを決定する．また，k の値によって，全データ中の部分文字列の最低出現回数を指定する．

例えば，$n = 5$ かつ $k = 10$ のとき，D から文字 5-gram フレーズを抽出する．そして，D 中に，出現回数が 10 回未満の文字 n-gram フレーズが現れないように匿名化する．なお，ここで記述した処理の詳細は以下の項で説明する．

3.3 部分文字列抽出

ここでの目的は，D 中の出現回数が k 回未満の部分文字列を抽出することである．そのため，次の処理に従って部分文字列を抽出する．

(1) 文字 n-gram フレーズの抽出
(2) 文字 n-gram フレーズの出現回数をカウント

以下，順を追って説明する．

3.3.1 文字 n-gram フレーズの抽出

$d_h(h = 0, ..., N-1)$ から，文字 n-gram フレーズを取り出す．d_h の文字数が $l_{dh}(> 0)$ 文字のとき，$p_{hi}(i = 0, ..., l-n+1) \in P$(P は文字 n-gram フレーズの全体集合) を取り出す．ただし，p_{hg}=$p_{hi}(i<g\leq l)$ の場合，p_{hg} は重複するので破棄する．

例として，表 1 を用いて文字 2-gram フレーズ p の抽出を説明する．d_0 の文字数 l_{d0}= 12 より，文字 2-gram フレーズ p_0 は $l-n+1 = 12-2+1 = 11$ 個取り出せる．しかし，$i = 0, 3$ のとき，p_{00}=p_{03}=”福岡”と重複している (表 1 の p_0 の行の色のついたセル)．そのため，p_{03} は破棄する．したがって，処理を適用した $p_0{}'$ を d_0 から抽出した文字 2-gram フレーズ群とする．同様の処理を d_1, d_2 にも行い，抽出する．

3.3.2 文字 n-gram フレーズの出現回数をカウント

p の出現回数を調べる．p を key に格納していき，その都度 key の出現回数 m_{key}(初期値 0) に 1 加算していく．重複する p がある場合，p は key へ新たに格納せず，m_{key} に 1 加算する処理のみ行う．m_{key} が k 未満の p を部分文字列 $c \in C$(C は c の全体集合) とする．補足すると $C \subseteq key \subseteq P$ という関係が成り立っている．

表 2 を用いて，処理の例を示す．ここで用いる D，P は表 1 と共通である．$k = 2$ のとき，出現回数 $m_{key} < k = 2$ となる key を抽出すればよい．まず，$p_0{}'$ から文字 n-gram フレーズを key に格納し，m_{key} に 1 加算する．次に，p_1 の文字 n-gram フレーズをみていく．このとき，p_{10}=”福岡”は，すでに key_0 に格納されている．そのため，”福岡”を新たに key に格納はしない．ただし，m_{key0} には 1 を加算する．p_2 に対しても同様の処理を行う．最終的に，m_{key0}= 2 より，key =”福岡”は匿名性を満たすとして部分文字列として利用しない．一方，表 2 の m_{key} の列で色のついたセルは $m_{key} < k = 2$ を満たす．これらは，部分文字列 c として抽出する．

3.4 共通部分文字列の置換

文字数が l 文字のデータ d に対し，要素数 l，初期値 0 の配列 A_d を用意する．次に，d と C を比較していく．d の j 番目の文字 $s_j(j = n-1, ..., l-n-1)$ が s_{j+n-1} まで連続して c と一致するとき，$A_d[j]$ から $A_d[j+n-1]$ までの要素に 1 を加える．C と比較後，A_d の要素が 1 以上になっているインデックスに対応する s_j を＊に置換する．これにより，匿名化が達成される．

表 3 に比較する部分文字列 C を，表 4 に共通部分文字列の置換の例を示す．示す．入力データ d_1=”福岡県北九州市早瀬区新垣 5” と部分文字列 C を比較する．$n = 2$ より，2 文字連続で C と d_1 が一致するインデックス j を求めればよい．これは，表 4 の色つきのセルにあたる．したがって，j =2, 3, 4, 5, 7, 11, である．ゆえに，$j = 2$ のとき，$A_{d1}[2]$, $A_{d1}[3]$ にそれぞれ 1 を加える．同様に，$j = 3$ のとき，$A_{d1}[3]$, $A_{d1}[4]$ に，$j = 4$ のとき，$A_{d1}[4]$, $A_{d1}[5]$ に，$j = 5$ のとき，$A_{d1}[5]$, $A_{d1}[6]$ に, $j = 7$ のとき，$A_{d1}[7]$, $A_{d1}[8]$ に， $j = 11$ のとき，$A_{d1}[11]$, $A_{d1}[12]$ に，それぞれ 1 を加える．A_{d1} の要素が 1 以上になるのは，j = 2, 3, 4, 5, 6, 7, 8, 11, 12 のときである．したがって，d_1 は $d_1{}'$ =”福岡＊＊＊＊＊＊＊区新＊＊”と匿名化される．

4. 実験

本実験では，提案手法の結果を分析し，評価する．

提案手法を評価するにあたって，本研究では「適切な匿名化」状態を「匿名化された文字列を完全に特定できないが，内容は理解できる・分析は可能である」状態と定義する．そして，適切な匿名化であるかを評価するために，特定不可能性と理解可能性という二つの指標を用いる．

匿名化処理は，はじめにで述べたように「個人情報を含んだ収集データやその解析結果の公開」を可能にすることを目的としている．つまり，匿名化データを用いてなんらかの分析や理解ができる必要がある．すなわち，過度の匿名化はデータ中の情報量を過度に減少させ，分析や理解ができなくなるため有効ではない．さらに，匿名化が不十分なデータは個人情報流出を引き起こすため妥当ではない．したがって，手法によって適切な匿名化状態を実現できるかを評価する必要がある．そこで，本研究では匿名化された文字列を完全に特定できないことを測る指標として特定不可能性を，内容は理解できる・分析は可能であることを測る指標として理解可能性を用いる．

実験データとして，適切な匿名化状態を観測しやすい事業所名を利用した．事業所名は，事業の内容や形態をあらわす部分文字列が繰り返し出現する特徴をもつ．たとえば，株式会社や医療法人などがある．一方，出現回数の低い部分文字列は，事業所名を特定できる情報である．そのため，出現回数の低い部分文字列を匿名化すれば，「事業所を特定することはできない（特定不可能性）が，その事業の内容はわかる（理解可能性)」状態になる．このように事業所名は適切な匿名化状態を観測しやすい．

4.1 実験 1：事業所名による匿名化実験・評価

4.1.1 実験方法

実験データとして，福岡市の事業所名 387 データを利用した．これらに対し，提案手法である n-gram・k-匿名化を

表 1　文字 n-gram フレーズ抽出

h	d_h	i	0	1	2	3	4	5	6	7	8	9	10	11
0	福岡県福岡市早通区新谷 3	p_0	福岡	岡県	県福	福岡	岡市	市早	早通	通区	区新	新谷	谷 3	-
0	福岡県福岡市早通区新谷 3	p_0'	福岡	岡県	県福	岡市	市早	早通	通区	区新	新谷	谷 3	-	-
1	福岡県北九州市早瀬区新垣 5	p_1	福岡	岡県	県北	北九	九州	州市	市早	早瀬	瀬区	区新	新垣	垣 5
2	福島県福島市瀬区新垣	p_2	福井	井県	県福	井市	市瀬	瀬区	区新	新垣	-	-	-	-

表 2　部分文字列抽出

key	m_{key}	c
福岡	2	-
岡県	2	-
県福	2	-
岡市	1	岡市
市早	2	-
早通	1	早通
通区	1	通区
区新	3	-
新谷	1	新谷
谷 3	1	谷 3
県北	1	県北

key	m_{key}	c
北九	1	北九
九州	1	九州
州市	1	州市
早瀬	1	早瀬
瀬区	2	-
新垣	2	-
垣 5	1	垣 5
福井	1	福井
井県	1	井県
井市	1	井市
市瀬	1	市瀬

表 3　2-gram・k-匿名化において比較する部分文字列

岡市	新谷	北九	早瀬	井県
早通	谷 3	九州	垣 5	井市
通区	県北	州市	福井	市瀬

表 4　2-gram・k-匿名化の例

j	0	1	2	3	4	5	6	7	8	9	10	11	12
d	福	岡	県	北	九	州	市	早	瀬	区	新	垣	5
Ad	0	0	1	2	2	2	1	1	1	0	0	1	1
d'	福	岡	*	*	*	*	*	*	*	区	新	*	*

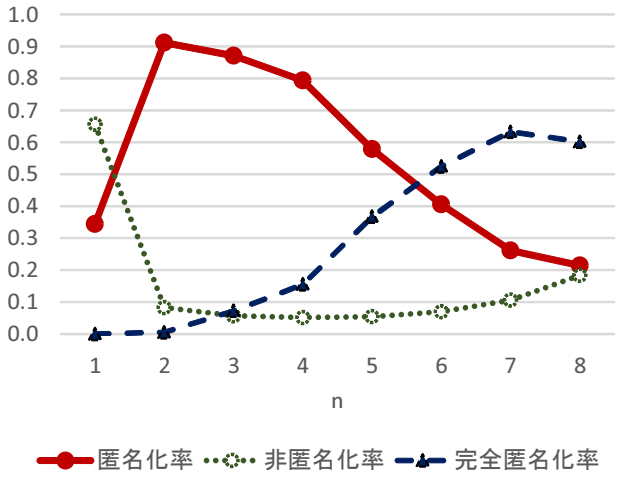

図 2　n-gram・2-匿名化における匿名化率の変化

行う．評価指標として，匿名化率，文字匿名化率を用いる．そして，n や k の値によってこれらの指標がどのように変化するか比較する．評価指標は以下のように定義する．

匿名化率

匿名化率を定義する前に，非匿名化率，完全匿名化率を定義する．非匿名化率とは，全データに対する一文字も匿名化されなかったデータ数の割合である．たとえば，d=”福岡県”が匿名化されず”福岡県”のままであったとき，d は非匿名化されたと呼ぶ [*1]．完全匿名化率とは，全データに対する全文字が匿名化されたデータ数の割合である．d=”福岡県”の全文字が匿名化されて”＊＊＊”となったとき，d は完全匿名化されたと呼ぶ．そして，全データにおける匿名化率を，次のように定義する．

匿名化率 = 1−(非匿名化率 + 完全匿名化率)

本実験における，適切な匿名化状態は，先に「匿名化された文字列を完全に特定できない（特定不可能性）が，内容は理解できる・分析は可能である（理解可能性）」状態と定義した．そのため，一文字も匿名化されない非匿名化状態は，前者の特定不可能性に反する．また，全文字匿名化という完全匿名化状態は，後者の理解可能性に反する．このことから，非匿名化や完全匿名化は不適切な状態であるといえる．したがって，全体から非匿名化率と完全匿名化率をのぞいた匿名化率が，適切に匿名化された割合を表す．

文字匿名化率

文字匿名化率は，全データの文字数に対し，匿名化された文字数の割合を表す．つまり，全データの文字数に対する＊の数の割合である．

4.1.2　結果

図 2 は，$n = 2, 3, \cdots, 8$，$k = 2$ のときの n-gram・2-匿名化の完全匿名化率・非匿名化率・匿名化率を示したものである．匿名化率は，$n = 2$ のとき最大となり，$n \geq 3$ のとき減少する．これは，完全匿名化率が増すからである．完全匿名化率は，$n = 2, 3, \cdots, 7$ のとき増加し，$n = 8$ のとき減少する．非匿名化率は $2 \leq n \leq 4$ のとき減少しているが，$n \geq 5$ のとき増加する．例えば，$n = 5$ のとき，共通部分文字列として 5 文字未満の文字列は抽出されない．そのため，5 文字未満のデータは匿名化されない．結果，非匿名化率が増加する．

図 3 は，n-gram・k-匿名化の匿名化率の変化を示したものである．n =2, 3, 4, 1 の順で匿名化率が高い．$n = 1$ は，$2 \leq k \leq 20$ の範囲で増加，$n = 1$ のとき匿名化率は最大値 1 をとる．$n \geq 21$ の範囲で減少していく．$n = 2$ は，$2 \leq k \leq 4$ の範囲で増加．$k \geq 5$ の範囲で減少していく．$n \geq 3$ のとき，

[*1] なお，非匿名化データを含んでいたとしても，必ずしも k-匿名化の定義には反しない．なぜなら，一文字も匿名化されていなくとも他に類似した文字列が出てくることで，データを $\frac{1}{k}$ 以上に特定できないからである．

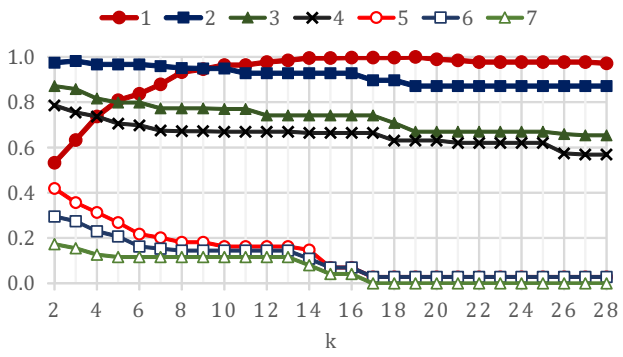

図 3　n-gram・k-匿名化における匿名化率の変化

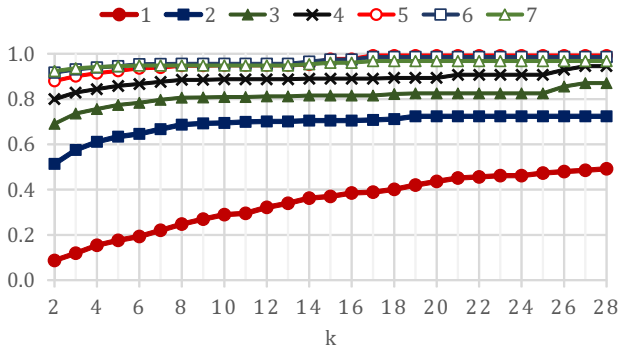

図 4　n-gram・k-匿名化における文字匿名化率の変化

k の増加とともなって匿名化率は減少していく．

図 4 は，n-gram・k-匿名化の文字匿名化率の変化を示したものである．$n = 1$ のとき，文字匿名化率が低い．このことを，1-gram・k-匿名化は文の保存率が高いとも言い換えられる．n の値が増すごとに，文字匿名化率は増加している．また，k の値が増加した場合にも文字匿名化率が増加している．

4.1.3　考察

n の値が増加するにつれて，匿名化率は増加する．しかし，匿名化率が最大値を過ぎると，完全匿名化率が増加し，非匿名化率と匿名化率は減少する．そして，ある値を境に非匿名化率が増加し，完全匿名化率と匿名化率が減少する．n の値がある値より小さいとき，匿名化率は最大値をとるまで増加し，その後減少する．n の値がある値より大きいとき，k の増加につれて匿名化率は単調減少する．n の値が小さいほど，文字列の保存率が高い．n や k の値の増加につれて，文字匿名化率は単調増加する．

匿名化率の増減は完全匿名率と非完全匿名率の増減が関係している．匿名化率が増加するのは，出現回数 k の値が増すことで，非匿名化率が減少するからである．匿名化率が減少するのは，出現回数 k の値が増すことで，完全匿名化率が増加するからである．

以上のように，実験 1 の結果からわかることは，k や n の値によって，部分文字列がどれだけ匿名化されるか，ということである．しかし，特定不可能性や理解可能性を達成できているかを評価することはできない．そのため，別の指標を用いて特定不可能性や理解可能性といった匿名性を測る必要がある．したがって，適切な匿名化状態であるかを測るために，実験 2 の被験者実験による評価を行う．

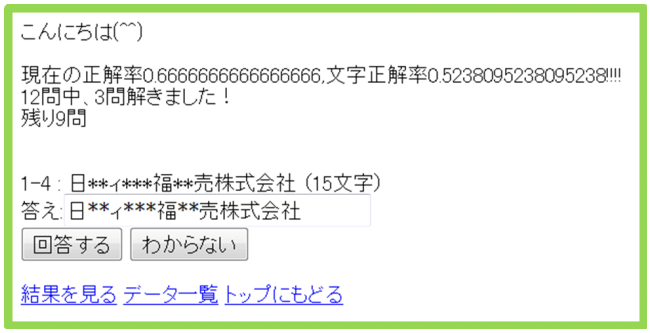

図 5　回答画面例

4.2　実験 2：被験者実験による評価

実験 1 では，k や n の値によって，部分文字列がどれだけ匿名化されたかを評価した．しかし，実際に特定不可能性と理解可能性を達成した適切な匿名化状態になっているかはわからなかった．そこで，適切な匿名化状態であるかを測るために，被験者実験による評価を行う．

4.2.1　方法

実験データの説明をする．元データとして福岡市の事業所名 387 データを利用する．元データに対し，1-gram・2-匿名化したデータと 2-gram・2-匿名化したデータを作成する．適切に匿名化された 1-gram・2-匿名化データ 133 データから 3 データ，適切に匿名化された 2-gram・2-匿名化データ 353 データから 9 データ取り出す．これらを合わせた 12 データを実験データとする．なお，元データの重複は許す．

実験方法として，被験者は問題回答型システムを用いて匿名化された文字列を特定する．図 5 に，システムの回答画面例を示す．システムはまだ出題されていない出題文集合からランダムに 1-gram・2-匿名化されたデータか，2-gram・2-匿名化されたデータを出題する．被験者は出題文中の＊に変換された匿名済み文字列を前後の文字列から推測する．そして，入力フォーム内に記入された出題文の＊部分を推測した文字に置き換え，回答する．わからない場合は「わからない」ボタンを押すか，回答に＊を含んだまま「回答する」ボタンを押す．

評価に用いる指標について説明する．特定不可能性を測る指標として文字特定率を用いる．文字特定率を定義するために，先に匿名済み文字数と正解文字数を説明する．匿名済み文字数は，出題文中の匿名化済みの文字数をあらわす．つまり，出題文中の＊の個数をあらわす．正解文字数は，被験者が匿名済み文字を特定できた文字数をあらわす．

そして，文字特定率は，匿名済み文字数に対する正解文字数の割合である．これは，被験者がどのくらいの割合文字を復元できたかを示す．

さらに，被験者の回答から，被験者は事業所の事業内容を理解できるかを人手で評価する．被験者が事業の内容を出題文から理解できるならば○，理解できないなら×と評価する．この指標を用いて，匿名後のデータから事業内容を理解できるかという「理解可能性」を測る．

4.2.2 結果

表 5，表 6 は実験結果の例をまとめたものである．n はどの文字 n-gram・2-匿名化によって匿名化されたかをあらわす．$n = 1$ のとき文字 1-gram，$n = 2$ のとき文字 2-gram である．匿名化データの列は，実際に被験者に出題した問題文である．回答の列は，被験者の回答をのせている．正解の列では，匿名化前の元データをのせている．全体の結果として，特定率は平均 25%，事業内容の理解ができているのは 62%であった．追って，結果の詳細を記述していく．

表 5 は，適切に匿名化された例である．被験者の回答をみると，その事業所名を特定できていない．また，事業内容の理解については，出題文から $i = 0$ は税務署，$i = 1$ は株式会社，$i = 2, 3$ は医療系，$i = 4$ は学校関係ということがわかる．$i = 2$ の回答を取り上げる．被験者は「医」という文字を回答している．被験者の回答である「医院」と正解データの「病院」はともに医療系の事業内容である．この結果から，被験者は匿名後のデータから，医療系の事業であると理解できている．しかし，事業所を特定する「原土井」の文字列は特定できていない．このことから，提案手法が「事業の内容はわかるが，その事業所を特定することができない状態」を達成できていることがわかる．

表 6 は，不適切に匿名化された例である．$i = 0$ は，ほとんどの文字列が匿名化されている．そのため，事業所を特定することができないと同時に，事業内容を理解することができない．被験者の回答をみると，事業内容が運送系であることを理解していないことがわかる．この匿名化データは「事業の内容はわかるが，その事業所を特定することができない状態」を達成できていない．$i = 1$ は，被験者の回答をみると，被験者は事業所名を特定できておらず，かつ事業の内容が金融系であることも理解できていないことがわかる．匿名化データでは，「金」という文字が匿名化されているため「金庫」という文字列が出現しない．残された「庫」という文字から文字列「金庫」を推測するのは難しいと考えられる．この匿名化データは「事業の内容はわかるが，その事業所を特定することができない状態」を達成できていない．$i = 2$ は，被験者の回答をみると，被験者は事業所を特定する「千早」の文字列を特定できていないが，事業の内容が医療系であることも理解できていない．匿名化データでは，「病」という文字が匿名化されているため「病院」という文字列が出現しない．残された「院」という文字から文字列「病院」を推測するのは難しいと考えられる．この匿名化データは「事業の内容はわかるが，その事業所を特定することができない状態」を達成できていない．

4.2.3 考察

不適切に匿名化されたものの中 n に，ほとんどの文字列を匿名化したことによって理解できない状態になったものがあった．これは，出現頻度の低い文字 n-gram フレーズによって事業所名が構成されているからである．この問題への対処方法として，出現頻度の高い部分文字列を残す方法が考えられる．他に，匿名化したい文字列と残しておきたい文字列が逆転して匿名化されたものがあった．これは，出現頻度の高い部分文字列で事業所名が構成されているからである．さらに，事業の内容を表す文字 n-gram フレーズの出現頻度が低かったことも関係している．例えば，「千早病院」は「早病」の出現頻度が低いために，事業内容を表す「病院」の「病」が匿名化されている．この問題に対処する方法として次の方法が考えられる．まず，保持しておきたい属性を設定しておき，それに属する単語を非匿名化フレーズとして用意する．そして，n-gram フレーズを匿名化する際，非匿名化フレーズを含んでいるならば匿名化しない．さきほどの例で言えば，「千早病院」を匿名化したとき，「病」が匿名化される．しかし，非匿名化フレーズに「病院」が含まれているならば，匿名化された「病」を復元する．結果，事業内容を表す「病院」は匿名化されない．

文字特定率からは，特定不可能性を測ることはできなかった．これは，文字特定化率の高いデータは，文字匿名化率が低く，さらに前後の文字が推測できる事業の内容をあらわす文字列が匿名化されているからである．また，出現頻度の高い部分文字列で構成された事業所名であることも関係している．したがって，文字特定化率は必ずしも特定不可能性を表していなかった．そのため，文字特定率にかわる，特定不可能性を測る指標を考える必要がある．

事業所の種類の理解からは，提案手法の理解可能性を算出することができた．この指標を用いて，他の手法と比較することが可能になると考えられる．

5. おわりに

本研究では，文字 n-gram フレーズの出現回数に基づいたテキストの k-匿名化を実装し，評価した．その結果，k の値の変化と匿名化された文字数の割合が n の値によって異なることを明らかにした．さらに，適切な匿名化状態を定義し，文字特定率から特定不可能性を，人手評価で理解可能性を測った．

今後の課題として，次の四つがあげられる．

- 本研究で観測された不適切な匿名化への対処法を考案し，システムを改良する．
- ソーシャルメディアのテキストデータとして Twitter

表 5　適切な匿名化の例

i	n	匿名化データ	回答	正解	正解文字数	匿名済み文字数	文字特定率	事業内容の理解
0	1	**税務署	福岡税務署	香椎税務署	0	2	0%	○
1	2	*****岡株式会社	九州大学福岡株式会社	アビスパ福岡株式会社	1	5	20%	○
2	2	医療法*****院	医療法人***医院	医療法人原土井病院	1	5	20%	○
3	1	医療法人原*井病院	医療法人原田井病院	医療法人原土井病院	0	1	0%	○
4	2	福****学	福****学	福岡女子大学	0	4	0%	○

表 6　不適切な匿名化の例

i	n	匿名化データ	回答	正解	正解文字数	匿名済み文字数	文字特定率	事業内容の理解
0	2	九**********	九州大学理工学研究科	九州運輸局福岡運輸支局	1	10	10%	×
1	2	商******庫 福****ンタ****	商******庫 福****ンタ****	商工組合中央金庫 福岡流通センター出張所	0	14	0%	×
2	2	国家公務** ＊済組合連合****院	国家公務員* ＊済組合連合****院	国家公務員等 共済組合連合会千早病院	1	7	14%	×

などを実験データとして利用する．

- 文字特定率に代わる特定可能性を測る評価指標を作成する
- 従来手法と比較し，提案手法の有効性を確認する．

謝辞　本研究の一部は，NAIST ビッグデータプロジェクトおよび JSPS 科研費 23700113 によるものである．

参考文献

[1] Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 843–852. ACM, 2008.

[2] Haruno Kataoka, Natsuki Watanabe, Keiko Mizutani, and Hiroshi Yoshiura. Dcnl: Disclosure control of natural language information to enable secure and enjoyable e-communications. In *U-and E-Service, Science and Technology*, pp. 131–140. Springer, 2009.

[3] Hoang-Quoc Nguyen-Son, TRAN Minh-Triet, Hiroshi Yoshiura, Noboru Sonehara, and Isao Echizen. Anonymizing personal text messages posted in online social networks and detecting disclosures of personal information. *IEICE TRANSACTIONS on Information and Systems*, Vol. 98, No. 1, pp. 78–88, 2015.

[4] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[5] 荒牧英治, 増川佐知子, 宮部真衣, 森田瑞樹. テキストの k-匿名化. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2012, No. 9, pp. 1–8, 2012.

[6] 松前恵環. Sns におけるプライバシーの期待と保護のあり方: Lj ストゥラホラヴィッツの「プライバシーの社会ネットワーク理論」 を手がかりに. *Journal of Global Media Studies*, Vol. 13, pp. 75–84, 2014.