テキストデータの収集の程度を測る指標に関する研究

大橋洸太郎

目 次

第 1 章　序章 **1**
1.1　問題 1
1.2　研究目的 1

第 2 章　自由記述のカテゴリ化に伴う観点の飽和度としての捕獲率 **2**
2.1　はじめに 2
2.2　方法 2
2.3　適用例 4
2.3.1　データ 4
2.3.2　結果 4

第 3 章　自由記述における単語の種類数の捕獲率 **5**
3.1　はじめに 5
3.2　方法 5
3.2.1　DeLury 法 5
3.3　適用例 6
3.3.1　データ 6
3.4　結果 6

第 4 章　自由記述における汲みつくしの指標としての遭遇率の提案—パレート分布を用いた分析— **7**
4.1　はじめに 7
4.2　方法 7
4.2.1　頻度を説明するモデル 7
4.2.2　各種パレート分布の PDF 8
4.2.3　遭遇率 8
4.2.4　データと加工 8
4.2.5　推定方法 9
4.3　結果 9

第 5 章　Web 文書収集の飽和度の計算と Web 探索 **10**
5.1　はじめに 10

5.2 方法 10
5.2.1 DeLury 法における分散と s.e. 10
5.3 研究 1：分析の手順 11
5.4 研究 1：結果と考察 11
5.5 研究 2：分析の手順 12
5.6 研究 2：結果と考察 12

第 6 章 授業評価における知見収集の飽和度 13
6.1 はじめに 13
6.2 方法 13
6.2.1 用いたデータ 13
6.2.2 データの加工方法 13
6.2.3 得られた知見の分析方法 14
6.3 結果 14
6.3.1 “授業 F”の知見 14
6.3.2 捕獲率の計算結果 14

第 7 章 就職活動の成功体験に関する知見の収集と分析 15
7.1 はじめに 15
7.2 方法 15
7.2.1 データの収集 15
7.2.2 データの処理 15
7.2.3 記事からの知見の抽出 16
7.3 結果 16
7.3.1 抽出された知見 16
7.3.2 Schnabel 法による捕獲率の計算結果 16

第 8 章 総合考察 17

引用文献 18

第1章　序章

1.1　問題

テキストデータを収集する目的は，調査仮説の論点に関する知見を充足させることにある。しかしながら実際の調査においては，テキストを収集し続けた場合に，その内容をどこまで読み進めたとしても新しい知見や観点が得られなくなるということは稀であり，知見の完全な充足には果てがない。このため研究者はデータ収集のある段階で，“課題に関する論点は十分に収集された”と判断し，収集を打ち切る必要に迫られる。これには 1.“課題に関する論点は十分収集された”という基準は研究者間で同一とは限らない，2. 研究の成果を読む査読者や一般の読者達は，研究者と同程度にはテキストデータ収集の飽和の程度を理解することは難しい，という 2 つの問題がある。

これらの問題を解決するためには，テキストにおける知見収集の飽和度について客観的な指標が不可欠であると考える。テキストデータ収集の飽和度の指標が開発されれば，研究者は客観的な指標を元にデータ収集を打ち切ることが可能となり，手元のデータが，その後の分析を始めるに当たって十分であるかを容易に，そして客観的に判断すること可能となる。以上より飽和度の指標の開発は，研究者，読者の両者に有益な結果をもたらすと考えた。

1.2　研究目的

本稿の目的はテキストデータ収集の飽和度を示す指標“捕獲率”の開発，提案を目的とする。この指標がテキスト資料を扱う分析に広く応用可能であることを示すために，1. 様々な標本サイズにおいて適用可能な方法の開発，2. 様々なテキストデータの処理方法に対応した応用例の提示，3. 様々な種類のテキストデータへの応用の 3 点に考慮して提案，応用を試みた。

第2章　自由記述のカテゴリ化に伴う観点の飽和度としての捕獲率

2.1　はじめに

第 2 章は捕獲率についての提案が初めてなされる章である。自由記述の収集が十分さの程度を測る方法として，本章では捕獲率が提案される。

2.2　方法

以下に表記法を示す。

1. 捕獲回数 (知見を得る機会の回数) I 回。添え字は $i(i = 1, \cdots, i, \cdots, I)$ を使用。i 回目の捕獲にて
2. 印を標識 (mark) と呼び, 標識の付いた知見数 (既知となった知見の数) は m_i $(m_1 = 0)$
3. 捕獲 (capture) した知見数 (既知, 未知に関わらず, i 回目の捕獲において得られた知見の総数) は c_i
4. 再捕獲 (recapture) された知見数 (既に得られている知見の数) は r_i $(r_1 = 0)$
5. 新しく (new) 捕獲された知見数 (これまでになかった未知であった知見の数) $n_i = c_i - r_i$ $(n_1 = 0, m_{i+1} = m_i + n_i)$
6. 標識率 (i 回目の捕獲において, 既知となっている知見の割合) $p_i = r_i/c_i$

また捕獲回数 I は，研究者が限りある研究資源 (時間や研究費) の中で捕獲率の計算を試みた時点での，手元のデータでの全捕獲回数を指している。

本手法を用いる目的は知見 (資源・魚・カード) の総数 N を推定することであり, それを利用して捕獲率 (飽和度)$Cr = m_I/N$ を推定することである。誤差を考慮しない比例式 $N : m_i = c_i : r_i$ を解けば

$$\hat{N}_0 = m_i c_i / r_i \tag{2.1}$$

という式を得る。これは $r_i = 0$ のとき (全部, 新知見である好ましい捕獲のとき) 定義されず, 推定量として選ぶと分散が定義されない。そこで本研究では推定量として

$$\hat{N} = m_i(c_i + 1)/(r_i + 1) \tag{2.2}$$

を用いる。これを Petersen の修正式といい (Schaefer 1951), 分散は

$$V[\hat{N}] = \frac{m_i^2(c_i + 1)(c_i - r_i)}{(r_i + 1)^2(r_i + 2)} \tag{2.3}$$

と導かれている (Schaefer 1951, Jones 1964)。

ここでは相当多数回の捕獲 (知見の収集) を行うので, それらを利用した知見数の推定量

$$\hat{\hat{N}} = \Sigma r_i^* \hat{N}_i \qquad (r_i^* = (r_i + 1)/\Sigma(r_j + 1)) \tag{2.4}$$

を利用する。この方法を Schnabel 法 (Schnabel, 1938) といい, その際の推定量の 1 つである。ただし 2 回目の捕獲から I 回目の捕獲まで, 全てのデータを利用すると初期の偏りの影響を受け易く, 安定しない。 このため本論文ではバーンインして I 回目の捕獲から数回遡ったデータ (遡る回数をラグ回数と呼ぶ) を利用する。捕獲間の再捕獲数は互いに独立であり, その重み付き平均が $\hat{\hat{N}}$ であるから, その分散は

$$V[\hat{\hat{N}}_i] = \Sigma \left(\frac{m_i^2(c_i + 1)(c_i - r_i)}{(r_i + 2)} \right) / \left(\Sigma(r_i + 1)\right)^2 \tag{2.5}$$

と求めることができた。捕獲率 Cr は研究者にとって高いほうが都合がよい指標であるから, 本論文では控えめに, 知見総数の 95%信頼区間の上側限界を利用して, 以下のように推定することを提案する。

$$\hat{Cr} = \min(1,\ m_I/(\hat{\hat{N}} + 1.96 \times \sqrt{V[\hat{\hat{N}}_i]})) \tag{2.6}$$

2.3　適用例

2.3.1　データ

第 2 章では応用例として，株式会社日経 BP によるブランドジャパン 2011(日経 BP コンサルティング, 2011) のデータを用い, 毎年上位にランクインしているブランド「ブランド A (映画会社) 」と「ブランド B (教育機関) 」に関する自由記述意見から, 各ブランドに関するイメージの観点 (知見) を収集した。「ブランド A」は 2 年分合計 421 名分,「ブランド B」は 5 年分合計 518 名分の自由記述を分析の対象とした。

2.3.2　結果

「ブランド A」と「ブランド B」について，それぞれ 83 個, 84 個の知見を収集した。またブランド A の場合, 8 割を超えるのは 17 回目の捕獲であった。これは 85 枚の自由記述を読んだ時点であり, 精読に当てた全時間の 20 % であった。ブランド B の場合は 8 割を超えるのは 10 回目の捕獲であった。50 枚の自由記述を読んだ時点であり, 精読に当てた全時間の 10 % であった。捕獲率 Cr を参照すると, 精読に要する時間を大幅に節約することができることが示された。

第3章　自由記述における単語の種類数の捕獲率

3.1　はじめに

第 3 章では，収集対象が等確率で得られるという Schnabel 法の仮定に依存しない DeLury 法を用いて，前章よりも大きな標本サイズを想定した場合における捕獲率の推定方法について論じる。また，自由記述データの処理には自動コーディング技術を応用することとした。そのため，本章では自由記述から得られた“知見”ではなく，自動コーディング処理の結果得られた“単語 (名詞や形容詞)”を捕獲対象とした。

3.2　方法

自由記述データの収集の程度を示すために，資源量推定の一手法である DeLury 法 (Delury, 1947) を応用し，自由記述における名詞と形容詞の種類数を資源量として推定した。データの加工にはテキストマイニングツールを用い，単語の種類数の母数の推定値に対する，現在までに得られている単語の種類数の多さを以て自由記述におけるデータ収集の飽和の程度と考える。

3.2.1　DeLury 法

表記法は前章の Schnabel 法と同様であり，まず DeLury 法では以下の 2 つの仮定が導入される。

$$n_i = p_i N_{i-1},\ \ p_i = q x_i \tag{3.1}$$

これらの仮定の下，$y_i = n_i/x_i$ とすると，

$$y_i = qN - qm_{i-1} \tag{3.2}$$

のように qN を切片，$-q$ を回帰係数，y_i を基準変数，m_{i-1} を予測変数とした単回帰モデルとなる。N の推定値についてはデータの収集の順序による影響を防ぐため，ランダムな順序で並べ替えたデータを用いて一定回数 N を推定し，その結果から単語の種類の総数の推定値と標準誤差の平均 $\bar{N}$ と s.$\bar{\text{e}}$. を求める。捕獲率は前章と同様に，以下のように定義することを提案する。

$$\hat{Cr} = \min(1,\ m_I/(\bar{N} + 1.96 \times \text{s.}\bar{\text{e}}.)) \tag{3.3}$$

3.3 適用例

3.3.1 データ

適用例として，株式会社日経 BP によるブランドジャパン 2011(日経 BP コンサルティング, 2011) から，自動車会社 “ブランド A”(3686 名分)，レジャー関連企業 “ブランド B”(3370 名分) の印象について尋ねた自由記述における名詞と形容詞を分析対象とした。

3.4 結果

両ブランドの最終捕獲回における標識数 m_I(得られていたデータ内の単語の全種類数) と，順序をランダムに並べ替えた 1000 回分の計算の結果，ブランド A の名詞の捕獲率が 0.880，形容詞が 0.924，ブランド B の名詞の捕獲率が 0.845，形容詞が 0.940 だった。両ブランドの両品詞共に十分な単語の収集がなされていたと考えられる。

第4章　自由記述における汲みつくしの指標としての遭遇率の提案—パレート分布を用いた分析—

4.1　はじめに

例えば捕獲率が 80%の場合でも次の捕獲で知見や単語が得られる確率がまだ十分に高ければ，収集を続けた場合に十分な収穫が得られる。その一方で，次の捕獲で知見や単語が得られる確率が非常に低い場合には，これ以上の収集を続けても新たな単語が得られる見込みは少ない。このため，次の捕獲で知見や単語が得られる確率を指標化し，捕獲率と共に用いることができれば，自由記述の収集の程度について更なる検討を行うことができる。第 4 章ではそのような指標として遭遇率を提案した。

4.2　方法

4.2.1　頻度を説明するモデル

単語の出現頻度を説明するモデルとしては，ジップ分布 (Zipf, 1932) が挙げられる。ジップ分布とパレート分布は出現頻度の総数を N とするとき，両者はべき則をもつ共通の確率密度関数のもとで同等な関係が成立する (Adamic, & Huberman, 2002)。パレート分布には様々な形状のものがあるため，今回は第一種のパレート分布に加え，第二種のパレート分布および一般パレート分布を用いて単語の出現頻度の説明を試みた。

4.2.2 各種パレート分布の PDF

単語の出現頻度の分布を説明するための 3 種類のパレート分布を示す。まず第一種のパレート分布の PDF は，母数 $\alpha > 0, \beta > 0$ について，

$$f(x|\alpha, \beta) = \frac{\beta\alpha^{\beta}}{x^{\beta+1}}, \quad x \geq \alpha \tag{4.1}$$

と表わされる。続いて第二種のパレート分布の PDF は，母数 $a > 0, b > 0$ について，

$$f(x|a, b) = \frac{ab^{a}}{(x+b)^{a+1}}, \quad x > 0 \tag{4.2}$$

と表現される。そして一般パレート分布の PDF は，母数 c について $c \neq 0$ の場合，

$$f(x|k, c) = \frac{1}{k}\big(1 - \frac{cx}{k}\big)^{1/c-1} \tag{4.3}$$

$c = 0$ の場合，

$$f(x|k, c) = \frac{1}{k}\exp\big(-\frac{x}{k}\big) \tag{4.4}$$

となる。ここで $c \leq 0$ のとき $x > 0$，$c > 0$ のとき $0 < x < k/c$ である。

4.2.3 遭遇率

実在のデータから得られた n 種類の単語の出現頻度から各パレート分布を当てはめた結果，最もよく適合したモデルを単語の出現頻度を表す分布として採用し，その累積分布関数を求める。n 番目までの頻度で出現する単語が得られる確率を $F(n)$ と表記すれば，更なる探索一試行における未知の単語との遭遇率は，以下のように推定される。

$$EP = 1 - F(n) \tag{4.5}$$

4.2.4 データと加工

本研究では適用例として，第 3 章と同じ株式会社日経 BP コンサルティング (2011) からブランド C と D を用いた。

4.2.5　推定方法

推定方法には，マルコフ連鎖モンテカルロ (markov chain monte carlo, MCMC) 法によるベイズ推定を用いた。その際に事前分布としては，母数の定義されない領域を除いて無情報となるように，以下のような一様分布を設定した。

$$\beta, a, b, k \sim U(0, 1.0e+6), c \sim U(-1.0e+6, 0) \tag{4.6}$$

ただし，第一種のパレート分布の母数 α については，定義域が $x \geq \alpha > 0$ であることから，この値は単語の頻出順位の最小値 1 であるとして固定した。

4.3　結果

すべての母数について Geweke(1992) の指標が-1.96 から 1.96 の内に収まっていたため，マルコフ連鎖の収束が示唆された。また各分布の母数の推定値から計算した適合度，DIC を比較した結果，値の小ささから名詞については第二種と一般パレート分布が，形容詞については第一種のパレート分布がより単語の頻出具合を説明できていると判断した。

今回のデータから，ブランド C の名詞は 1694 種，形容詞は 152 種，ブランド D の名詞が 1390 種，形容詞は 149 種確認された。また未知の単語との遭遇率はブランド C の名詞が 0.044，形容詞が 0.062，ブランド D の名詞が 0.051，形容詞が 0.068 となり，最大で 7 パーセント程であることが分かった。十分に小さな値である事から，データの収集は十分であることが分かった。

第5章　Web文書収集の飽和度の計算とWeb探索

5.1　はじめに

第5章では，第3章と第4章で提案された捕獲率と遭遇率を用いて，Web上のデータを対象とした研究を行う。まず分析対象として，これまでに扱ってきた自由記述アンケートの回答ではなく，収集されたWeb文書をデータとして用い，手元のデータ内の捕獲率，遭遇率を計算した例を示す。続いて手元の文書内のリンク情報を用いて得られたリンク先のWeb文書を加えることで，さらなる文書の収集を行う方法を示す。

5.2　方法

本章では，第3章で提案されたDeLury法を用いて捕獲率の計算を行った。ただし前章までに比較して非常に多くのデータを扱うため，3章のファントム変数を用いるプログラムではNのs.e.が算出されない場合がある。このため次の小節でDeLury法における分散とs.e.の式を導入し，分析の結果プログラムからNのs.e.が算出されなかった場合は直接的にs.e.を計算することとする。

5.2.1　DeLury法における分散とs.e.

第3章の表記法を用いるとDeLury法の分散は，Krebs(1999)やSokal & Rohlf(1995), Zar(1996)などから以下となる。

$$\mathrm{V}[N] = \frac{s^2}{q^2}\left[\frac{1}{I} + \frac{(N-\bar{m}_{i-1})^2}{\sum(m_{i-1}-\bar{m}_{i-1})^2}\right] \tag{5.1}$$

ここで，

$$s^2 = \sum \frac{(y_i - q(N - m_{i-1}))^2}{(I-2)} \tag{5.2}$$

である。よって N の標準誤差 (s.e.) は以下となる。

$$\text{s.e.} = \sqrt{\text{V}[N]} \tag{5.3}$$

先述のように，本節では 3 章と同じ DeLury 法のプログラムで N の推定値を求めるものの，標本数があまりに多くなるため，計算機の問題上 N の分散が出力されない場合がある。その場合には (5.3) 式を用いて直接 N の s.e. を算出し，3 章と同じく以下のように単語の総種類数の 95%信頼区間の上側限界を利用して捕獲率 Cr を定義する。

$$\hat{Cr} = \min(1,\ m_I/(\bar{N} + 1.96 \times \text{s.e.})) \tag{5.4}$$

5.3 研究 1：分析の手順

研究 1 では，まずはじめに種ファイルとして用いた研究室 E(心理学，統計学を専攻する大学の研究室) について Web 文書の収集を行った。その結果，この研究室のホームページである 60 個の Web 文書が収集され，これらを種ファイルと考えた。次に，得られたテキストデータについて形態素解析を実行し，今回はテキスト内から名詞の単語のみを抽出した。最後に，クリーニングを終えたデータを用いて捕獲率と遭遇率の計算を行った。捕獲率はデータの順序をランダムに入れ替えて 100 回計算した結果の平均値を指標として用いた。

5.4 研究 1：結果と考察

研究室 E の 60 ファイル分（27100 語）の捕獲率を計算した結果，種ファイルに関しては捕獲率が 0.784 であることが分かった。50%を上回る捕獲率が計算されたことから，この研究室についての Web 文書は単語の種類数推測の観点から名詞が十分に収集されていると判断される。

また，第 4 章同様に各パレート分布のデータへの当てはまりを比較したところ，名詞については，DIC の値の低さから研究 1 の Web 文書についても第一種

のパレート分布より第二種・一般パレート分布を用いた方が，よりモデルがデータに適合していることが分かった。このデータにおける遭遇率は 0.042 と十分に低いことから，これらの 60 個の Web 文書だけでも，研究室 E についての情報収集が十分であることが，名詞の捕獲率と遭遇率の観点から示唆された。

5.5 研究 2：分析の手順

研究 2 では，まず種ファイル内にあったリンク情報からリンク先の Web 文書を種ファイルとの合計が 1000 個になるように 940 個収集し，この中から日本語を主な言語とする 451 個の Web 文書を分析対象とした。次に得られた葉ファイルについて形態素解析を実行し，テキスト内から名詞の単語のみを抽出した。最後に，クリーニングを終えたデータを用いて捕獲率と遭遇率の計算を行った。捕獲率の計算に当たっては，得られた葉ファイルから内容的に種ファイルと関連のある Web 文書を選択し，それらを種ファイルに加えてデータの順序を入れ替えて，一定回数計算した。

5.6 研究 2：結果と考察

分析の結果，DIC の値の低さから，研究 2 の Web 文書についても第一種のパレート分布より第二種・一般パレート分布を用いた方が，よりモデルがデータに適合していることが分かった。遭遇率については，選択したファイルは，種ファイルのみを用いた場合よりは値が若干大きくなっていた。選択したファイルを用いた場合，種ファイルに加えて捕獲率を 10%程度上昇させることができた。

第6章　授業評価における知見収集の飽和度

6.1　はじめに

第 6 章では，第 2 章と第 3 章で提案された捕獲率を用いて，授業評価のデータを対象とした研究を行う。得られた 83 名分の授業評価データから，まず第 2 章と同様に KJ 法を用いて知見を収集する。続いて Schnabel 法を用いた知見の総種類数の結果を示し，捕獲率の計算を試みる。

6.2　方法

本章では，第 1 章で提案された Schnabel 法を用いて捕獲率の計算を行う。また比較のため，第 3 章で提案された DeLury 法を用いて捕獲率の計算を行った結果を示し，2 つの方法による知見の総種類数の推定結果について考察する。

6.2.1　用いたデータ

本章ではデータとして，私立大学における心理統計学の授業 “授業 F”に対する授業評価の回答を用いた。評価は授業を 1 年間受講した 83 名の生徒に対して行われ，第 1 項目が 1 時間目，第 2 項目が 2 時間目に関する教材の質，授業の進度，宿題の内容，難易度，その他実習に関して，なるべく教材の改善に結びつく視点での回答を求めた自由記述形式の質問項目となっていた。

6.2.2　データの加工方法

第 2 章と同様に自由記述からの知見の抽出を行った。アンケートの文章中，授業評価として取り上げるべきか否かの判断と新知見として取り上げるべきか否

かの判断は，“授業 F”を担当した教員と TA の合計 3 名の合議によって行われた。

6.2.3　得られた知見の分析方法

得られた知見について，収集した 83 名分の自由記述で収集が十分であったかどうかを捕獲率の計算によって確認する。ただし本章の場合，第 2 章よりも自由記述の枚数が少ないため，第 2 章のように相当多数回の捕獲は行っていないため，(2.4) 式においてラグの回数を 0 回とした場合に $\hat{\hat{N}} = \hat{N}$ となることを利用して，以下を捕獲率の指標として分析を行う。

$$\hat{Cr} = \min(1,\ m_I/(\hat{N} + 1.96 \times \sqrt{V[\hat{N}_i]})) \tag{6.1}$$

次にこの指標を用い，捕獲率の値が過半数の 0.5 以上であった場合に収集は十分であったとみなし，得られた知見を利用してさらなる授業評価の分析を行う。分析対象は必修科目を受講した回答者のうち，次年度に心理学の研究室に配属された回答者と，その回答者から得られた知見とした。

6.3　結果

6.3.1　“授業 F”の知見

“授業 F”については，53 個の知見が収集された。また “授業 F”が必修である生徒 80 名からは知見 1 から 50 まで，外部の聴講生である生徒 3 名からは知見 1 から 53 までの内容の知見が得られた。

6.3.2　捕獲率の計算結果

83 枚の回答を 3 枚毎にまとめて 28 回の捕獲があったものとし，Schnabel 法を用いて捕獲率を計算した。その結果 28 回の捕獲による捕獲率は 0.930 と，これらの 53 個の知見で総知見数の 90%以上の知見を網羅していることが分かった。また DeLury 法を用いて捕獲率の計算を行ったところ，推定値が数値的に大きな差が見られなかった。また第 3 章と第 5 章で紹介した各 s.e. の比較の結果にも大差はなく，どちらの結果を用いても捕獲率の計算には問題ないことが示唆された。

第7章　就職活動の成功体験に関する知見の収集と分析

7.1　はじめに

第 7 章ではインタビュー法を用いてデータを取り，得られた知見について捕獲率を計算する。また得られた知見を変数として，データマイニング手法を適用して更なる分析を行った分析には最近 2 年間に就職活動を終えた学生の体験談を尋ねたインタビューデータを用い，得られた知見を変数として決定木と連関規則の分析手法を適用し，どのような意見を呈したインタビュー対象者が自身の就職活動により満足感を得ていたか，また呈された知見間にどのような関連性がみられるかを確認した。ただし本稿では，連関規則と決定木の結果は割愛し，データの収集が十分であったことのみを提示する。

7.2　方法

7.2.1　データの収集

データの収集はインタビュー法を用いて行った。調査対象者は 2013 年または 2012 年に就職活動を行った学生 23 名だった。

7.2.2　データの処理

すべての会話を録音し，文章化を行った。その際にデータのクリーニングを行った。次に編集版の記事を作成した。そしてこの編集版の記事内の知見について，収集されたデータの量で知見の収集が十分であるかどうか，捕獲率を計算することによって示す。本章もインタビュー対象者の数が 23 名と少数であっ

たため，第 6 章と同様にラグの回数は 0 として (6.1) 式の捕獲率を指標として用いた。

7.2.3　記事からの知見の抽出

編集版の記事から，第 2 章と同様に知見を抽出した。ただし第 2 章と異なり，90 分のインタビュー記事は非常に文章量が多かったため，まず記事から知見の候補となる箇条書きを抜き出し，その中から重要な知見を採用した。

7.3　結果

7.3.1　抽出された知見

23 名のインタビューを行った結果，就職活動に有益であると判断された知見が合計 416 種，大分類して 31 種類見つかった。今後，これらの知見を C1〜C416 と表記する。

7.3.2　Schnabel 法による捕獲率の計算結果

知見の大分類の種類についての捕獲率を計算したところ，捕獲率がほぼ 1.000 となり，これらの 31 個の知見で総知見数のほぼすべてを網羅していることが分かった。また，23 名分から得られた知見を 1 人分ずつ増やして捕獲率を計算していった結果，捕獲率は 0.897 となり，これらの 416 個の知見で総知見数の大部分を網羅していることが分かった。

第8章　総合考察

本稿では，これまでテキストデータ収集の飽和の程度を評価する指標の開発と応用を試みてきた。様々な標本サイズ，データの内容に対応した捕獲率，遭遇率を提案し，自由記述，Web 文書，授業評価アンケート，インタビューデータといった応用事例を通じて，これらの指標の有効性を確認してきた。

本稿で提案した捕獲率や遭遇率は，これまでに存在しなかった“課題に関する論点は十分収集された”という基準を客観的に示すことができる有益な指標である。この指標を確認することで，査読者や一般読者は，理論構築に際して研究者が十分に知見やキーワードとなる単語を収集していたことをより明確に知ることができる。また研究者も十分データを収集したことを，捕獲率の観点から客観的に示すことができる。この意味において本研究の意義は深いと考える。

自然言語処理の技術が年々向上し，ユーザーフレンドリーで優良なテキストマイニングツールが次々と発表される中，テキストを分析対象とした研究は益々多くみられるようになってくるだろう。そのような中で，データ収集をどこかで打ち切り，データ収集の程度を示すことが必要となったときに，本手法が有効に機能すれば幸甚である。捕獲率の計算手法は実在の海での様々な経験を活かして発達してきた資源量推定法に依るところが大きい。情報の海でのこの手法の多くの応用を経て，テキストデータの事情に沿った新たな手法の提案を目指すことが，今後の課題として期待される。

引用文献

Adamic, L. A. & Huberman, B. A. (2002). Zip's law and the Internet. Glottometrics, 3, 143-450.

Delury, D. B.(1947). "On the estimation of biological populations," *Biometrics* Vol.3, 145-167.

Jones, R. (1964). A review of methods of estimating population size from marking experiments. *Rapp. p.-v. Reun. Cons. perm. int. Explor. Mer.*, 155, 202-209.

Krebs, C. J. (1999). Ecological methodology. Second edition. Addison-Welsey, Menlo Park, California, USA.

日経 BP コンサルティング (2011). ブランドジャパン 2011 総合報告書 [解説書]. 日経 BP コンサルティング.

Schaefer, M.B. (1951). Estimation of size of animal populations by marking experiments. U.S. Dept. Int. Fish Wildl. Serv. Bull. 69, 187-203.

Schnabel, Z.E. (1938). The estimation of total fish population of a lake. *American Mathematical Monthly, 45*, 348-352.

Sokal, R. R., and F.J. Rohlf. 1995. Biometry. Third Edition. New York: W. H. Freeman and Company.

Zar, J. H. (1996). Biostatistical analysis. Third edition. Prentice-Hall International, Englewood Cliffs, New Jersey, USA.

Zipf, G.K. (1932). Selected studies of the principle of relative frequency in Language. Cambridge, MA.: Harvard University Press.