テキストデータの理論的サンプリング

Theoretical Sampling of Text Data

松村 真宏
Naohiro Matsumura
大阪大学大学院経済学研究科
Graduate School of Economics, Osaka University

1. はじめに

本稿では，テキストマイニングの分析対象となるテキストデータの分量について考察する．これまでは基本的に「多ければ多いほどよい」と見なされてきたので，入手できるデータは全て使うことが前提であり，分量が多すぎるときはサンプリングして使われることが多かった．しかし，分析に必要なテキストの分量を見積もることができれば，データ収集およびデータ分析のコストを最適化できる可能性がある．そこで本稿では，テキストデータの理論的サンプリングに関する基礎的考察を行う．

2. 理論的サンプリング

文献や面接や自由記述文といった質的データを対象として，特定の概念を表すカテゴリを抽出して体系化するアプローチに質的研究法がある．カテゴリが出尽くした状態が「飽和」したと判断され，例えば 1 テーマにつき同質の対象者を 20 人から 30 人程度に面接すれば飽和すると言われている．このように，カテゴリが飽和するまでサンプルを逐次的に追加するアプローチは「理論的サンプリング」と呼ばれる．

理論的サンプリングの考え方は，Glaser と Strauss がデータから理論を生成するための方法論として提唱したグラウンデッド・セオリー・アプローチ (Grounded Theory Approach; GTA) において提案されたものであり [1]，豊田は理論的飽和や理論的サンプリングに数理的表現を与えた [2]．本稿では，豊田の提案した数理モデルに基づいて，テキストマイニングの観点からテキストデータの理論的サンプリングを試みる．

3. テキストデータの理論的サンプリング

本稿では，テキストデータを増やしても新たな語種が現れなくなったときをテキストデータが「理論的に飽和」した状態であるとみなす．また，母集団の語種における既出の語種の割合を「理論的飽和度」とする．

理論的飽和度を求めるためには母集団における語種数が推定できればよい．豊田 [2] に基づくと，母集団の語種数は以下の手続きで推定できる．$i-1$ 回目までのサンプリングで得た語種数を m_{i-1}，i 回目のサンプリングで得た語種数を c_i，i 回目のサンプリングで再び得られた語種数（$i-1$ 枚目までのサンプリングで得られた語種と重複する語種数）を r_i とする．このとき，母集団における語種数を N_i とすると，$N : m_{i-1} = c_i : r_i$ の関係より N_i は以下の式で推定できる．

$$N_i = m_{i-1}c_i/r_i \tag{1}$$

この N_i は $r_i = 0$ のときに定義できないので，ピーターセンの修正式により母集団の語種数の推定値 $\hat{N}_i$ を以下の式で求める．

$$\hat{N}_i = m_{i-1}(c_i+1)/(r_i+1) \tag{2}$$

さらに，複数回のサンプリング結果を用いて平滑化するシュナーベル法により，母集団の語種数の推定値 $\hat{\hat{N}}_i$ は以下の式で推定できる．

$$\hat{\hat{N}}_i = \sum_i r_i^* \hat{N}_i \tag{3}$$

$$r_i^* = r_i / \sum_j r_j \tag{4}$$

ここで $\hat{\hat{N}}_i$ の分散を $V[\hat{\hat{N}}_i]$ とすると，95%信頼区間の上側限界 $\hat{\hat{N}}_i^{upper}$ は以下の式で表される．($\hat{N}_i$ の 95%信頼区間の上側限界 $\hat{N}_i^{upper}$ も同様に求まる)

$$\hat{\hat{N}}_i^{upper} = \hat{\hat{N}}_i + 1.96 \times \sqrt{V[\hat{\hat{N}}_i]} \tag{5}$$

この上側限界を使えば，サンプリングの理論的飽和度は推定値の 95%上側限界値に対する実測値の割合として以下の式で表される．

$$(m_{i-1} + c_i - r_i)/\hat{\hat{N}}_i^{upper} \tag{6}$$

この理論的飽和度を基準にしてテキストデータの必要量を見積もることを「テキストデータの理論的サンプリング」と呼ぶことにする．

4. 実験と考察

「iPhone」を含む日本語のツイートデータ[1]から 100 ツイートずつランダムサンプリングして，形容詞の語種数の実測値と母集団の推定値を求めた結果を図 1 に示す．形態素解析器には MeCab[2] を用いた．横軸はサンプル数，縦軸は語種数であり，実測値，$\hat{N}_i^{upper}$，$\hat{\hat{N}}_i^{upper}$ をプロットしている[3]．

理想的には母集団の予測値は一定であることが望ましいが，図 1 を見るとサンプル数の増加に従って母集団の予測値も増加していることがわかる．また，サンプル数が増えるに従って実測値および予測値は収束に向かっていることもわかる．また，$\hat{\hat{N}}$ のほうがスムーズな推定結果になっていることもわかる．

---

[1] TTC (http://mtmr.jp/ttc/) を用いて収集した「iPhone」を含む 14,289 件の日本語ツイートデータ．http://mtmr.jp/data/20111005--iPhone.zip から取得可能．

[2] http://mecab.sourceforge.net/

[3] シュナーベル法および分散を求める際には，各時点から遡った 10 回のサンプルを対象とした．

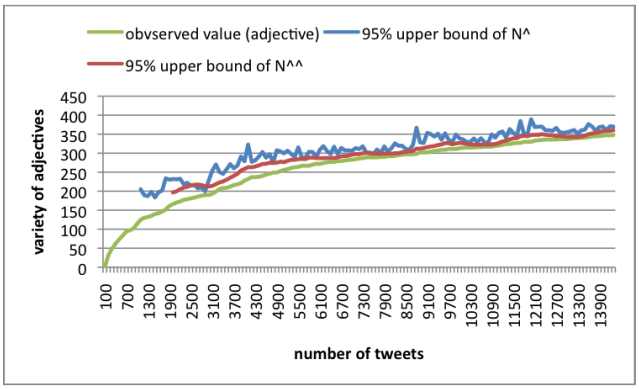

図 1: 形容詞の語種数の実測値と母集団の推定値

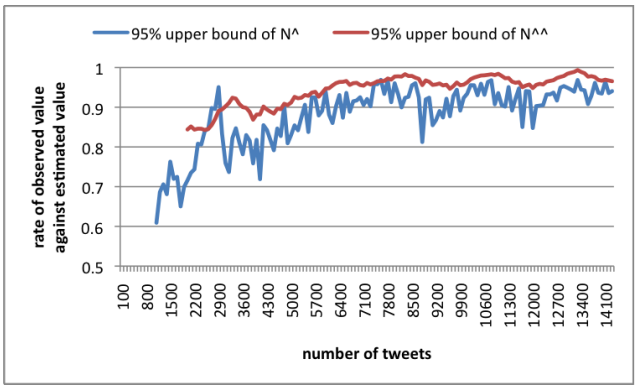

図 2: 形容詞の語種数の理論的飽和度

次に，理論的飽和度のグラフを図 2 に示す．横軸はサンプル数，縦軸は理論的飽和度である．図 2 をみると，$\hat{N}$ ではサンプル数が 5,000 件を超えたあたりから安定して 90%を超えていることがわかる．したがって，理論的サンプリングによると，このデータセットから 90%以上の形容詞を抽出したいときには，5,000 件サンプリングすればよいと判断できる．

しかし，5,000 件サンプリングした時点での形容詞の語種数の実測値は 249，母集団の推定値は $\hat{N}_{50}^{upper} = 275$，$\hat{\hat{N}}_{50}^{upper} = 262$ であるが，14,300 件サンプリングした時点での語種数の実測値は 348 である．仮にこれが形容詞の真の母集団だとしても 5,000 件サンプリングしたときの飽和率は 71.6%にすぎない．したがって，理論的飽和度と実際の飽和度には大きなズレがあることがわかる．

この原因はいろいろ考えられるが，主な要因としては語の出現頻度が等確率ではないことが考えられる．具体的には語の出現頻度は Zipf の法則に従うことが知られている．また，語と語の生起確率も互いに独立ではないことも原因の一つだと考えられる．

今回は形容詞を対象として実験を行ったが，形容詞は語種数が少ないために収束しやすい．今回用いたデータセットの他の品詞の語種数の推移（図 3 参照）を見ると，形容詞や副詞は収束に向かっているが，名詞は伸び続けており収束する気配がない．名詞はテキストマイニングにおいて重要な品詞であるので，名詞を網羅的に収集するためにはより多くのサンプル数が必要であろう．

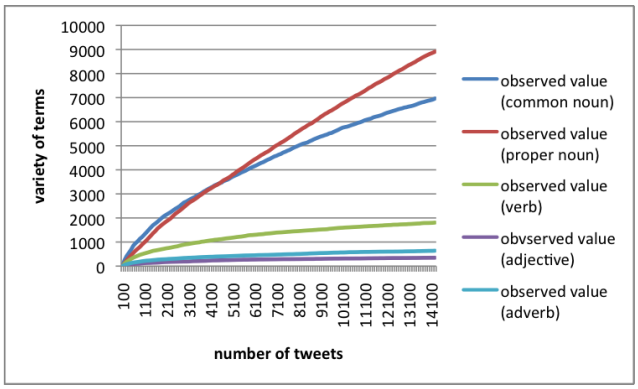

図 3: 品種別の語種数の推移

なお，豊田 [2] でも述べられているが，ここで得られた理論的飽和率は「同じ結果が得られるという意味で、信頼性に対応する概念」であり,「本来欲しかった知識の集まりの何割を有したかという意味で妥当性」にはならないことに注意する必要がある．

5. まとめ

本稿では，テキストデータの理論的サンプリングに関する基礎的考察を行った．実測値と推定値との間にズレはあるが，理論的飽和度が分析に必要なテキストデータの分量の目安になることを示した．

質的研究で抽出対象となるカテゴリー数に比べて，テキストマイニングで抽出対象となる語種数（特に名詞）は遥かに多いので，飽和状態に至るにはより大きなデータセットを必要とする．限られたデータセットの中から語の出現傾向や評判情報といった大局的な知見を得ることがテキストマイニングのアプローチであり，飽和状態に至るデータセットは必ずしも必要ではない．しかし，分析対象のデータセットの飽和状態を知ることは，得られる知見の適用範囲を知る上で重要であろう．

豊田 [2] は等確率性を仮定しない微分方程式によるモデルも提案しているが，テキストデータの性質を考慮すれば Zipf の法則を組み込んだモデルも考えられる．また，語種数はトピック毎に異なると考えられるが，一般的なデータセットを用いて理論的飽和度を予め求めておけば，テキストマイニングに必要なテキストデータの分量の一般的な目安が得られるだろう．

参考文献

[1] Glaser, B.G., Strauss, A.L.: The Discovery of Grounded Theory, Strategies for Qualitative Research, ALDINE PUB. Co., pp. 45–77 (1967)

[2] 豊田秀樹：質的研究の理論的サンプリングにおける理論的飽和度，日本教育心理学会第 53 回総会自主企画 25-J-01 (2011)