確率的潜在意味解析による集団匿名化法における情報損失と実質的個人識別リスクの評価

○山下真一郎[1] 本村陽一[1,2] 櫻井瑛一[2] 竹中毅[3]

[1]東京工業大学 [2]産業技術総合研究所サービス工学研究センター [3]経済産業省

yamashita.s.ae@m.titech.ac.jp

概要 トピックモデルの一種である確率的潜在意味解析(pLSA)を用いて集団匿名化することで情報損失を少なくしつつ，パーソナルデータを利活用するための実質的個人識別を不可能にする手法を提案する．また情報損失を平均情報エントロピーを用いて定義し，属性の一般化による情報損失とpLSAによる情報損失を比較した．

キーワード **確率的潜在意味解析**，情報損失，実質的個人識別

1 はじめに

顧客 ID を持つ POS システムや共通ポイントカード，電子マネーなどの普及によって大量の購買履歴や行動履歴が ID とともに集積される時代が到来した．こうした ID 付きの大量データを利活用し，経営や利便性などに役立つ有望な知見を抽出することが大いに期待されている．しかしその一方で個人情報が漏洩し悪用された場合の社会的影響は深刻であり，プライバシー保護の観点から従来は個人情報保護法による保護が求められており，その場合には氏名への到達可能性の有無が主要な論点であった．そのため顧客リストから個人名や電話番号などの個人を特定可能な属性(識別子)のみを消去する(単純匿名化)による対応が行われてきた．しかし近年，氏名には到達しないが個人を識別しうる実質的個人識別性という概念がプライバシー保護を必要とする大規模データ解析の判断基準として議論され始めている[1]．そこでは年齢や性別などの属性の組み合わせから，個人情報でなくてもデータから個人が識別可能になることを問題にしている．この問題に対応するための規準として，*k*-匿名性がある[2]．これはデータを集計することで集団匿名化し，集計結果の最小単位が k 人(k>1)であることで実質的個人識別を不可能にできることを保証する．ただし，この際個人識別の可能性が低くなると同時にトレードオフとして情報損失が問題になる．そこで言語処理分野で用いられるクラスタリング手法の一種である確率的潜在意味解析を用いて集団匿名化することで安全にパーソナルデータを利活用するための実質的個人識別を不可能にする手法を提案する．そして情報損失を比較する実験を行う．

2 提案手法

提案手法ではデータとして企業内 ID で連結された顧客の属性情報(ID，年齢，郵便番号，性別)を含む顧客リストと顧客が購買した店舗の情報を集計した購買店舗履歴があることを想定している．従来の k-匿名性では顧客リストの属性を一般化することにより集団匿名化する．本研究では購買店舗履歴を基にして確率的潜在意味解析により集団匿名化し，実質的個人識別を不可能にする．

2.1 確率的潜在意味解析

確率的潜在意味解析(以降 pLSA：*probabilistic Latent Semantic Analysis*)とは，当初自然言語処理分野で文書と単語の共起頻度から潜在的なトピックを抽出する手法として T. Hofmann により提唱された[3]．

文書 $d=\{d_1,d_2,\ldots,d_M\}$，単語 $w=\{w_1,w_2,\ldots,w_N\}$，話題 $c=\{c_1,c_2,\ldots. c_K\}$としたとき，文書 dと単語 wの間の関係は文書 d が与えられた時の話題 c である確率 $P(c \mid d)$ と話題 c が与えられたときの単語 w である確率 $P(w \mid c)$ で表される．これらの関係はベイズの公式を用いた変形によって式(1)と表現される．

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) \qquad (1)$$

$P(d|c)$と$P(w|c)$は EM アルゴリズムによって計算する．本研究では文書d-単語wに潜む話題cではなく，顧客(*user*)-店舗(*store*)に潜む関係(*segment*)を用いる．

2.2 情報損失

通常，集団匿名化を行うと情報損失が発生する．本研究では *user* を *segment* 毎に集団匿名化した際の情報損失 Z を式(2)~式(4)で定義する．

この時，ユーザをセグメント毎に分割する手法として確

率的潜在意味解析を用いることで情報損失をできるだけ少なくする.

$$Z = H' - H \tag{2}$$

$$H' = -\sum_{d}\sum_{w} P(w|c(d)log_2 P(w|c(d)) \tag{3}$$

$$H = -\sum_{d}\sum_{w} P(w|d)log_2 P(w|d) \tag{4}$$

2.3 実質的個人識別リスク

ユーザをセグメント毎に集団匿名化した際に, それぞれのセグメントに入っているユーザの最小単位がk人であるとき実質的個人識別リスク R は式(5)で定義する.

$$R = \frac{1}{k} \times 100 \quad [\%] \tag{5}$$

3 情報損失比較実験

本研究では大規模ショッピングモールで蓄積された顧客 ID, 年齢, 郵便番号, 性別の 4 つの属性が含まれた顧客リストと顧客が訪問した店舗について記録された顧客-店舗の共起行列データを使用して情報損失比較実験を行う. 情報損失が少ない手法がより情報の有用性を残し, かつ個人識別を不可能にする手法であると言える.

3.1 実験 1 *k*-匿名性が同じ場合

まず顧客リストの属性の一般化によって, *k*-匿名性(*k*=2, 3)を満たすように顧客の集団匿名化を行う. 次に顧客-店舗の共起行列データを pLSA によって顧客リストを一般化した場合と同様の *k*-匿名性(*k*=2, 3)を満たすように顧客の集団匿名化する. そして顧客リストの集団匿名化とpLSA による集団匿名化においてセグメント数と情報損失を比較する.

3.2 実験 2 セグメント数が同じ場合

k-匿名性(*k*>1)を満たすように顧客リストの属性の一般化による集団匿名化を行う. 次に顧客リストを集団匿名化した場合のセグメント数と同数のセグメント数(50)で購買店舗履歴を pLSA によって集団匿名化する. そして顧客リストの集団匿名化と pLSA による集団匿名化において実質的個人識別リスク R と情報損失を比較する.

3.3 実験結果

実験1及び実験 2 における結果を図 1~3 に示す. 実験結果より, pLSA を用いた顧客の集団匿名化の方が属性の一般化による集団匿名化よりも情報損失が少なくなった.

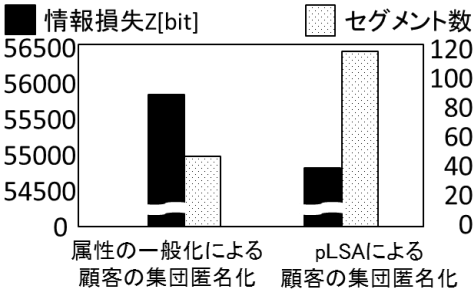

図1 実験1における R=50%の時のセグメント数と情報損失

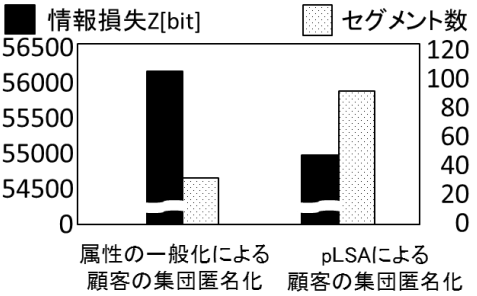

図2 実験1における R=33%の時のセグメント数と情報損失

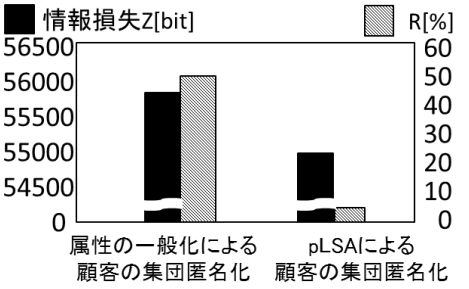

図3 実験2におけるセグメント数 50 の時の実質的個人識別リスク R と情報損失

4 おわりに

pLSA を用いた集団匿名化を提案し, 他の集団匿名化との情報損失を比較した. 本研究では、集団匿名化を目的として pLSA を用いたが, 顧客が利用する店舗の予測を目的として pLSA を用いる場合にも、情報損失があると仮定して、情報損失が小さい顧客セグメントを使うほうが良いと思われる.

今後のプライバシー保護の流れを受けて集団匿名化が避けられない場面においてできるだけ情報損失を少なくすることが, これからの大規模データの活用のためには重要である. 本研究で示した情報損失を防ぎ, それを評価する方法が大規模データを活用するための一つの枠組みとして役立つことが期待できる.

参考文献

[1] 総務省,「パーソナルデータの利用・流通に関する研究会」報告書(2013)

[2] *L.Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal on Uncertainly, Fuzziness and Knowledge-based Systems,* 10(5),*pp.*571-588 ,2002

[3] *T.Hofmann, probabilistic Latent Semantic Analysis, Proceeding,UAI'*99 *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp.*289-296, 1999