1L2-OS-17a-4

確率的潜在意味解析による集団匿名化法における来店店舗予測精度の評価

山下 真一郎[*1] 本村 陽一[*1*2]
Shinichiro Yamashita Yoichi Motomura

[*1] 東京工業大学 Tokyo Institute of Technology
[*2] 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

In this paper, we evaluated the coming shop prediction precision before and after to be anonymous group using probabilistic Latent Semantic Analysis.

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ている．このような 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．そこで山下ら[3]は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した．本研究では k-匿名性を満たすように確率的潜在意味解析を用いて集団匿名化された購買店舗履歴データを用いた来店店舗予測手法を提案し，その予測精度を集団匿名化を施さない元の実質的個人識別性のある購買店舗履歴データを用いた来店店舗予測精度と比較した．

2 手法

顧客が購買した店舗の情報を集計した購買店舗履歴（図1）があることを想定している．本研究では購買店舗履歴データを元データとし，確率的潜在意味解析により得られる潜在セグメント毎に個人データを集計することで集団匿名化し,その結果実質的個人識別が不可能なノンパーソナルデータを生成する．

購買店舗履歴				
ID	shop1	shop2	shop3	shop4
27	1	1	9	2
45	8	4	1	1
67	9	3	1	1
88	1	9	2	1
13	2	8	1	1
78	1	2	8	2
56	2	1	1	9
32	1	1	1	7

図1 購買店舗履歴 イメージ図

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，二種のデータ集合に含まれる共起関係を分析する次元圧縮・自動分類のためのアルゴリズムである．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[4]．

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 $P(c \mid d)$

連絡先：山下真一郎，産業技術総合研究所，東京都江東区青海 2-3-2，03-3599-8513，shinichiro.yamashita@aist.go.jp

と話題 c が与えられたときの単語 w である確率 $P(w|c)$ 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.

$$P(w,d) = \sum_{c} P(c)P(d|c)P(w|c) \quad (1)$$

またこの式の対数尤度関数は(2)式のようになる. これを最大化するような P(d|w) を探すことになるが, これには EM アルゴリズムを用いる.

$$\mathrm{L} = \sum_{w}\sum_{d} n(w,d)\log P(w,d) \quad (2)$$

カテゴリ数を一意に決定する必要がある場合の基準に次式に示す AIC(赤池情報量基準) を用いることとする. AIC が最小値をとる時のカテゴリ数を最適カテゴリ数とする.

$$AIC = -2lnL + 2k \quad (3)$$

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(*shop*)に潜む関係(*segment*)を用いる.

2.2 来店店舗予測精度

式(4), (5)で来店店舗予測精度 *CSPP* を定義する. ある *user* が所属する categoly の中で最も来店確率が高い *shop* とその *user* の最も来店確率が高い *shop* が同一だった場合, その *user* の$score_u$を 1 とする. そしてすべての *user* の$score_u$の和を来店店舗予測精度 *CSPP* とした.

$$CSPP = \frac{\sum_{all\,u} \; score_u}{N_{user}} \quad (4)$$

$$score_u = \begin{cases} 1 : \underset{s}{argmax}P(s|c_u) = \underset{s}{argmax}P(s|u) \\ 0 : \underset{s}{argmax}P(s|c_u) \neq \underset{s}{argmax}P(s|u) \end{cases} \quad (5)$$

但し$u: user \quad s: store \quad c: category \quad N_{user}$: ユーザ数

3. 来店店舗予測精度比較実験

3.1 実験データ

本実験では大規模ショッピングモールで蓄積されたデータを使用した. 集計したデータの情報を表1に示す.

表1

顧客数	27102 人
対象店舗数	188 店舗
取引総数	1612475 件

3.2 実験方法

Step1:

実質的個人識別性を有している顧客毎のパーソナルデータを生成し, このデータを与えた pLSA の実行によってクラスタリングした後, そのデータにおける来店店舗予測精度を求める.カテゴリ数は AIC を用いて決定する.

Step2:

2-匿名性を満たすように pLSA を用いて 実質的個人識別が不可能となるノンパーソナルデータを生成する. このデータをさらに 2 回目の pLSA の実行によってクラスタリングした後, そのデータにおける来店店舗予測精度を求める. カテゴリ数は Step1 のカテゴリ数と同一にする.

Step3 :

Step1 で得られた来店店舗予測精度と Step 2 で得られた来店店舗予測精度を比較し, 実質的個人識別が不可能となるようにノンパーソナルデータ化することで, どの程度予測精度が劣化するかを評価する. 実験結果は当日発表する.

4. おわりに

確率的潜在意味解析を用いた集団匿名化法における来店店舗予測精度の劣化を評価した. 集団匿名化によって実質的個人識別を不可能にすることが重要であると同時に集団匿名化されたデータにおける劣化を少なくすることが今後のパーソナルデータ利活用社会に求められる.

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] 山下真一郎,本村陽一,吉田真,竹中毅,“実質的個人識別を不可能にする情報損失の少ない集団匿名化法” ,行動計量学会,pp218-221,2013

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