D-004

文脈からの推定を考慮した非構造化テキストデータの匿名化手法
Anonymization technique for Unstructured text data considering inference from context

前田若菜 † 鈴木優 † 吉野幸一郎 † 中村哲 †
Wakana Maeda Yu Suzuki Koichiro Yoshino Satoshi Nakamura

1. はじめに

総務省の調査 [3] より，非構造化テキストデータの増加が予想されている．そのため，これらのデータ活用が期待されている．しかし，個人情報を含んだテキストデータについては，データの公開や委託分析をするうえで個人情報が漏洩しないようにデータを加工する必要がある．したがって，元データの状態をなるべく保ち，かつ個人情報が保護されるようにデータを加工する匿名化技術が必要である．

テキストデータの匿名化の方法の一つに，匿名化すべき対象をリスト化し，それを参照してテキストデータ中の該当箇所を「＊＊＊」のように置換する黒塗り (redaction) がある．例えば，電子カルテの自由記述文において，患者の名前を匿名化する場合を考える．このとき，患者の名前リストを作成し，自由記述文とリストを照らし合わせ，該当する文字列を黒塗りすることで名前を匿名化する．この方法の利点は，リストさえ準備できれば，複雑な処理を行わずに匿名化ができる点にある．一方で，テキストデータ中のリストと一致した文字列は全て黒塗りされるため，テキストデータ中の残存文字数は減少する問題がある．

そこで本研究では，参照リストとのパターンマッチングを用いた部分文字列の k-匿名化手法を提案する．この手法では，リストに対してパターンマッチングを行った際，k 件以上ヒットするように秘匿対象文字列の部分文字列を黒塗りする．残存文字数を確保するために，本手法では秘匿対象文字列の全文字列でなく部分文字列を黒塗りする．ただし，この場合は残存する文字列から秘匿対象文字列を推測できる可能性が生じる．そのため，本手法では再特定のリスク増加を防止するため，k-匿名性の指標を用いている．これにより，参照リストと照合したときに匿名化済み文字列を一意に識別できなくなる．たとえば，参照リストが漏洩あるいは容易に照合できる場合，再特定を試みることができる．しかし，匿名化済み文字列をリストと照合したとき，マッチする項目が k 個以上であれば，一意に識別することが不可能である．具体的には，リストとパターンマッチングを行った際．ヒット件数が k 件以上になるような部分文字列を特定し，黒塗りする．

この手法の利点は，参照リストを用いた従来の匿名化と比較して残存文字数の減少を抑制できる点である．秘匿対象文字列の全文字列を黒塗りするのではなく，部分文字列を黒塗りするため，残存文字数の減少を抑制できる．

2. 関連研究

既存の非構造化テキスト匿名化手法を大別すると二つの方法論がある．一つはパターンマッチング，もう一つは機械学習である．また，両方を用いたハイブリット式の方法もある．ただし，Meystre らの調査 [1] によれば，匿名化手法の大部分は機械学習を用いず，パターンマッチングやルール，辞書のような単語リストのみに依拠している．これは，アノテートされたトレーニングデータが不要なことと，ルールや辞書を追加することで比較的容易にパフォーマンスを改善できることに由来する．ただし，マッチした文字列を除去するために，テキストデータ中の残存文字数が減少する問題がある．

匿名性の評価について，識別の困難さを示す k-匿名性 (k- anonymity) という指標がある [2]．これは，ある情報から個人を k 人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．本研究では，この指標に基づいて匿名性を評価する．

3. 提案手法

本研究では k-匿名性を満たす状態を秘匿対象文字列の部分文字列を黒塗りしたものを参照リストに対してパターンマッチングを行った結果，ヒット数が k 件以上になった状態をさす．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．

はじめに，(1) 入力データとパラメータを設定する必要がある．入力データとして匿名化したい文書集合と参照リストを用いる．さらに，パラメータ設定によって何文字フレーズずつ部分文字列を黒塗りするか，参照リストとパターンマッチングして何件以上ヒットする状態にするかを決める．

処理としてはまず，(2) 入力文書集合中の秘匿情報文字列を参照リストを元に抽出する．次に，(3) 抽出した秘匿情報文字列に対し，k-匿名性を満たすような部分文字列の黒塗り箇所を求める．その後，(4) 入力文書集合における秘匿情報文字列を部分文字列を黒塗りしたものに置換していく．

以下，詳細を記述する．

(1) 入力データとパラメータ設定

入力データとして匿名化したい N 個の文書集合 $D = \{d_0, ..., d_{N-1}\}$ と参照リスト $L = \{l_0, ..., l_{M-1}\}$ を用い

† 奈良先端科学技術大学院大学情報科学研究科，Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

る．さらに，パラメータとして文字 n-gram フレーズの n(≥ 1)，k-匿名化の $k(\geq 2)$ の値を設定する．n によって何文字フレーズずつ部分文字列を黒塗りするか決定する．k の値によって，参照リストとパターンマッチングした結果，何件以上ヒットする状態にするかを決める．

例 あるテキストデータ中の事業所名を参照リストを用いて匿名化したいと考え，事業所名リストとして Wikipedia のページタイトル一覧を用意した．さらに，パラメータとして $n = 1$，$k = 3$ と設定した．つまり，文字 1-gram フレーズずつ部分文字列を黒塗りし，参照リストとパターンマッチングした結果，3 件以上ヒットする状態になるよう設定した．

(2) 秘匿情報抽出

ここでは，入力文書集合から参照リストを用いて秘匿情報を抽出する．具体的には，入力文書集合について形態素解析を行い，抽出した名詞が参照リストに含まれるか調べる．参照リストに含まれていれば，抽出した名詞は秘匿情報であると判断する．秘匿情報と判断された文字列は秘匿情報配列 S に挿入する．

例 入力した文書集合に対し，入力したリストを参照してテキストデータ中の事業所名を調べた．その結果，”NAIST”という事業所名が検出された．

(3)k-匿名性の検証

ここでは，(1) の処理で抽出した m 個の名詞を格納した秘匿情報配列 S に対して処理を行う．$|s_i| = l$ を $s_i(\in S)$ の文字列の長さ，$s_i[t : t + n - 1]$ を文字列 s_i の t 番目から $t + n - 1$ 番目までの n-gram フレーズとする．この配列 S のすべての要素に対し，$t = 0$ から $t = l - n$ の範囲で t を 1 ずつ増やして処理を行う．$s_i[t : t + n - 1]$ を任意の 1 文字を表すワイルドカード（正規表現であれば”.”にあたる）に置換し，パターンマッチングを行う．パターンマッチングを行った置換後の文字列を $K_{i,t}$，ヒット件数を $V_{i,t}$ とする．$0 \leq t \leq l - n$ の範囲で，$minV_{i,t} < k$ ならば，置換文字列を増やして再度処理を行う．$minV_{i,t} \geq k$ ならば，$a_i = \{K_{i,t'} | t' = argminV_i(\geq k)\}$ を匿名化後文字列配列 $A = [a_0, ..., a_{m-1}]$ に挿入する．

例 部分文字列として 1-gram フレーズ匿名化するため ($n = 1$)，黒塗り候補位置は 5 箇所 ($0 \leq t \leq 4$) になる．そこで，いずれかの位置を黒塗りした場合のリストとのパターンマッチング結果を求める．

パターンマッチングの結果，”*AIST”のとき {JAIST, KAIST, NAIST} の 3 件がヒット，”N*IST”，”NA*ST”のとき {NAIST} の一件，”NAIS*”では {NAISG, NAIST} の 2 件がヒットした．なお，この結果からわかることは，”NAIST”を 1-gram フレーズ匿名化する場合，”N*IST”や”NA*ST”ではリストを参照すると元の文字列が NAIST であると特定可能なことである．再特定のリスクが高いことがわかる．

現在，パラメータを $k = 3$ と設定しているため，”NAIST”の匿名化として 3 件以上ヒットした”*AIST”を採用する．

(4) 秘匿情報文字列の置換

入力文書 $d(\in D)$ における秘匿情報文字列 s_i を匿名後文字列配列 a_i に置換する． 全ての入力文書に対して処理を行い，匿名化済みデータとして出力する．

例 入力文書中の”NAST”を”*AIST”に置換し，匿名化済みデータとして出力する．

4. おわりに

本研究では，参照リストとのパターンマッチングを用いた部分文字列の k-匿名化手法を提案した．残存文字数を確保するために，本手法では秘匿対象文字列の部分文字列を黒塗り (redaction) する．ただし，残存する文字列による秘匿対象文字列の再特定を防ぐため，本手法では参照リストとのパターンマッチングにおいて k-匿名性を満たすように黒塗り箇所を選定するよう設計している．

本稿では，提案手法の構成について記述した．今後の課題として，手法の実装，パラメータ設定の考察および有効性の確認がある．そのため，パラメータの値による残存文字数の変化を測定することと，秘匿情報保護については，残存文字列による再特定のリスクを抑えることができているかを測定する．

謝辞

本研究の一部は NAIST ビッグデータプロジェクトによるものである．

参考文献

[1] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, Vol. 10, No. 1, p. 1, 2010.

[2] 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.

[3] 総務省. 情報流通・蓄積量の計測手法の検討に係る調査研究. 2013.