

行政コールセンターへの問い合わせデータを対象とした住民の要望分析

川島 寛乃[†] 河野 慎[†] 米澤 拓郎[†] 中澤 仁[†]

[†] 慶應義塾大学 中澤研究室 〒 252-0882 神奈川県藤沢市遠藤 5322

E-mail: †{hirono,makora,takuro,jin}@ht.sfc.keio.ac.jp

あらまし 行政業務の効率化として住民の要望を把握するため、本研究では行政の運営するコールセンターに着目する。行政コールセンターにおける住民の要望を含んでいる問い合わせの記録より、それを抽出することを試みる。分析手法としては、問い合わせの背景にある潜在的な問題を探るためにトピックモデルである潜在ディリクレ配分法 (LDA) を用いる。このときある問い合わせはそれぞれ異なる単語分布をもつ複数の潜在的な問題から生成させると仮定し、モデルを適用させる。実験では藤沢市の市役所コールセンターの問い合わせ履歴のデータを用い、パープレキシティを評価指標として 1 つの文章として扱う範囲の粒度を変えながら潜在意味統計、潜在ディリクレ配分法で分析した結果を比較する。また、単語の出現分布をワードクラウドに可視化して比較し、市民の抱える問題の可視化を行う。キーワード 統計的潜在意味解析, トピックモデル, コールセンター履歴

1. はじめに

少子高齢化社会の到来により、働き手の不足や税金収入の低下が確実となっており、行政業務の効率化の実現は今後ますます重要となる。行政の効率化を達成するために重要な点の一つとして、住民の生活状況や要望を迅速かつ容易に把握することが挙げられる。これにより限りある行政の人的・資金的資源の最適配分を行い、効率の良い行政計画が可能となる。特に、行政業務の中でも重要な市民対応を効率よく行うためには、日々変化する住民の要望を把握し、住民が今後必要とするであろう情報を事前に住民に伝達したり、役所の住民対応窓口の職員人数を動的に変更する必要がある。現状、住民の要望を把握する手段としては、定期的な市民向けのアンケート調査や報告書などの統計結果 [27]、行政のホームページを経由した電子的な問い合わせ、役所もしくはコールセンターへの問い合わせ電話などが挙げられる。これらの複数の情報源から得られる住民の声を日々解析することにより、日々変わる住民の要望を把握・予測し、行政業務に活かすことが可能となると考えられる。

本研究では、住民の声が収集可能な情報源の中でも、特にコールセンターの問い合わせに着目する。コールセンターは自治体の委託によって開設され、一元的な住民の電話問い合わせ窓口として、日々住民からの質問に回答を行っている (図 1 参照)。自治体に関する情報は電子化が進み、自治体ホームページにも様々な情報が存在するが、求める情報にたどり着けない住民も多く、また情報機器の操作になれない高齢者の住民にとっては、電話での問い合わせは未だ重要な手段となっている。よって、コールセンターへの問い合わせ情報のトピック分析やトレンド分析を行うことで、幅広い住民層の要望の把握と、その要望の変化を学習することにより将来発生するであろう要望の予測が可能となると考えられる。また、コールセンターは住民にとって利便性のよい行政サービスではあるが、予算の問題から開設している自治体は限られており、例えば総務省による平成 22 年度の調査によると、全国約 1700 の自治体のうち 60

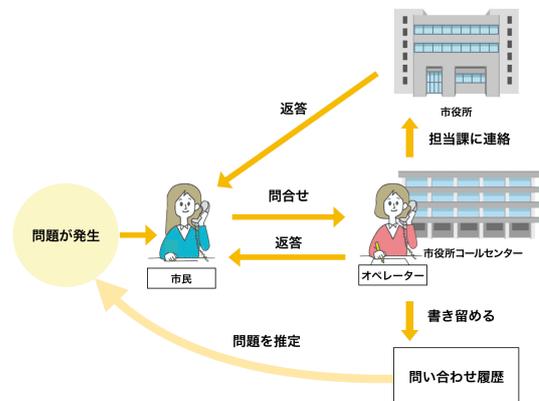


図 1: 問題に遭遇した際に市民は問い合わせ先として市役所コールセンターに連絡し、その内容をオペレーターが書き留めたものが問い合わせ履歴データとなる。本稿では問い合わせ履歴から市民が抱えている問題の推定を行う。

の自治体程度しか開設していない。よって、コールセンターが開設されている自治体の情報を分析し近隣の市区町村と共有可能とすることにより、コールセンターを開設していない自治体においても間接的に住民の要望を把握できる可能性が存在する。これらのことから行政のコールセンターにおける問い合わせ履歴の分析は重要であり、本研究が取り組む理由である。

これまでもコールセンターの分析の取り組みとして、特定の企業等における問い合わせを対象としたものが挙げられる [17] [20]。これら企業が設置するコールセンターにおいては、問い合わせの多くは特定の企業のサービスや商品についてが大半を占めている。従って、問い合わせの内容自体も使われる語彙単語も特異性の高いものになるため、限定された環境を対象とした手法が多い。一方、行政が設置するコールセンターの分析を深く行った例は、特に国内においては著者らが調べた限り存在していない。住民からの問い合わせ内容は、製品などを扱う企業とは異なった問い合わせ内容・頻度であることが予測さ

れるため、行政が設置するコールセンター情報の分析をどのように行うべきか、は重要な課題となる。本研究では、対象データセットとして、藤沢市コールセンターの問い合わせ履歴を利用する。藤沢市コールセンターで記録される各問い合わせレコードでは、2017年10月の記録を参照すると、1件あたり平均して132文字と短い文章であり、1日当たりの平均問い合わせ件数は106件ほどである。これらの性質を持つ文章からトピック推定等を行う分析方法を構築することが、本論文における主な貢献である。

コールセンターが記録する文章の分析手法として、本研究ではトピックモデルである潜在ディリクレ配分法 (LDA) を用いる。LDAは、文章はそれぞれ異なる単語分布を持つ複数のトピックから成ると仮定した手法であり、本稿ではある問い合わせが複数の潜在的な問題から発生するとしてモデルを適用する。ここで出てくる問題は潜在的なものであり、それぞれの潜在的な問題における単語の出現分布を元に、その問題の本質を推定する。評価指標にはモデルの予測性能を表すパープレキシティを用いる。一方、上述したように、コールセンターで記録される問い合わせの記録は、それぞれ1件あたり132文字と短いという特徴がある。そこで、トピックモデルの1つの文章として扱う範囲について粒度を、1件を1つの文章として扱う、同じ日の問い合わせ全体を1つの文章として扱う、1か月の中で同じ時間帯の問い合わせ全体を1つの文章として扱う、というように変化させ、それぞれ結果を比較する。またさらにトピック数を変化させた場合のパープレキシティを比較する。以上の条件で実験を行った時、トピック数をあげるほどパープレキシティの値は低くなり、モデルの予測性能が高まったことが示された。

2. アプローチ

市役所コールセンターの問い合わせ履歴から市民の問題を発見するための分析手段として、本稿では統計的潜在意味解析である、潜在ディリクレ配分法 (LDA) [3] を用いる。

2.1 潜在ディリクレ配分法 (LDA)

潜在ディリクレ配分法 (latent dirichlet allocations: LDA) は1つの文章が複数のトピックから成ることを仮定した確率言語モデルであり、文章の内容を分析する際に有効な手法である [3]。本研究では、コールセンターへの各問い合わせは市民の生活上にある何らかの問題から発生すると見なし、LDAを適用させる。またこのときにLDAの考え方にに基づき、問い合わせに含まれる単語は、各問題の持つ単語分布に従うとする。以下で具体的な説明を述べる。

コールセンターへの問い合わせを D とし、そのうちの1件の問い合わせの文章を問い合わせ d として扱う。市民の問題 k における単語 v の出現確率を $\phi_{k,v}$ とし、単語の出現分布を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。問い合わせ d の i 番目の単語 $w_{d,i}$ はこの単語分布 ϕ_k に従って生成される。

また1つの文章は複数の問題から成るため、文章 d で問題 k が出現する確率を $\theta_d = (\theta_{d,1}, \theta_{d,2}, \dots, \theta_{d,k})$ で表す。 $w_{d,i}$ の問題は潜在変数 $z_{d,i}$ として示す。

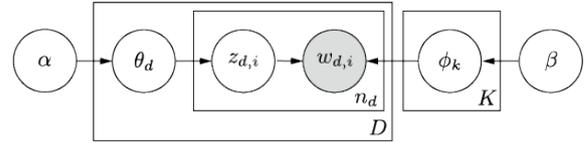


図 2: 問い合わせ履歴に適用する LDA のグラフィカルモデル。 θ_d, ϕ_k のパラメータを α, β とする。

このとき θ_d や ϕ_k は確率ベクトルであるので、確率ベクトル上の確率分布である Dirichlet 分布による生成を仮定すると、

$$\theta_d \sim \text{Dir}(\alpha) (d = 1, \dots, D), \phi_k \sim \text{Dir}(\beta) (k = 1, \dots, K) \quad (1)$$

と表せる。 $\alpha = (\alpha_1, \dots, \alpha_K)$ は K 次元ベクトル、 $\beta = (\beta_1, \dots, \beta_V)$ は V 次元ベクトルで、共に Dirichlet 分布のパラメータである。

単語 $w_{d,i}$ や問題 $z_{d,i}$ は多項分布を生成分布として仮定する。

$$z_{d,i} \sim \text{Multi}(\theta_d), w_{d,i} \sim \text{Multi}(\phi_{z_{d,i}}) (i = 1, \dots, n_d) \quad (2)$$

このグラフィカルモデルを図 2 に示す。

3. 実験

本稿の実験では、データセットに実際の藤沢市のコールセンターの問い合わせ履歴を使用する。1つの文章として扱う範囲の粒度を変えて分析を行う。分析手法には3章で述べたLDAを用いる。トピック数を変化させた際の評価指標には、モデルの予測性能を示すパープレキシティを用いる。

3.1 データセット

データセットには、実際の藤沢市のコールセンターの問い合わせ履歴を使用する。表1にその一例を示す。問い合わせ履歴には、問い合わせをはじめとした全部で8つの項目が含まれている。このデータの項目の一つである問い合わせ内容に、問い合わせ時の通話内容をコールセンターのオペレーターが書き留めた文章が記されている。またその他の項目で、受付種別、問題タイプ、完了区分、回答タイプはそれぞれ

- 受付種別: 電話のみ
- 問題タイプ: 2択 (問い合わせ・取次)
- 完了区分: 2択 (センター対応・エスカレ対応)
- 回答タイプ: 5択 (藤沢市 HP (FAQ)・藤沢市 HP (FAQ 以外)・センター内情報・所管轄対応・なし)

のように選択式で記入が可能となっている。1回の通話を1件としたとき、2016年1月から12月までの問い合わせの件数及び問い合わせの文章の単語数、1件辺りの単語数を、表2に示す。

3.2 評価指標

実験で扱うLDAに対し、モデルの予測性能を表すパープレキシティを使った評価を行う。パープレキシティは、あるトピックに基づいて文章中の単語を予測するとき、最も確からしい単語の選択肢の数を示す。すなわち、数が小さくなるほど、予測できる単語の数が絞られるため、モデルの予測性能を表す。まず学習用とテスト用に分割したデータの単語をそれぞれ w_d^{train} ,

表 1: 藤沢市役所コールセンター問い合わせ履歴データの一例.

受付日時	問題タイプ	完了区分	回答タイプ	関連所管轄	問い合わせ内容
2017/9/1 8:07	問い合わせ	センター対応	藤沢市 HP (FAQ 以外)		内線 2345 に繋いでください
2017/9/1 8:34	問い合わせ	センター対応	藤沢市 HP (FAQ 以外)		8 月 25 日に会社をやめたので、国民健康保険の加入方法を教えてください。退職の書類が昨日届いたので社会保険を辞めたかわかりません。
2017/9/1 8:39	取次ぎ	センター対応	センター内情報	市民自治防犯交通安全課	下水から水が漏れているようです。私は住宅管理会社の者です。対応は可能でしょうか。お客様ご本人様からご連絡するようにお伝えします。

表 2: 問い合わせ履歴データセットについて

種類	問い合わせ件数	文字数
10 月問い合わせ履歴	3,276	433,802
1 日の平均	106	13,994
1 件の平均	1	132

w_d^{test} , また文章 d の i 番目の単語を $w_{d,i}$ とするとき, モデル M を用いた予測分布を $p(w_{d,i}^{train} | M)$ として表すことができる. このとき, w_d^{test} の対数尤度は

$$\mathcal{L}[w^{test} | M] = \sum_{d=1}^D \sum_{w_{d,i} \in w_d^{test}} \log p(w_{d,i}^{test} | M) \quad (3)$$

となる. したがって, パープレキシティは,

$$perplexity[w^{test} | M] = \exp\left\{-\frac{\mathcal{L}[w^{test} | M]}{\sum_{d=1}^D n_d^{test}}\right\} \quad (4)$$

によって算出される.

3.3 結果

前述のデータセットと評価指標のパープレキシティを用い, 以下の 2 点の条件を変えて LDA を用いた比較実験を行った.

- トピック数
- 1 文書として扱う問い合わせの範囲

まずトピック数を $K = \{5, 10, 15, 20, 25, 50, 100, 150, 200, 250\}$ とした時のパープレキシティの変化を図 3 に示す. 次に, 1 つの文章として扱う範囲を 1 件の問い合わせ, 1 日毎の問い合わせ, 1 ヶ月毎の問い合わせ, 8:00~9:00・9:00~10:00 など 1 時間毎の問い合わせの 4 種類の粒度で区切って $K = 10$ として LDA を適用させた. その結果を図 4 に示す. またその際の各トピックの単語分布を表 3 に示す. 本稿では各トピックの単語出現確率が高いものの上位 5 件ずつを表にまとめ, 単語の左側に出現頻度を記す. 1 件の文書におけるトピック分布からもっとも確率の高いものをその文章のトピックとした時, 各トピックごとの該当する文章の件数は図 5 のようになる. 各トピックの単語分布に基づく単語出現回数をワードクラウドを用いて可視化したものを図 6 に示す.

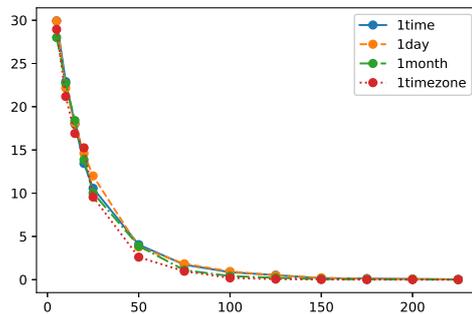


図 3: 1 件の問い合わせを 1 文章として扱ったとき, トピック数を変化させたときの LDA のパープレキシティの変化.

3.4 トピック数

モデルの予測性能を表すパープレキシティは, トピック数が増加するに従って, 減少している. つまり, トピック数を大きくするほどパープレキシティの値は下がり, 予測性能は良くなったと言える. しかしトピック数を増やした場合, 図 4 のように, 複数のトピックの単語分布が大きく被った. これは複数のトピックの持つ単語分布が近いということであり, 同じような単語分布からそれが何を表すトピックなのかを推定しなければならない. そのため複数のトピックの単語分布が大きく被る状態はトピックの推定にあたって不適である. そこで, 本稿では以降, トピック数を $K = 10$ および $K = 20$ として分析した.

3.5 1 文章として扱うデータの範囲

本稿ではコールセンターのデータは短文であることを考慮し, 様々な粒度でデータを区切って実験を行った. その結果トピック数を $K = 10$ としたときの図 4 (e)・(g) と同様に, 1 つの文章として扱うデータの範囲が 1 ヶ月, 1 時間帯である場合, 文章中のトピックの割合が 1 つ, もしくは 2 つになった. このトピックの割合の分布が, 文章におけるトピック分布である. 文章中の単語のトピックを考えた時, 任意のトピックの割合が大きくなった場合, 多くの単語が同じトピックの単語分布から生成されていることになる. 本来ならば異なるトピックに出現する確率が高い単語が同じトピックの単語分布に含まれると, ト

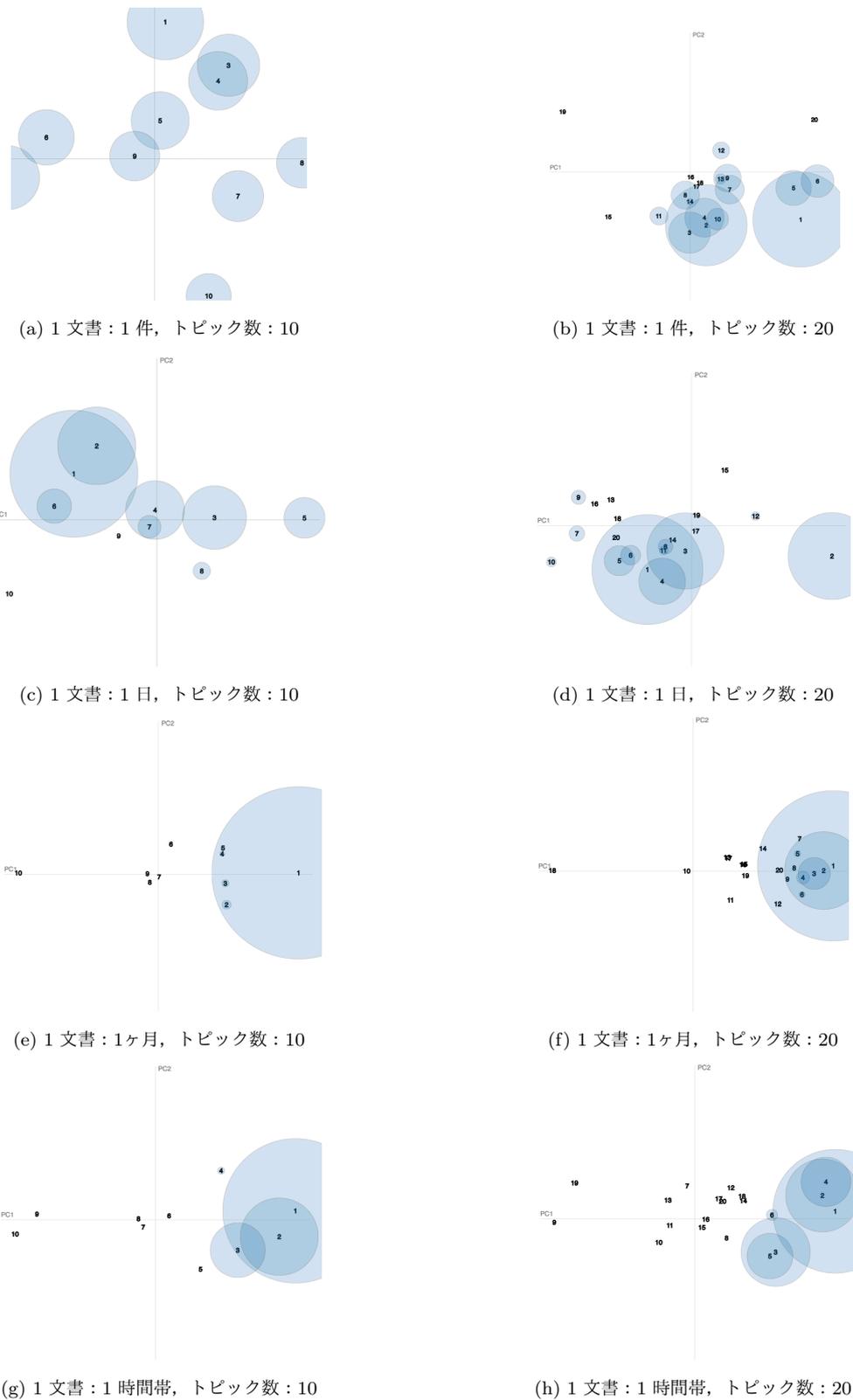


図 4: 1 文書として扱う範囲の粒度を変えた際の各トピックの分布図。

ピックが何を表しているのかを端的に理解することができない。そのため、トピックの推定することは困難となる。それを避けるため、本稿では実験結果より、問い合わせデータを扱う際の区切る大きさは、1 件及び 1 日の粒度が望ましいとする。

3.6 各トピックの単語分布

図 6 に示した複数のトピックのワードクラウドからは、各ト

ピックにおける重要単語を理解することができる。ここから、例えば (a) は藤沢・本籍・転入・申請・手続きなどが出現確率が高くなっているため、このトピックは住居に関連する問い合わせであると推定できる。また (d) は投票・期日・入場・センター・選挙などの出現確率が高くなっているため、このトピックは選挙を表していると推定できる。

表 3: 1 文章として取扱う問い合わせの範囲の粒度を変え、LDA で分析をした際の各トピックの頻出単語 5 つ (10topic)。

	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9
1 件 LDA (10)	0.029 確認	0.029 市	0.037, 的	0.056 投票	0.023 投票	0.033 課	0.035 市	0.027 課	0.016 投票	0.029 保険
	0.021 書	0.023 藤沢	0.029 課	0.025 者	0.021 市民	0.031 担当	0.030 藤沢	0.024 電話	0.015 券	0.023 確認
	0.018 証明	0.022 書	0.021 不	0.020 券	0.019 前	0.030 市	0.019 確認	0.019 後	0.014 前	0.023 証
	0.017 市民	0.018 課	0.018 完了	0.017 確認	0.016 期日	0.030 片瀬	0.018 月	0.019 着信	0.013 期日	0.022 手続き
	0.014 センター	0.018 確認	0.017 機械	0.017 所	0.016 センター	0.020 藤沢	0.018 住民	0.017 確認	0.013 整理	0.018 市
1 日 LDA (10)	0.031 投票	0.031 市民	0.028 投票	0.035 市民	0.034 市民	0.028 市民	0.025 男性	0.035 男性	0.031 投票	0.037 投票
	0.027 市民	0.027 横浜	0.019 前	0.030 男性	0.028 投票	0.021 横浜	0.025 市民	0.031 市民	0.024 市民	0.032 市民
	0.022 横浜	0.019 男性	0.017 券	0.029 横浜	0.028 横浜	0.019 入電	0.021 入電	0.025 入電	0.019 前	0.019 確認
	0.020 入電	0.019 入電	0.015 市民	0.027 入電	0.023 入電	0.019 投票	0.019 課	0.020 市	0.018 入電	0.018 男性
	0.016 藤沢	0.018 市	0.014 市	0.016 課	0.020 男性	0.018 確認	0.017 横浜	0.018 横浜	0.018 藤沢	0.017 入電
1ヶ月 LDA (10)	0.032 市民	0.035 市民	0.032 市民	0.029 入電	0.031 市民	0.033 市民	0.034 市民	0.030 市民	0.032 市民	0.025 市民
	0.023 横浜	0.026 投票	0.021 投票	0.027 投票	0.028 投票	0.025 男性	0.026 投票	0.023 投票	0.029 横浜	0.023 投票
	0.020 男性	0.024 入電	0.018 男性	0.025 横浜	0.019 市	0.025 投票	0.023 男性	0.022 横浜	0.020 男性	0.023 横浜
	0.018 投票	0.018 確認	0.018 入電	0.023 市民	0.019 入電	0.021 横浜	0.019 横浜	0.021 男性	0.020 投票	0.023 男性
	0.017 確認	0.017 市	0.017 横浜	0.020 男性	0.018 男性	0.021 入電	0.018 入電	0.016 確認	0.018 入電	0.017 確認
1 時間帯 LDA (10)	0.028 市民	0.023 市民	0.029 投票	0.018 市民	0.026 市民	0.028 市民	0.036 市民	0.023 市民	0.038 投票	0.036 市民
	0.024 投票	0.021 横浜	0.027 市民	0.016 横浜	0.019 横浜	0.027 男性	0.022 入電	0.020 入電	0.031 市民	0.026 横浜
	0.020 男性	0.019 投票	0.024 横浜	0.012 入電	0.017 投票	0.021 横浜	0.020 横浜	0.019 投票	0.024 入電	0.024 男性
	0.020 入電	0.018 市	0.020 男性	0.012 男性	0.016 入電	0.021 入電	0.020 男性	0.017 男性	0.021 男性	0.022 投票
	0.020 横浜	0.018 入電	0.018 入電	0.010 投票	0.015 確認	0.021 投票	0.018 投票	0.017 確認	0.021 横浜	0.021 入電

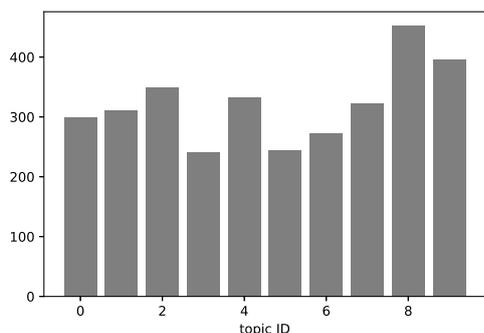


図 5: 各トピックに含まれる文章の件数

4. 考察

本章では前章で示した実験結果についての考察を行う。

まず、問い合わせを分析する上で重要となる、設定すべきトピック数について述べる。実験では 1 文章として扱う範囲を 1 件とし、トピック数を $K = 10$ とするのがふさわしいという結果に至った。また今回図 6 (c)・(d) のように、確認・入電・市民・センターなど問い合わせを要約した時に使用される言葉の出現確率が高く、各トピックの主要語となってしまうケースが発生した。市民の要望や問題点を抽出するために出現確率の高い主要語の影響を抑えるために、BIC [18] やバイクラスタリング [13] などの手法が存在する。これらを用いることでモデルの質の測り方を工夫し、トピック数を最小限にすることが今後の課題となると言える。

次に、推定したトピックの単語分布から得られる市民の要望について述べる。本実験では LDA の分析結果によって得られた単語分布が示すトピックの命名を行った。例えば図 6 では (a) 住居, (b) 施設利用, (c) 手続き方法, (d) 選挙などのト

ピックがあった。

(d) 選挙は昨年 10 月 22 日 (日) に衆議院総選挙が行われたことにより、その会場や投票などに関する情報を求める要望が多くあったために出現したトピックであると考えられる。特に単語分布の上位には、期日前投票の「期日」や、投票方法に関係する「入場」「整理券」などの単語があり、それらについての問い合わせが多かったことが伺える。

一方で、本稿ではデータセットの各データに対して LDA を用いた分析を試みたが、これは 1ヶ月の問い合わせデータに対する静的な分析結果にすぎない。そのため、時間的なトピックの推移を追ったり [2], 1ヶ月以上の範囲で月ごとや季節など長期の周期性を持つ住民の要望を抽出することはできていない。

そのため、トピックの追跡や時系列を考慮したモデルの構築、および扱うデータの範囲の拡張などが今後の研究の課題となる。

5. 関連研究

この章では関連研究として、統計的意味解析手法を使用した研究及び市民・顧客の声を解析した研究について述べる。

5.1 LDA を用いた文章解析

近年、様々な分野で統計的解析手法が使われており、現在文章の意味を解析するための手法としてはトピックモデルの LDA が多く用いられている。潜在意味解析 (LSA) は 1980 年代後半に検索分野で登場した [5] [15]。これは文章と語彙の共起行列を特異値分解によって近ランク近似行列することで、同義語の検索を可能にしたものである。しかし特異ベクトルが何を示しているのか解釈がわかりにくく、またトピックは軸が直行するように形成されるため、制約として強すぎるなどの問題が生じた。そこでこれらの問題を解決するため、1999 年に Hoffman によって確率モデル PLSI として再定式化された [10]。PLSI では文章ごとに複数のトピックを持つ、トピックごとに異なる単



(a) topic1: 住居

(b) topic4: 施設利用



(c) topic6: 手続き方法

(d) topic8: 選挙

図 6: 1 件の問い合わせを 1 文書、トピック数を 10 としたときの各トピックの単語分布の一例

語の生成分布を持つという性質を仮定したモデルである。さらに 2003 年以降 PLIS に階層ベイズモデル (hierarchical Bayes model) による拡張をした潜在ディリクレ法 (LDA) が提案され [3], 多くの研究分野で用いられる手法となった。LDA は推論時に変分ベイズを用いる手法が提案されていたが、その後ギブスサンプリングを用いた推論 [8] や変分ベイズの精度を向上させた手法 [1] [23] など提案された。またオンラインで入力されるような文章をデータとする時の扱い [24] や LDA の教師あり学習 [9] [12] [16] などについても研究が進められている。

現在ではこれらの手法は定着化しており、モデルの拡張や構造の変更によってより分野に適用させることを目的とした研究が多く進められている。LDA を用いた分析の研究としては、小説や論文の著者推定 [14] [21], Twitter のツイートからのユーザーの興味推定 [19] など取り組まれている。また、画像の分類などに使われている例もあり、これは画像を文章に落とし込むことでトピックの推定を可能にしている [4] [6]。通常 LDA では何か 1 つの文章のまとまりを 1 文章として取り扱う場合が多いが、本研究の問い合わせのデータのように 1 件あたりの文字数が比較的少ない場合には、複数件を 1 つの文章として扱うなどの工夫がされる場合がある [25]。本研究では 1 文章として扱う範囲を粒度を変えて実験を行い、それぞれ結果を比較した。

5.2 市民・顧客の声の分析

企業や団体において、満足度の向上に繋げるための顧客の声の分析は古くから取り組まれている [7]。分析対象としてはアン

ケートや調査の結果が主流であったが [11] [22] が、顧客からの問い合わせの窓口であるコールセンターに着目し、問い合わせのデータから顧客の抱えている問題を発見しようとする研究も成されている。コールセンターの問い合わせデータを用いた研究の一例としては、顧客からの電話をどの担当部署が受け取るべきかを判別する研究 [20] や、キーワードの分布から文章の類似度を計算する研究 [17] などが挙げられる。また、含まれる語句のうち重要性が高いとみなされる語句を抽出し、その語句間の関係性をマップとして表示することで、大量に記述されている概念の構造の視覚化が取り組まれている [26]。しかし、これらの研究はいずれもルールベースに基づく自然言語処理の研究の延長として進められており、コールセンターの環境に基づいて作成されたキーワードの辞書をベースに解析を行なっている。そのためモデルの汎用性は低く限定的で、環境ごとに辞書を作成して調整を行わなければならない。

本研究はこれらの関連研究と同じように、問い合わせの内容を分析するという目的を持っているが、アプローチ方法としてルールベースの分析手法は用いず、確率モデルを用いた。それゆえ分野や規模を問わず、様々な環境下のコールセンターの問い合わせデータについて一様に分析を行うことが可能となっている。また市民の声の集まる市役所のコールセンターのデータを扱っているため、都市や市町村を超えた共通性が存在し、複数の都市や市町村のデータを用いることができた場合も同じモデルで分析を行い、結果を比較することが可能である。

6. ま と め

本研究では市民の声から問題を発見する過程で市役所コールセンターの問い合わせ履歴に着目し、アンケートや世論調査では捉えきれない問題を見つけ出すことを目的とする。問い合わせ履歴の分析には LDA を用い、実装したモデルはパープレキシティを評価基準として比較した。また問い合わせの内容は短文であるため、実験では一つの文章として扱う長さを 1 件・1 日・1 ヶ月・1 時間帯ごとに変化させて結果を出力した。このとき結果の考察から、トピック数を増やすと予測性能は上がるが各トピックの単語分布が被ってしまうこと、1 文章として扱う問い合わせの範囲は 1 件及び 1 日の問い合わせ全体が適していること、各トピックの単語分布からはそれぞれ重要単語が抽出できることなどを理解することができた。

今後の課題としては、トピックの追跡や時系列を考慮したモデルの構築、および扱うデータの範囲の拡張などが挙げられる。また今後、行政コールセンターの特徴として挙げられる全国での共通性を生かし、1 つのコールセンターの問い合わせ分析結果から他の各都市や自治体においても適用することのできる部分を見出すことを目指したい。

謝 辞

本研究の一部は国立研究開発法人理化学研究所に支援頂いた。

文 献

- [1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34. AUAI Press, 2009.
- [2] David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, Vol. 10, No. 71, p. 34, 2009.
- [3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.
- [4] Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1903–1910. IEEE, 2009.
- [5] Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 281–285. Acm, 1988.
- [6] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 2, pp. 524–531. IEEE, 2005.
- [7] Abbie Griffin and John R Hauser. The voice of the customer. *Marketing science*, Vol. 12, No. 1, pp. 1–27, 1993.
- [8] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, Vol. 101, No. suppl 1, pp. 5228–5235, 2004.
- [9] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–274. ACM, 2009.
- [10] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999.
- [11] Dawn Iacobucci, Amy Ostrom, and Kent Grayson. Distinguishing service quality and customer satisfaction: the voice of the consumer. *Journal of consumer psychology*, Vol. 4, No. 3, pp. 277–303, 1995.
- [12] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In *Advances in Neural Information Processing Systems*, pp. 835–843, 2009.
- [13] Bhargav Kanagal and Vikas Sindhwani. Rank selection in low-rank matrix approximations: A study of cross-validation for nmfs. In *Proc Conf Adv Neural Inf Process*, Vol. 1, pp. 10–15, 2010.
- [14] Marie Katsurai, Ikki Ohmukai, and Hideaki Takeda. Topic representation of researchers’ interests in a large-scale academic database and its application to author disambiguation. *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 4, pp. 1010–1018, 2016.
- [15] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, Vol. 25, No. 2-3, pp. 259–284, 1998.
- [16] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128, 2008.
- [17] Dunja Mladenic. Text-learning and related intelligent agents: a survey. *IEEE intelligent systems and their applications*, Vol. 14, No. 4, pp. 44–54, 1999.
- [18] Art B Owen, Patrick O Perry, et al. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics*, Vol. 3, No. 2, pp. 564–594, 2009.
- [19] Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. Online topic model for twitter considering dynamics of user interests and topic trends. In *EMNLP*, pp. 1977–1985, 2014.
- [20] Hideo Shimazu, Akihiro Shibata, and Katsumi Nihei. Case-based retrieval interface adapted to customer-initiated dialogues in help desk operations. In *AAAI*, Vol. 1994, pp. 513–518, 1994.
- [21] Masato Shirai and Takao Miura. On domain independence of author identification. *Intelligent Data Engineering and Automated Learning-IDEAL 2011*, pp. 9–16, 2011.
- [22] Robert Teehan and Walter Tucker. A simplified lean method to capture customer voice. *International Journal of Quality and Service Sciences*, Vol. 2, No. 2, pp. 175–188, 2010.
- [23] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pp. 1353–1360, 2007.
- [24] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 937–946. ACM, 2009.
- [25] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pp. 338–349. Springer, 2011.
- [26] 岡本青史, 関口実, 三末和男, 西野文人. カスタマーセンター支援システム. *人工知能学会誌* Vol.15, pp. 1027–1034, 2000.
- [27] 渡邊勉. 地域に対する肯定観の規定因: 愛着度, 住みやすさ, 地域イメージに関する分析. *地域ブランド研究*, Vol. 2, pp. 99–130, 2006.