混合 Polya 分布による多項文書頻度のモデル化

貞光 九月*　　山本 幹雄*　　内山 将夫**

*筑波大学 {sadamitsu@milab.,myama@}is.tsukuba.ac.jp
**通信総合研究所 mutiyama@crl.or.jp

1　はじめに

ある単語を含む文書数を数えた文書頻度 (df: document frequency) はコーパスから得ることができる基本的な統計量の一つである。文書頻度は情報検索等の応用において重要であり、その確率的なモデルが果たす役割は大きい [Robertson&Walker94]。単語の出現頻度は話題やスタイルによって大きく変化するため、文書頻度は単純な二項分布で近似したものと大きくずれることが知られている。そこで話題やスタイルの分布を階層的な確率モデルで取り込んだモデルが提案されている (例えば、負の二項分布)[Church&Gale95, Jansche03]。

本稿では、文書頻度を複数の単語を対象とするように拡張した多項文書頻度を定義し、これに対する混合 Polya 分布に基づくモデルを検討する。Polya 分布は近似的に負の二項分布の多変量版と見なすことができ、単項文書頻度モデルの自然な拡張となっている。

以下では、まず多項文書頻度を定義し、その応用例として文書頻度を用いた相互情報量を検討する。次に多項文書頻度の確率モデルを定義・検討した後に、実際のコーパスを用いて多項文書頻度の実測値と比較する。

2　多項文書頻度

2.1　多項文書頻度の定義

ある単語 w に対する (単項) 文書頻度 $df(w)$ は、文書コレクション中で w を少なくとも 1 つは含む文書の数である。また、文書中の出現頻度まで考慮に入れた文書頻度 $df(k;w)$ は、w をちょうど k 個含む文書の数である。このように従来の文書頻度は、単一の単語に対する概念であったが、複数の単語に対する概念へ自然に拡張できる。多項文書頻度 $df(w_1^n)$ は、n 個の単語 $w_1^n = w_1, ..., w_n$ をすべて 1 個以上含む文書の数とする[*1]。特に単語数 n が決まっている場合は、「n 項文書頻度」と呼ぶことにする。

さらに、文書中の出現頻度を考慮した多項文書頻度 $df(k_1^n; w_1^n)$ は、単語 $w_1^n = w_1, ..., w_n$ がそれぞれ $k_1^n = k_1, ..., k_n$ 回出現した文書の数を表すものとする。以下が成り立つ。

$$df(w_1^n) = \sum_{k_i>0} df(k_1^n; w_1^n)$$

また、各文書における対象としている単語の総数を決めた場合の文書頻度 $df(k; w_1^n)$ を次のように定義する。

$$df(k;w_1^n) = \sum_{k_i>0, (\sum_i k_i)=k} df(k_1^n; w_1^n)$$

n 未満の k の値に対しては、上式は未定義である。

2.2　文書頻度に基づく相互情報量

文書頻度を相互情報量に適用した例を示す。相互情報量は二つの単語間の出現頻度に基づき共起の程度を見る指標であるが [Church&Hanks89]、ここでは出現頻度を文書頻度に置き換えた次のような量 $MI_{df}(w_1, w_2)$ を考える [山本 (英) 他 04]。ここで、D は文書の総数である。

$$MI_{df}(w_1, w_2) = \log \frac{D df(w_1^2)}{df(w_1) df(w_2)}$$

表 1 に文書頻度を用いた相互情報量の大きなペアの例を示す。表の左は「野球」との MI_{df}[bit] が大きくなる上位 10 単語である。右は「しか」の場合の上位 10 単語である。 ただし、「しか」の場合は機能語との共起を見たいので、出現回数が 2000 回以上の単語に限定してある。「じゃ」「本当に」「なぜ」等は「しか」と連接していない場合が多いので、従来の $n$gram 統計では得にくい単語である。

3　多項文書頻度のモデル化

3.1　単項文書頻度に対するモデル

Church らは、単語の出現確率の変動を吸収するために、ポアッソン分布のパラメータに対する確率分布を仮定し、階層的なモデルによって文書頻度をモデル化した [Church&Gale95]。例えば、パラメータの分布としてガンマ分布を仮定した以下のような負の二項分布はかなり正確に文書頻度を近似できる。

$$P_{NgBn}(k; p, n) = \binom{-n}{k} p^n (1-p)^k$$

[*1] 記法の単純化のために $df(\cdot)$ の引数の数や型で (単項) 文書頻度か多項文書頻度か、あるいは続いて定義される出現頻度を考慮した多項文書頻度であるかどうかなどを区別する。

表 1: MI_{df} 値の大きな単語ペアの例 (毎日新聞 1999 年版)

w_2 (w_1=野球)	MI_{df}	w_2 (w_1=しか)	MI_{df}
高野連	5.33	くらい	2.01
球児	5.29	わずか	2.00
軟式	5.25	こう	1.98
硬式	5.18	じゃ	1.97
黒獅子旗	5.18	分から	1.97
神宮球場	5.13	本当に	1.96
ペナントレース	4.99	ぐらい	1.93
コミッショナー	4.96	言え	1.90
球界	4.89	なぜ	1.89
始球	4.88	のに	1.89

これは対象とする単語が文書中に k 回出現する確率で、文書の総数 D を掛けることで $df(k)$ を予測できる。p と n は負の二項分布のパラメータである。

同様のことは二項分布を利用しても可能である。二項分布の場合はパラメータが単語の確率であるためベータ分布を仮定すると、その合成分布はベータ二項分布 $P_{BeBn}(k;\alpha,\beta,n)$ となる [Jansche03]。

$$P_{BeBn}(k;\alpha,\beta,n) = \binom{-\alpha}{k}\binom{-\beta}{n-k}/\binom{-\alpha-\beta}{n}$$

ここで、n は文書の平均長である。ベータ二項分布の場合、文書の平均長を固定して文書確率を近似するので、この点で負の二項分布よりも不利である。しかし、$p=\beta/(n+\beta), k=\alpha$ を一定に保ったまま $n\to\infty$ とすれば、ベータ二項分布は負の二項分布に分布収束する。n は新聞記事で数百 (単語/記事) であるから、2 つの分布はほぼ等価 (相互にパラメータを変換可能) として扱ってよい。

3.2 多項文書頻度に対するモデル

3.2.1 混合 Polya 分布

前節で述べた単項文書頻度の確率分布は単変量分布であったのに対して、多項文書頻度のモデルは多変量分布である必要がある。ベータ二項分布の元となった分布 (二項分布とベータ分布) を多変量化すると (それぞれ、多項分布とディレクレ分布)、以下のような Polya 分布 (別名、ディレクレ多項分布) が得られる。ここで、$\boldsymbol{\alpha}$ は n 個のパラメータ $\alpha_1,...,\alpha_n$、$k=\sum_i^n k_i$、$\alpha=\sum_i^n \alpha_i$ である。

$$P_P(k_1^n;\boldsymbol{\alpha},k) = \binom{k}{k_1!...k_n!}\frac{\Gamma(\alpha)}{\Gamma(\alpha+k)}\prod_{v=1}^{n}\frac{\Gamma(k_v+\alpha_v)}{\Gamma(\alpha_v)}$$

残念ながら Polya 分布は単語同士の共起の強さをモデル化できない。そこで、複数の話題を導入し、話題ごとに Polya 分布を設定し、それらの混合分布、すなわち混合 Polya 分布としてモデル化する。m を混合数、$\boldsymbol{\lambda}=\lambda_1,...,\lambda_m$ を混合重み ($\sum_i^m \lambda_i=1$)、$\boldsymbol{\alpha}_1^m=\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_m$ をそれぞれ (i 番目) の話題ごとの Polya 分布のパラメータ ($\boldsymbol{\alpha}_i=\alpha_{i1},...,\alpha_{in}$) とすると、次のように混合 Polya 分布 (PM: Polya Mixtures) を定義できる。

$$P_{PM}(k_1^n;\boldsymbol{\alpha}_1^m,\boldsymbol{\lambda},k)=\sum_{i=1}^{m}\lambda_i P_P(k_1^n;\boldsymbol{\alpha}_i,k);$$

混合 Polya 分布は多項分布のパラメータに対して混合ディレクレ分布を仮定した場合の合成分布である。詳しくは文献 [山本 (幹) 他 03] を参照のこと。

3.2.2 頻度付多項文書頻度 (a):$df(k_1^n;w_1^n)$

本節では、混合 Polya 分布を用いて、$\hat{df}(k_1^n;w_1^n)=D\times P(k_1^n;w_1^n)$ [*2]をモデル化する。

実際のモデル化では対象とする単語すべて (例えば 2 万単語) を同時に混合 Polya 分布 (単語数が 2 万単語とすると 2 万変量) でモデル化し、n 項文書頻度を計算する場合は対象の n 単語に周辺化すればよい。例えば、もともと N 単語 (w_1^N) を対象とする N 変量混合 Polya 分布を 2 単語 ($\acute{w}_1^2$) に周辺化する場合を考える。2 単語はそれぞれ、$\acute{w}_1=w_a,\acute{w}_2=w_b$ であるとする。$(k_1^N)_{-ab}$ は k_1^N から k_a と k_b を除いた確率ベクタとすると、周辺化された確率 $P(\acute{k}_1^2;\acute{w}_1^2)$ は次のように計算できる。

$$\begin{aligned}P(\acute{k}_1^2;\acute{w}_1^2) &= \int P_{PM}(k_1^N;\boldsymbol{\alpha}_1^m,\boldsymbol{\lambda},k)d(k_1^N)_{-ab}\\ &= P_{PM}(\acute{k}_1^3;\acute{\boldsymbol{\alpha}}_1^m,\boldsymbol{\lambda},k)\\ &= \sum_i^m P_P(\acute{k}_1^3;\acute{\boldsymbol{\alpha}}_i,\boldsymbol{\lambda},k) \qquad (1)\end{aligned}$$

ここで、$\acute{\boldsymbol{\alpha}}_i=\{\acute{\alpha}_{i1},\acute{\alpha}_{i2},\acute{\alpha}_{i3}\}$、$\acute{k}_3$ は対象とする 2 単語以外の単語数を表し $\acute{k}_3=k-\acute{k}_1-\acute{k}_2$ である。パラメータは、$\acute{\alpha}_{i1}=\alpha_{ia},\acute{\alpha}_{i2}=\alpha_{ib},\acute{\alpha}_{i3}=\sum_{j\neq ab}\alpha_{ij}$ となる ($\boldsymbol{\lambda}$ は同じ)。一般の n 項文書頻度の場合も、同様に単純な操作で周辺分布が求まる。文書長 k は記事の平均長 L を設定する。

3.2.3 頻度付多項文書頻度 (b): $df(k;w_1^n)$

本節では $\hat{df}(k;w_1^n)=D\times P(k;w_1^n)$ をモデル化する。混合 Polya 分布の確率変数の成分和 ($z=k_1+k_2$ など) の分布は、その対象とする複数単語を 1 つの単語とみなし、1 変量に周辺化した場合と等しい。例えば、2 つの単語 $\acute{w}_1=w_a$ と $\acute{w}_1=w_b$ の出現回数の和 z の分布は以下のようになる。

$$P_{sum}(z;\acute{w}_1^2)=P_{PM}(k_1=z,k_2=k-z;\acute{\boldsymbol{\alpha}}_1^m,\boldsymbol{\lambda},k)$$

[*2] 以下、$df(\cdot)$ に対応する確率を $P(\cdot)$ で表現することとする。$P(\cdot)$ に文書総数 D を掛けると $\hat{df}(\cdot)$ となる。

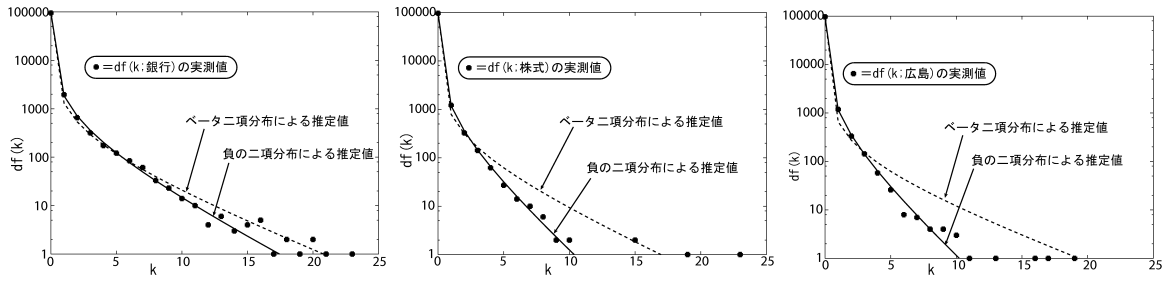

図 1: 単項文書頻度 $df(k;w)$ の実測値と推定値 (負の二項分布とベータ二項分布)

ここで、$\boldsymbol{\acute{\alpha}}_i = \{\acute{\alpha}_{i1}, \acute{\alpha}_{i2}\}$ であり、$\acute{\alpha}_{i1} = \alpha_{ia} + \alpha_{ib}$, $\acute{\alpha}_{i2} = \sum_{j \neq ab} \alpha_{ij}$ となる。この成分和の分布 $P_{sum}(z; \acute{w}_1^2)$ は、どちらかの単語の出現回数が 0 の場合 ($\acute{w}_1$ が z 回、$\acute{w}_2$ が 0 回等) を含むので、この確率 ((1) 式を使う) を引けば求めるべき確率となる。

$$P(k; \acute{w}_1^2) = P_{sum}(k; \acute{w}_1^2) - P(k_1 = k, k_2 = 0; \acute{w}_1^2) - P(k_1 = 0, k_2 = k; \acute{w}_1^2) \quad (2)$$

3.2.4 多項文書頻度: $df(w_1^n)$

本節では、$\hat{df}(w_1^n) = D \times P(w_1^n)$ をモデル化する。$P(w_1^2)$ は、少なくとも一方の単語が出現しない確率を (1) 式, (2) 式を組み合わせて求め、その確率を 1 から引くことによって求まる。以下のようになる。

$$P(w_1^2) = 1.0 - \{P(k=0; w_1) + P(k=0; w_2) - P_{sum}(k=0; w_1^2)\} \quad (3)$$

$P(k; w_1)$ および $P(k; w_2)$ は、(1) 式を 1 単語に周辺化した確率である。

4 多項文書頻度の実測とモデルの評価

4.1 実験データ

毎日新聞 1999 年版 (総単語数約 3000 万、記事数約 10 万)[毎日 00] を用いて多項文書頻度の実測とモデル推定、およびモデルの評価を行った。対象単語は主に「銀行」「株式」「広島」とし、二項ペアは { 銀行, 株式 } と { 銀行, 広島 } の 2 つのペアに注目して実測・検討を行った。毎日新聞 1999 年版におけるそれぞれの単語の出現頻度および文書頻度は表 2 に示すとおりである。「株式」と「広島」は、出現頻度および (単項) 文書頻度がほぼ等しい単語である。

表 2: 各単語の出現頻度と文書頻度: 毎日新聞 99 年版 (98211 記事, 約 3000 万単語)

単語	出現頻度 (tf)	文書頻度 (df)
銀行	7623	3467
株式	2985	1807
広島	2922	1781
銀行, 株式	-	402
銀行, 広島	-	36

モデルのパラメータ推定にも毎日新聞 1999 年版全体を用いた。単項文書頻度に関する負の二項分布のパラメータは文献 [Church&Gale95] の method2 を用い、ベータ二項分布のパラメータは [政瀧他 98] の方法を用いて求めた (両者ともモーメント法である)。混合 Polya 分布の混合数は 100、対象単語は高頻度 2 万単語を用いた。一つの混合分布で 2 万単語のモデルを作成したため高い精度は望めないが、前節で述べた周辺化によって 2 万単語中の任意の n 項の文書頻度の分布を容易に得ることができる。単語組ごとにモデルを求めようとすると、2 単語 (二項文書頻度) であっても語彙サイズの 2 乗オーダのモデルを推定しなければならない。なお、本報告で用いた 100 混合のモデルは文献 [山本 (幹) 他 03] で述べている 100 混合のモデルをそのまま利用している。

4.2 実測値と推定値の比較

図 1 はそれぞれの 3 単語の各単項文書頻度 $df(k; w)$ の実測値と、それを負の二項分布およびベータ二項分布でモデル化した値を示す。図 1 より、負の二項分布の方が精度が高いことが分かるが、3.1 節で述べたようにこれは文書長の分布を考慮しているか否かの差による。

図 2 は $df(j, k;$ 銀行, 株式$)$ と $df(j, k;$ 銀行, 広島$)$ の実測値と混合 Polya 分布を用いて計算した推定値 (3.2 節 (1) 式) である。「銀行」と「株式」は強い共起関係にあり、「銀行」と「広島」は弱い。混合 Polya 分布によるモデルは、頻度をやや小さめにモデル化しているが、共起関係はそれぞれのペアの特徴に応じて捉えていることが分かる。ちなみに、「株式」と「広島」は単独ではほぼ同じ分布をしているため (図 1)、それぞれの単語を独立だとみなすと、「銀行」との共起頻度はほぼ同じにモデル化されるはずである。この点から考えて、混合 Polya 分布は文書レベルにおける複数単語の共起傾向を捉える

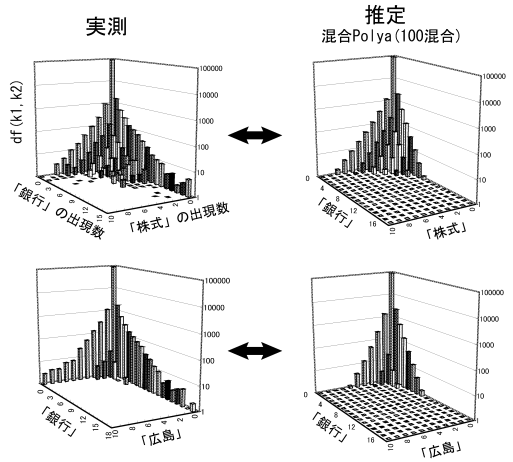

図 2: $df(k_1^2; w_1^2)$ の実測値と推定値

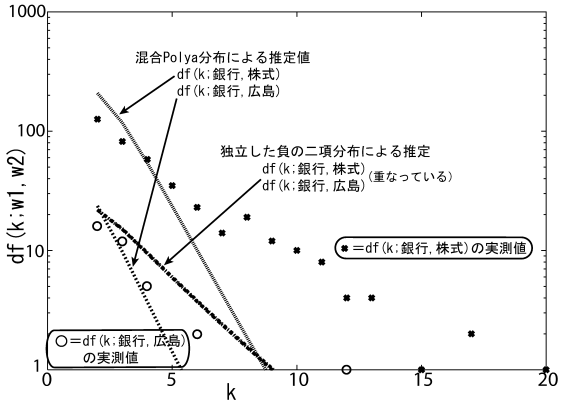

図 3: $df(k; w_1^2)$ の実測値と推定値

確率モデルの候補と成りえるものである。

図 3 は $df(k;$銀行,株式$)$ と $df(k;$銀行,広島$)$ の実測値、混合 Polya 分布による推定値 (3.2 節 (2) 式)、およびペアとなっているそれぞれの単語が独立に生起すると仮定した場合の文書頻度の推定値を示す。独立と仮定する場合は、単項文書頻度のモデル (図 1 の負の二項分布) を用いて求めた。$df(k; w_1^2)$ の場合でも 2 つのペアの共起性は大きく異なるが、独立を仮定した場合は、2 つのペア共にほぼ同じ分布となり、共起性のモデル化をまったくできないことが分かる。混合 Polya を用いた場合は、図 2 と同様、絶対値は小さめであるが、それぞれのペアの特徴を明確に区別できている。

図 4 はランダムに選んだ約 10000 の単語ペアごとに計算した二項文書頻度 $df(w_1^2)$ の実測値と混合 Polya 分布による推定値 ((3) 式) を散布図にしたものである。一つのモデルで 2 万単語をモデル化した場合、精度は残念

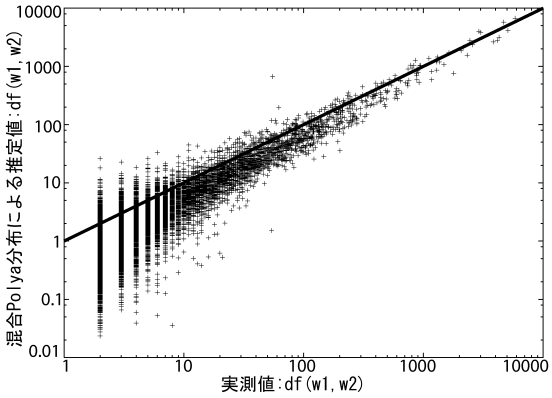

図 4: 二項文書頻度 $df(w_1^2)$ の実測値と推定値

ながらそれほど高くないことが分かる。

5 おわりに

多項文書頻度を定義し、混合 Polya 分布に基づく確率モデルを提案した。今回の実験では、2 万単語をたった一つの混合 Polya 分布でモデル化したにもかかわらず、ある程度の共起性をモデル化できていることを実験で確認した。今後は、多項文書頻度や混合 Polya 分布を利用して、共起性の強い単語集合の特徴付けや抽出を行う予定である。

参考文献

[Church&Hanks89] K.W.Church and P.Hanks. 1989. Word association norms, mutual information and lexicography. In ACL 27, pages 76-83.

[Church&Gale95] K.W.Church and W.Gale. 1995. Poisson mixtures. Natural Language Engineering 1 (2), pages 163-190.

[Jansche03] M.Jansche. 2003. Parametric models of linguistic count data. In ACL 41, pages 288-295.

[Robertson&Walker94] S.E.Robertson and Walker S. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proc. of SIGIR94, pages 232-241.

[政瀧他 98] 政瀧, 匂坂, 久木, 河原. 1998. 最大事後確率推定による N-gram 言語モデルのタスク適応. 信学会論文誌 D-II, 11(J81), pages 2519-2525.

[毎日 00] 毎日新聞社. 2000. CD-毎日新聞 1999 年版. 日外アソシエーツ.

[山本 (英) 他 04] 山本英子, 木田, 神崎, 井佐原. 2004. 共起情報に基づく呼応関係自動抽出法の検討. 情処学会研究報告, 2004-NL-159.

[山本 (幹) 他 03] 山本幹雄, 貞光, 三品. 2003. 混合ディレクレ分布を用いた文脈のモデル化と言語モデルへの応用, 情処学会研究報告, 2003-SLP-48, pages 29-34.