統計データ収集のためのフォーカストクローラ

和久井拓斗† 加藤 誠††

† 筑波大学 情報学群 知識情報・図書館学類 〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系 / JST さきがけ 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1811550@klis.tsukuba.ac.jp, ††mpkato@acm.org

あらまし 本研究では，ウェブ上に存在する統計データを効率よく収集するためのフォーカストクローリングアルゴリズムについて提案する．具体的には，探索/収集アプローチにおける探索フェーズにおいて，対象とするウェブサイト集合の中から統計データを有するウェブサイトを見分ける手法について扱う．本手法は，まず各ウェブサイトに対して一部のページ内容を確認し，各ウェブサイトの統計データの有無を判定する．同時に一部のページしか用いていないことによる判定の誤り確率を推定し，その大きさによって次の探索対象を決定する．実験では，統計データの有無に関してあらかじめラベル付けした実際のウェブサイトに対して，本手法に基づくクロールシステムを実装し評価を行った．

キーワード Web クローリング，フォーカストクローリング，統計データ，確率モデル，Web ページ分類

1 はじめに

統計データは政策や社会調査，企業などでの新規事業の立案時や，研究機関における基礎データとして有用であることは広く認識されている．しかし，一方で，公に人々に対して統計データへの高いアクセシビリティを提供することの重要性についてはあまり注目されていない．何より問題なのは，政府や地方自治体，企業などの様々な機関によって管理・運営されている統計データにのみフォーカスした，包括的な検索を可能にするプラットフォームがないことである．しかし，そのようなプラットフォームを実装する際には次のような技術的な課題にも直面する．それは，“日々変容するウェブ空間の中から，どのように統計データを効率良く収集するか”という問題である．

よって本論文では，統計データに対して高いアクセシビリティをユーザに提供するプラットフォームを作成することを念頭に，ウェブ上に存在する統計データを効率よく収集するためのフォーカストクローリング手法について提案する．前提として本フォーカストクローラには探索期と収集期のふたつのフェーズが存在する．探索期においては，クローラは全てのウェブサイトの中から統計データを有するウェブサイトを見分ける．収集期においてクローラは，探索期で統計データを有すると判定したウェブサイトから収集できる全ての統計データを収集する．その前提において，本論文で提案する手法は探索期のクローラに未知のウェブサイトをどのように分類させるかということである．提案するフォーカストクローラでは，収集済みのウェブページに基づいて，そのウェブページを含むサイトが統計データを含むかを推定し，その推定値によって次の探索対象を決定するような適応的な探索方針を用いる．

具体的には，各サイトに対する判定誤り確率を推定しながら，次のクロール対象を決定するという方針をとる．あるサイトに対して現時点までの収集済みウェブページ群を用いて，元となるウェブサイトが統計データを有するか否かの判定を行う．同時に，その判定結果が誤りである確率と仮にそのサイトをもう一度クロールした際の判定誤り確率の推定値を求める．現時点の判定誤り確率から仮にもう一度クロールした際の判定誤り確率の推定値を引いた値を δ とするとき，この δ の値に基づいて次のクロール対象となるウェブサイトを決定する方針をとる．このような方針をとることの目的は，予め与えられた探索クロール (探索期におけるクロール) 回数において，各サイトに対して誤った判定をしてしまった数の最小化をすることである．δ の値が大きいサイトを優先的にクロールすることは，判定誤り確率が大きく下がるサイトをクロールすることになり，この方針で探索クロールを続けることで，予め与えられた上限回数内において判定誤り確率を最小化できると考えられる．

提案手法の有効性の確認のため，ランダムにウェブページを選択していく方針をベースライン手法とし，2 種類の実験を行って提案手法と比較する．その際，各サイトには統計データの有無について独立したベータ分布が与えられているという仮定を置く．このような仮定を置く背景については 3 節で説明する．一つ目の実験は，ランダムなベータ分布のパラメータに基づく疑似的にクロール環境のシミュレーション．二つ目の実験は，実際のウェブページによって構成される実データに対する比較実験である．結果としては，どちらの実験においても，少ない探索コストしかかけられない環境下においては，提案手法の方がベースライン手法より有効であることを示すことになった．

本論文における我々の貢献を以下に示す：(1) 予め定められた探索クロール回数で，対象となるデータの有無についてサイト単位で判別する手法を提案した点．(2) 仮定した確率分布に基づき，ある時点での各サイトに対する判定が誤りである確率を求め，それに基づいて次のクロール対象を決定する手法を提案した点．(3) 提案手法に関しては，統計データに限らず様々なコンテンツ・データ形式の有無に関する探索クロールに適用でき，実際の汎用性が高い点．

本論文の構成は以下の通りである．2 節ではフォーカストクローラについて，またそこで生じる探索/収集問題についてのアプローチに関する関連研究について述べる．3 節では本研究の問題設定および手法の詳細を説明し, その上で提案アルゴリズムの有効性に関する数学的な証明を行う．4 節では提案アルゴリズムに基づく 2 種類の実験の結果を示す．最後に，5 節では本論文の結論と今後の課題を述べる．

2 関 連 研 究

本節では，フォーカストクローリングに関する関連研究，および，フォーカストクローラが抱える探索/収集問題を解決するためのアプローチに関する関連研究を紹介する．

2.1 フォーカストクローリング

ウェブクローラとは，プログラムに基づいて自動的にあらゆるウェブサイトを訪問し，ウェブページを大量にダウンロードするためのシステムのことを言う．その用途は多岐に渡り，代表的な例はウェブ検索エンジンの検索対象の収集であると言える．その目的は，有用なウェブページの集まりを収集することである．基本的なウェブクローリングのアルゴリズムは以下のように単純である．ウェブクローラは種集合という URL 集合が与えられた際に，各 URL によってアドレスされる全てのウェブサイトをダウンロードし，そのページに含まれるハイパーリンクを抽出し，またそのハイパーリンクによって辿れるウェブページをダウンロードするという一連の所作を可能な限り行うものである．この際，標準的なウェブクローラには，どの URL を優先的にクロールするという方針を選択する機能は一切備わっていない．

その一方でフォーカストクローラは，特定のトピックに関連するデータをなるべく多く収集し，関連しないページのクロールを避けることを目的として，標準的なウェブクローラに選択方針を決めるアルゴリズムを備えたものである．フォーカストクローラに関する概念は Menczer によって最初に言及されたと言える [1]．その後，Chakrabarti et al. によってフォーカストクローラという言葉が作られた [2]．Aggarwal et al. の研究ではクローリング中に収集した情報を取り入れて，その後にクロールするページを定める手法を行い，関連トピックに関するページの最大化を行った [3]．その研究では，ウェブページの木構造において，当該ページの一つ上の深さのページ（親ページ）や確認済みの同じ深さのページ（兄弟ページ）のコンテンツから抽出した特徴量や，未確認ページの URL 文字列から得たトークンを特徴量として用いている．

上記のようなフォーカストクローリングの主流の研究においては，トピックベースで分類を行うことを目指す場合がほとんどである (e.g. スポーツ，政治，etc.)．一方で，本研究が目的とするのはウェブサイトに統計データが含まれていそうか否かという観点でウェブページを評価する点である．これについて Meusel et al. の研究が関連する [4]．この研究では microdata や RDFa などのマークアップ言語で記述されたセマンティックデータが HTML に埋め込まれているページを収集することを目的としている．セマンティックデータが埋め込まれたページを提供するようなウェブサイトには特定の性質があるとし，クロール最中にその性質を学習しながら，次にクロールするサイトを適応的に決定するものである．ただ，この研究とは次の 2.2 項で紹介するバンディットベースの選択方針のアルゴリズムにおいて異なる．

2.2 探索/収集アプローチに基づくクロール方針

多くのフォーカストクローリングシステムにおいては，短期的に得られる報酬（関連ページ）を最大化することに目的が置かれていることが多い．これは，実際に目的とするデータを有するか否かの判定が難しいサイトに対しては有効にクロールできない可能性を生じさせる．結果的に，本来収集可能だったデータを，各時点での最適解を優先するあまり，最終的に大幅に消失することも発生する．これは，いわゆる「探索/収集問題」である．そして，この問題への対処方法を探索/収集アプローチと呼ぶ．探索/収集アプローチの代表例は，バンディットアルゴリズムや強化学習だと言える．何よりバンディットアルゴリズムが前提とする多クラスバンディット問題自体が探索/収集問題をモデル化する一般的な方法だと考えられている．

このような経緯からクローリングにおいても，探索/収集アプローチとしてバンディットアルゴリズムをベースとした手法や強化学習手法を適用することが近年注目されている．Meusel et al. の研究では，クローリング問題をバンディット問題に変換し，バンディットベースのアルゴリズムを適用することで探索/収集問題を克服した [4]．本研究においてもバンディットベースのアルゴリズムを用いるのだが，ベースとなる方策は異なる．Meusel et al. の研究では ϵ-greedy に基づく方策を用いているのに対して，本研究の提案手法は確率一致法や Thompson Sampling に基づいていると言える．また，強化学習手法をクローリングに適用した例もある [5] [6]．その他，クローリングへの適用ではないが，広告推薦システムなどにおいて，部分フィードバックに基づく多クラス分類をバンディットアプローチで行うモデルの研究は盛んである [7] [8] [9] [10].

これらの研究を参考に本研究においては，各ラウンドにおいてページ内容を入力とし，元となるサイトが統計データを扱うか否かを推定し，その上で部分的なフィードバックとして，その判定が誤りである確率をさらに推定し，それによって次の探索対象を決定するというバンディットベースのクロール方針を採る．

3 提 案 手 法

本節では，本研究が対象とする問題についての基本的な問題設定を行う．次に，提案アルゴリズムがもつ特殊な設定の説明と，その設定下における基本方針を説明する．

3.1 基本的な問題設定

S を n 個の要素からなるウェブサイト集合とする．各サイトはウェブページによって構成されており，その中に一つでも統

計データを含むページがあれば，元のサイトは統計データを持つサイトと定義する．その一方で，統計データを含むページが一つもなければ，元のサイトは統計データを持たないサイトとして定義する．T はクローラが全サイトに対して探索的クロールをすることができる回数の上限と定義する．t は 0 から T までの値をとり，ある時点までのクロール回数を表す．また，$t_1,\ldots,t_n$ は各 n 個のサイトに対するある時点までのクロール回数を表し，$t = t_1 + t_2 + \cdots + t_n$ となる．各時刻 t には，一つのウェブページをクロールすることができる．そして同時に，そのページを，元となるサイトが統計データを持つ確率 $p \in [0,1]$ を出力する分類器に通す．$i = 1,\ldots,n$ とするとき，あるサイト $s_i \in S$ に対して，t_i 回クロールした時点において，そのサイトが統計データを持っているかいないかの判定は $p_{i,1}, p_{i,2}, \ldots, p_{i,t_i}$ の平均値が 0.5 以上か未満かによって行う．我々の目的は，これらの条件を前提としたとき，どのような方針でウェブページを選択すれば，T 回クロールした後になるべく多くのサイトに対して統計データを持つか否かを正確に推定できるかという問題を解くことである．

3.2 提案アルゴリズムにおける特殊設定と基本方針

3.2.1 特 殊 設 定

我々はこの問題に対して，都度どのサイトのページをクロールするべきかを確率的に予測し決定するというアプローチをとる．まず，各サイトが統計データを持つか否かは，各サイトに対して一度だけベルヌーイ試行をした結果であると考える．統計データを持つサイトはベルヌーイ試行の結果が 1 であり，持たないサイトは 0 である．ただ，このベルヌーイ試行は，何の根拠もなしに行われた訳ではない．全サイトにはベルヌーイ試行の背景となる何かしらの独立同一分布 (i.i.d) が与えられており，それが事前分布となっていたため，各サイトに対する今回の結果が得られたと考える．ここで，本手法においては，全てのサイトには独立同一分布として“ベータ分布”が与えられていることを仮定する．そして，各サイトのベータ分布を事前分布とし，そのパラメータに基づいてベルヌーイ試行を行った結果が，そのサイトが実際に統計データを有するか否かという事実であるとする前提の下，以降の議論を進める．ウェブサイト $s_i \in S$ に対応して y_i は前述のベルヌーイ試行の結果を保持した値であるとする．s_i が統計データ有りであれば $y_i = 1$，無しであれば $y_i = 0$ とする．この情報を各 n 個のサイトに対応して保持した集合を Y とする．

上記の前提の下で，Y の各値を知ることは，S の全サイトの全ページを探索しない限りは観測不可能である．その上で，あるサイト s_i が統計データを有する確率を $P(y_i = 1) = \theta_i$ と表すとすると，サイト s_i に対するベータ分布のパラメータを α_i, β_i とした時，θ_i は以下のように表せる．

$$\theta_i = E[y_i|\alpha_i,\beta_i] = \frac{\alpha_i}{\alpha_i + \beta_i} \tag{1}$$

この時，$E[y_i|\alpha_i,\beta_i]$ は α_i,β_i に基づく y_i の値に対する期待値である．このことから，θ_i は y_i の値に対する期待値によって求められる．この θ_i に基づいて y_i の確率 $P(y_i|\theta_i)$ は次の式で表される．

$$P(y_i|\theta_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i} = \left(\frac{\alpha_i}{\alpha_i+\beta_i}\right)^{y_i}\left(\frac{\beta_i}{\alpha_i+\beta_i}\right)^{1-y_i} \tag{2}$$

しかし，現実にはサイト s_i が持つベータ分布自体を知ることはできないため，これを推定するしかない．よって以下の方法に基づいて推定を行う．まず，サイト s_i に対して t_i 回クロール済みであるとするとき，それらクロール済みのページを $s_{i,1}, s_{i,2}, \ldots, s_{i,t_i}$ のように表す．次に各クロール済みページに対して，ページの内容から元のサイトが統計データを有する確率を推定する分類器に通した値を，各サイトに対応するベータ分布から得た標本値として扱い，$p_{i,1}, p_{i,2}, \ldots, p_{i,t_i}$ のように表すとする．これらを用いて，サイト s_i に対する t_i 回目のクロール時点での θ_i の推定は，以下のように行う．

$$\hat{\theta}_{i,t_i} = E[y_i|\hat{\alpha}_{i,t_i},\hat{\beta}_{i,t_i}] = \frac{\hat{\alpha}_{i,t_i}}{\hat{\alpha}_{i,t_i}+\hat{\beta}_{i,t_i}} = \frac{1}{t_i}\Sigma_j^{t_i} p_{i,j} \tag{3}$$

ここで $\hat{y}_{i,t_i}$ をサイト s_i に対して t_i 回目までのクロールから得られた $\hat{\theta}_{i,t_i}$ に基づいて 0 か 1 を返すものであると定義する．具体的には，$\hat{\theta}_{i,t_i} < 0.5$ ならば $\hat{y}_{i,t_i} = 0$，$\hat{\theta}_{i,t_i} \geqq 0.5$ ならば $\hat{y}_{i,t_i} = 1$ を返す．$\hat{\theta}_{i,t_i}$ に基づく $\hat{y}_{i,t_i}$ の値の確率は次の式で表される．

$$P(\hat{y}_{i,t_i}|\hat{\theta}_{i,t_i}) = \hat{\theta}_{i,t_i}^{\hat{y}_{i,t_i}}(1-\hat{\theta}_{i,t_i})^{1-\hat{y}_{i,t_i}} \tag{4}$$

$$= \left(\frac{\hat{\alpha}_{i,t_i}}{\hat{\alpha}_{i,t_i}+\hat{\beta}_{i,t_i}}\right)^{\hat{y}_{i,t_i}}\left(\frac{\hat{\beta}_{i,t_i}}{\hat{\alpha}_{i,t_i}+\hat{\beta}_{i,t_i}}\right)^{1-\hat{y}_{i,t_i}} \tag{5}$$

3.2.2 基 本 方 針

このアルゴリズムの基本方針は $\forall s \in S$ に対する，統計データの有無の判定の誤りを最小化することである．判定の誤りを最小化する際には，各回の判定誤りの差を確認する必要がある．ここで，δ という値を用いる．δ は各サイトに対して，現時点における判定誤り確率から，仮に次にもう一度そのサイトをクロールした時の判定誤り確率の推定値を引いた値と定義する．その上で全サイトの δ の内，最大の値をもつ δ_i に対応したサイト s_i をクロールする方針を取れば，最終的な累積判定誤り確率は最小化されると考える．よって本手法は，予め与えられた探索クロール回数 T において，δ の値に基づく，確率一致法，Thompson Sampling をベースにした手法である．その上で，手法の詳細について以下で数学的に定式化する．

サイト s_i に対して，e_{i,t_i} は t_i 回目までのクロールに基づく，判定誤りに関する情報を保持するものである．e_{i,t_i} は，実際に統計データを持っているか否かの真の情報 y_i と t_i 回目までのクロールから得られる $\hat{y}_{i,t_i}$ が等しくないとき 1，等しいとき 0 を保持する．数式では，次のように表される．

$$e_{i,t_i} = \begin{cases} 1 & (y_i \neq \hat{y}_{i,t_i}) \\ 0 & (y_i = \hat{y}_{i,t_i}) \end{cases} \tag{6}$$

つまり，e_{i,t_i} は s_i への t_i 回目までのクロール時点で，s_i が統計データを持つか否かに対する推測の正誤に関する情報を記録する関数である．そして全ての n 個のサイトについて判定誤りの情報を保持したものを E と定義する．

e_{i,t_i} が 1 である確率 $P(e_{i,t_i} = 1)$ は数式では，以下のように表される．

$$P(e_{i,t_i}=1)=\begin{cases} P(y_i=1) & (\hat{y}_{i,t_i}=0) \\ P(y_i=0) & (\hat{y}_{i,t_i}=1) \end{cases} \tag{7}$$

しかし，$P(y_i)$ に関する実際の値は分からないので，$P(e_{i,t_i}=1)$ について次の式によって推定する．

$$\hat{P}(e_{i,t_i}=1)=\begin{cases} \hat{\theta}_{i,t_i} & (\hat{y}_{i,t_i}=0) \\ 1-\hat{\theta}_{i,t_i} & (\hat{y}_{i,t_i}=1) \end{cases} \tag{8}$$

ここまでは，実クロールに基づく順当な予測である．その上で，次にどのサイトをクロールするべきかを決定するために t_i+1 回目の結果を予測する必要がある．そのため t_i 回目までのクロール結果に基づいて，t_i+1 回目のクロールをすることで得られる $\hat{\theta}_{i,t_i+1}$ の予測値を z_{i,t_i} と定義する．

この z_{i,t_i} は，これまでの議論に基づけば，サイト s_i から得られた t_i 個の標本値 $p_{i,1},p_{i,2},\ldots,p_{i,t_i}$ の平均値 $\hat{\theta}_{i,t_i}$ を $t_i=1$ の時点から，$t_i=t_i$ の時点まで記録しておき，それらの値に基づく分布を考えたものである．すなわち z_{i,t_i} は，各サイトが統計データを持つ確率についての標本平均の分布であると言える．そして，元の標本がどのような分布に基づいていたとしても，標本平均の分布は中心極限定理に基づいて，正規分布に収束する．このことから，z_{i,t_i} は次のような正規分布になると仮定する．

$$z_{i,t_i} \sim N\left(E[y_i|\alpha_i,\beta_i], \frac{V[y_i|\alpha_i,\beta_i]}{t_i}\right)$$

この z_{i,t_i} に基づく $\hat{P}(e_i=1)$ の t_i+1 回目の予測確率を $\hat{P}(e_{i,t_i+1}=1|z_{i,t_i+1})$ として，次の式によって推定する．

$$\hat{P}(e_{i,t_i+1}=1|z_{i,t_i+1})=\begin{cases} z_{i,t_i} & (y_{i,t_{i+1}}=0) \\ 1-z_{i,t_i} & (y_{i,t_{i+1}}=1) \end{cases} \tag{9}$$

これを用いて，$\hat{P}(e_{i,t_i+1}=1)$ を推定すると以下のようになる．

$$\begin{aligned} \hat{P}(e_{i,t_i+1}=1) &= \int_0^1 \hat{P}(e_{i,t_i+1}=1|z_{i,t_i+1})P(z_{i,t_i})dz_{i,t_i} \\ &= \int_0^{\frac{1}{2}} z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} + \int_{\frac{1}{2}}^1 (1-z_{i,t_i})P(z_{i,t_i})dz_{i,t_i} \end{aligned} \tag{10}$$

付録において行う証明に基づけば，式 (8) から式 (10) を引いた値は必ず非負になる．明示的に言い換えれば，$\delta_{i,t_i}=\hat{P}(e_{i,t_i}=1)-\hat{P}(e_{i,t_i+1}=1)$ とするとき，δ_{i,t_i} は常に非負である．この δ_{i,t_i} が大きいサイトというのは，一回のクロールによってそのサイトの判定誤り確率が大きく下がるサイトであることを意味する．探索クロールに掛けられる回数 T が有限という制約下において，全てのサイトに対する判定誤り確率を可能な限り最小化したい場合，この δ_{i,t_i} が最大となるサイト s_i を優先的にクロールすることが最適な選択方針であると考えられる．

4 実　　験

本節では，提案手法の有効性の検証及び評価のために行った実験を紹介する．実験は 2 種類行った．一つ目は，ランダムに生成されるベータ分布に対して行ったシミュレーション．二つ目は，実際のウェブページから構成される実データに対して行った実験．各セクションの冒頭で各実験の目的について述べるが，端的に言えば，一つ目のシミュレーションは，提案手法で置いた確率分布に関する仮定がランダムに生成されたパラメータに基づくベータ分布において普遍的に通用するのか検証する目的で行った．二つ目の実データ実験は，提案手法とベースライン手法の性能を実際のクロール環境において比較し，提案手法の有効性を確かめる目的で行った．

4.1 ランダムなベータ分布に基づくシミュレーション

本節では，ランダムなベータ分布に基づくシミュレーションについて述べる．提案手法において，各サイトには，統計データの有無に関して独立同一分布としてベータ分布が与えられていることを仮定した．しかし，実際には必ずしもベータ分布に従うとは限らないため，この仮定が成立しない場合も想定される．あるいは，ベータ分布に従うとしても，一部の特殊なベータ分布には従うのみで，全てのベータ分布に普遍的に従うわけではないという可能性もある．よって，それらを検証するために，ランダムに生成されるパラメータに基づくベータ分布において，擬似的にクロール環境をシミュレーションし，本提案手法とベースライン手法の比較を行うことで，上述の問題について検証を行う．4.1.1 では，シミュレーションにおける環境設定について説明する．4.1.2 では，シミュレーション結果の提示と考察を行う．

4.1.1 シミュレーション環境設定

3.2.1 における特殊設定にあるように，本研究においては，全てのウェブサイトには独立同一分布としてベータ分布が割り当てられていることを仮定している．本シミュレーションにおいては次のような設定を行った．全サイト数は 1,000 件である．そして各サイトに対して，2 つの実数値 α,β を一様分布に従ってランダムに生成する．これによって，各サイトはその 2 値をパラメータとするベータ分布に従うことになる．そのベータ分布を事前分布として，ベルヌーイ試行を行い，その結果を各サイトが統計データを持っているか否かの情報とする．両手法において，ベータ分布のパラメータおよびベルヌーイ試行の結果は共有されている．ページ数という概念については，シミュレーションにおいては存在しないことにする．しかし，実装上は各サイトに対して，ベータ分布の標本 $p\in[0,1]$ を 1,000 個ずつ生成し，各サイトへのクロール回数をインデックスとして用いることで参照できるようにしている．仮に，与えられた探索回数上限の内で，一つのサイトに対する全ての標本を取得し尽くしてしまった場合は，それ以上そのサイトへの探索はできないものとする．

ベースライン手法と提案手法に共通する前提として，全てのサイトに対して一度は対応ページのクロールを行う．以降便宜的にこのことを「初期クロール」と名づける．今回は 1,000 件のサイトを設定しているので，初期クロールとして 1,000 回のクロールを行う．そしてこの初期クロールは，予め与えられた探索クロール回数の上限の中で行うものとする．よって，例えば探索クロールに対して 5,000 回の上限が課せられていた場合，そのうち 1,000 回は初期クロールを行い，提案手法やベースラ

イン手法におけるサイトの選択は残り 4,000 回のクロールにおいて行われるものとする．その上で，我々はベースライン手法として，クロールするサイトをランダムに選定する方法を設定した．ベースライン手法においては，取得可能なページが存在する限り，与えられた上限回数の範囲内であればいくらでもクロールし続けることができる．提案手法においては，初期クロールの結果を元に各サイトの δ を計算し，以降各 t 回において，その値が一番大きいサイトをクロールして，そのサイトの δ を更新する．与えられた上限回数の範囲内において，全てのサイトに対する δ の値が 0 になった場合，その時点で提案手法における探索クロールは終了する．

4.1.2 シミュレーション結果

図 1，図 2，図 3 はそれぞれシミュレーションにおける精度，適合率，再現率の推移のグラフである．今回は両手法において，20,000 回クロールしている時点で，精度と適合率は収束しており，再現率に関しては，提案手法は収束した一方ベースライン手法は単調増加していることが確認できたため，探索クロール回数上限 20,000 回までの推移を載せた．

本シミュレーションにおける前提として，始めにランダムに生成したベータ分布のパラメータが違えば，結果は異なったものになる．また，統計データを持っていることになっているサイトの数も試行毎に変化する．ベータ分布のパラメータ α,β は一様分布に従って生成されるため，試行毎におおむね 470 から 530 サイトが統計データを持つ状態になることが 10 回の試行から観測されている．

そのように，試行毎に多少の変化は生じるが，基本的な傾向は一貫した結果が得られた．それは，精度，適合率は両手法ともほぼ同じで，再現率に関しては最初の数千回までは提案手法の方が高く，ある程度回数が経過するとベースライン手法の方が提案手法を上回ることである．シミュレーション結果において提案手法とベースライン手法を比較する上では，精度および適合率は相対的に重要な指標ではないと考えられる．なぜなら，統計データの有無に関して均衡状態なので，初期段階とクロール回数が増えた段階での差がほとんど見られず図 1 の精度や図 2 の適合率のように横ばいになるからである．よって，重要な指標は再現率であると言える．図 3 の再現率において，初期クロールを含めて，最初の 3,000 回程度までは提案手法の方が再現率が高いことが示されている．このことから，探索クロールにかけられるコストが少ない条件下 (各サイトに対して平均 2,3 回程度) においては，提案手法の方が成績が高いという結果を示している．それは同時に，十分な探索コストが割り当てられている場合においては，ベースライン手法の方が適していることも示している．

上述の結果は，シミュレーションの全試行において同様の結果を得られたことから，本問題を扱うにあたって，事前分布としてベータ分布を用いることは有効であることが推測される．また，ベータ分布をどのようにランダムに生じさせたとしても同様の結果が得られることも確認できたと言える．

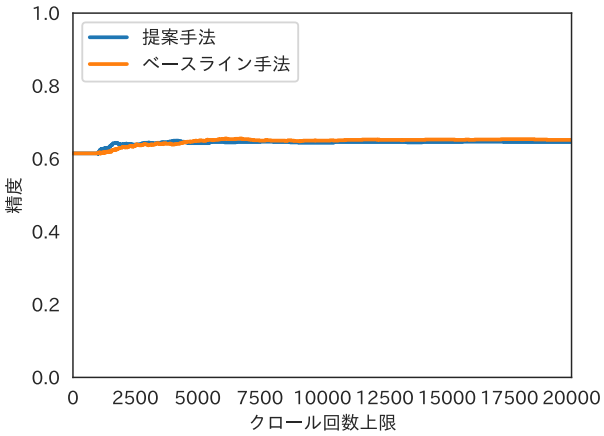

図 1　シミュレーションにおける精度の図

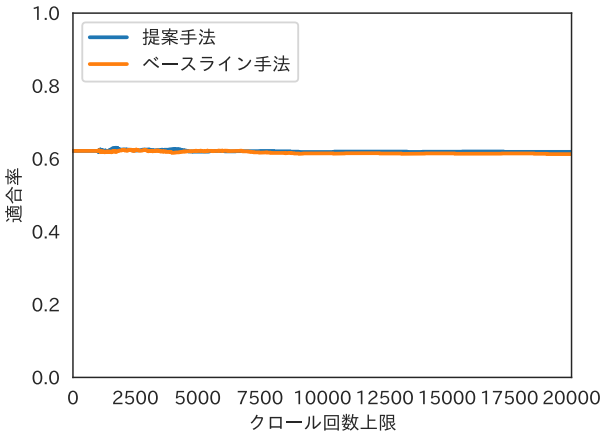

図 2　シミュレーションにおける適合率の図

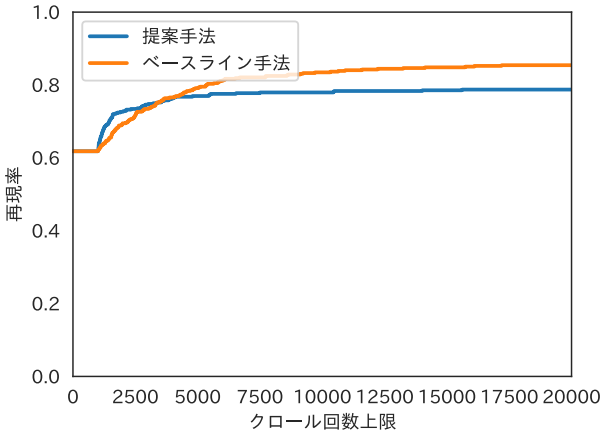

図 3　シミュレーションにおける再現率の図

4.2 実データを用いた実験

本項では，実際のウェブサイトから構成される実データに対して行った実験について述べる．4.2.1 では，実験に使用したデータセットについて，その収集方法と定量的な情報を述べる．4.2.2 では，本実験における設定について紹介する．そして，4.2.3 で実験結果の提示と考察を行う．

4.2.1 データセット

関連研究 [4] と同様に，ページのコンテンツやウェブグラフの変化，ウェブページのホストの可用性などの他の要因から切り離すために，実験には静的データセットを採用した．以下の実験で使用するすべてのデータセットは，Common Crawl Foundation によって提供された一般にアクセス可能な日本語ウェブページから抽出されたリンクに含まれる各サイトに対して，深さ 1 から 3 のページを網羅的に収集したものである．ページの深さはトップページからの最短距離として定義される．また，サイトによっては膨大な数のページが存在するため，深さの制約に加え，1 サイトあたりのクロールページ数の上限を 10,000 とした．クロールされた各ページはデータ分類器に入力され，そのページが統計データへのリンクを含んでいるかどうかを判定する．この分類器は，ファイル中の数字表現の割合などの標準的な特徴に基づく Gradient Boosted Decision Tree モデルで，2,505 ファイルで学習し，99.4%の精度を達成することが示された．その結果，16,819 サイト，3,904,590 ページが得られ，そのうち少なくとも 1,976 サイトが統計データファイルを含んでいる．

その中で，実際に使用したのは無作為抽出した 2,000 サイト，457,741 ページである．このうち統計データを含むサイト数は 239 個で，全体の 11.95%である．その上で，ページ内容を MeCab を用いて形態素解析し，そこに含まれる名詞から元サイトが統計データを持つ確率を推定する分類器をサポートベクターマシンを用いて作成した．特徴量はページ内容を tf-idf でベクトル化したものである．また，データセットの不均衡性を解消するため，コスト考慮型学習を行った．2,000 サイトのうち 800 サイト，187,204 ページを学習用データ，200 サイト，48,659 ページを精度と適合率の評価用データとして用いた．このうち，学習用データには 89 個 (11.125%) が統計データを含むサイト，評価用データには 25 個 (0.125%) の統計データを含むサイトであった．評価用データの各サイトに対して，各サイトの全てのページを用いて，サイトが統計データを持っている確率の平均値が 0.5 以上ならば当該サイトは統計データを持つ，0.5 未満ならば持たないという判定を行った．この判定結果に対して，各サイトが実際に統計データを持っているか否かの情報と比較することで，精度と適合率の評価を行った．結果として，精度は 89%，適合率は 92.38%となった．そして残りの 1,000 サイト，221,878 ページを実験用データとして用いた．このうち，統計データを含むサイト数は 125 個 (12.5%) であった．

4.2.2 実験設定

用いるデータが 4.2.1 で説明した実データであることを除いて，基本的に 4.1.1 のシミュレーション環境設定と全く同じ設定を置く．具体的には，初期クロールはサイト数の 1,000 回行うこと．ベースライン手法はサイトをランダムに選定する方針をとること．提案手法は，δ に基づくクロール方針をとり，探索クロール上限内で，全てのサイトに対する δ の値が 0 になった場合，提案手法の探索クロールは終了することなどである．また，ある時点で全てのページが取得済みになったサイトに関してはクロール不可能サイトとして扱い，いずれの手法においても，各時点においてクロール可能状態なサイトを選択する．

4.2.3 実験結果

前提として，同じデータセットを用いた場合，提案手法では精度，適合率，再現率の全てにおいて以下に示す結果とほぼ一意の値が得られるが，ベースライン手法においては，完全な無作為選択であるため，実行するたびに若干結果は異なる．

図 4，図 5，図 6 に示すのはそれぞれ本実験における精度，適合率，再現率の推移のグラフである．今回は両手法において，20,000 回クロールしている時点で全ての値においてほぼ収束していることが確認できたため，探索クロール回数上限 20,000 回までの推移を載せた．

精度に関しては図 4 の通り両手法において大局的には横ばいである．ただ，最終的な値は，両手法ともにクロール開始時より若干の低減が見られる．原因としては，データセット自体が非均衡データであるため，統計データがないという判定を多くすれば自然と精度が高くなる傾向にあることが挙げられ，クロールすればするほど，本来統計データを持っていないサイトに対しても，統計データを持っているという判定をする確率が高くなっていると考えられる．その点においては，提案手法がベースライン手法より 3,000 回ほど早い段階で低下していることから，提案手法の方が“本来統計データではないデータを，統計データありと誤判定しやすい傾向にある”と言える．その傾向は図 5 での適合率の低下からも明らかである．ここで適合率とは，“統計データありと判定したサイトが，本来統計データであった割合”を示すため，先述の傾向を直接的に示す．よって適合率が低下していることは，上記傾向を裏付ける．しかしここで重要なのは，探索期のクロールにおいては，この傾向は決して悪いことではないという点である．特に今回のような非均衡データにおいて，“実際には統計データを持っているサイトを持っていないと判定してしまうこと”の重みの方が，“実際には統計データを持っていないサイトを持っていると判定してしまうこと”の重みよりも圧倒的に大きい．後者の問題に関しては，収集段階のクロールにかけるコストが十分にかけられていれば，相対的に軽微な問題であり，探索期においてはむしろある程度は奨励されるべき傾向だと言える．

両手法の性能比較において一番重要な指標は再現率だと言える．図 6 に再現率を示す．ある程度の回数クロールをした段階においては，両手法の再現率において大きな差はない．重要なのは，クロール開始後 5,000 回以内における比較であると言える．その段階においては，シミュレーション同様，提案手法の再現率の高まり具合はベースライン手法より早いことが確認できる．この再現率の結果が説明するのは次のことである．(1) 与えられたページから元サイトが統計データを持っている確率

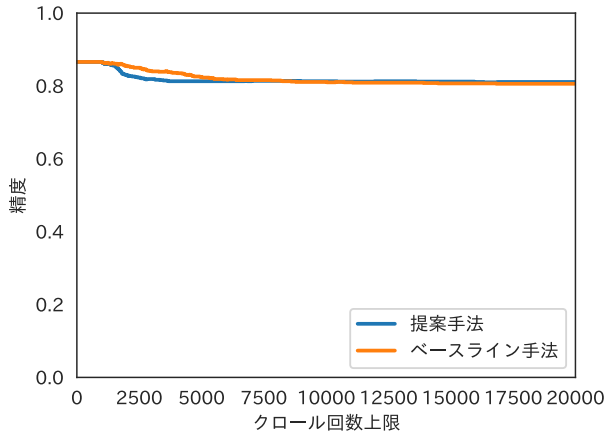

図 4　実データ実験における精度の図

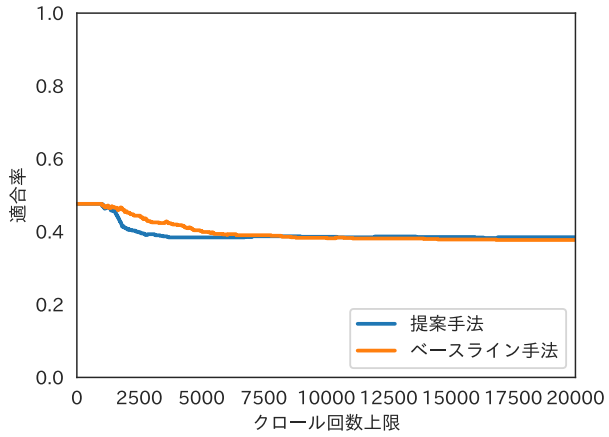

図 5　実データ実験における適合率の図

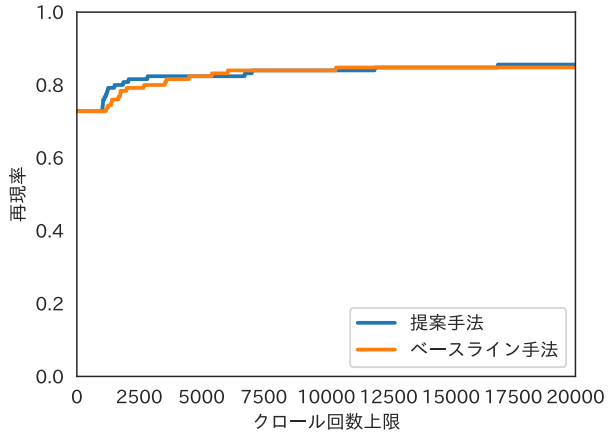

図 6　実データ実験における再現率の図

をある程度しっかりと予測することができれば，初期クロールの段階でもある程度高い再現率でサイトを分類することができる．(2) そこからさらに，少ない探索コスト (各サイトに対して平均 2,3 回程度) で，より多くの本来統計データを含むサイトを正しく分類するという点では，提案手法の方が良い成績を出している．特に，実際のシステムとしてフォーカストクローラを実装することを考慮したとき，(2) の観点は重要である．実際のウェブ上の数億のサイトをクロールの対象とする場合，探索にかけられるコストは決して多くはないことが考えられる．よって，本実験結果からは，探索クロールにかけられるコストが限られている前提においては，提案手法はベースライン手法より有効であることが確認できたと言える．

5　ま　と　め

本論文では，ウェブ上から統計データを収集することを目的としたフォーカストクローラを提案した．その提案手法は，バンディットベースの探索/収集アプローチに基づいて，次にどのサイトのページを取得するかを決定することを探索クロール方針とするものである．その上で，本手法は予め少ない探索クロール回数 T を設定し，その中でなるべく多くのサイトに対して正しい判定を行うことを目的としたものである．

提案手法では，全てのサイトには統計データの有無に関する事前分布となるベータ分布が割り当てられていると仮定し，クロールしたページを分類器に通すことで，元サイトが統計データを持つ確率を得ることができるものとした．クローラは，予め与えられた探索クロール回数において，各時点までのクロール済みのウェブページに基づいて，そのウェブページを含むサイトが統計データを含む確率を推定する．その確率から，そのサイトに対する現時点の判定誤り確率と次にそのサイトのページをクロールした際の判定誤り確率の推定値の差 δ を求め，この値に応じて次の探索対象を決定する．その意味で本クロール方針は，確率一致法や Thompson Sampling をベースとする方針であると言える．

提案手法の有効性の確認のため，ランダムにウェブページを選択していく方針をベースライン手法とし，提案手法に対して 2 種類の比較実験を行った．一つ目はランダムに生成したベータ分布のパラメータに基づくクローリングの疑似的なシミュレーション，二つ目は実際のウェブページを元にした実データに対する比較実験である．それらについて，精度，適合率，そして再現率を元に比較評価を行った．結果としては，いずれの実験においても，少ない探索クロール回数における提案手法の有効性を示す結果となった．

今後，より提案手法の有効性を正確に示すために，同じデータセットや同様の条件下において，他のバンディットアルゴリズムにおける方針 (ϵ-greedy や UCB など) を用いた際との比較実験を行うことを考えている．今回実験段階においては，本提案手法が予め与えられた探索クロール回数 T と δ という特殊なパラメータを用いてはいるが，確率一致法や Thompson Sampling に基づくバンディットベースのアルゴリズムであることを調査

しきれていなかった．そのため，実験段階で他のバンディットアルゴリズムの方針との比較を行うことができていなかったという点がある．よって，それについて今後詳しく調査したい．

謝辞　本研究は JSPS 科研費 18H03244，および，JST さきがけ JPMJPR1853 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Filippo Menczer. Arachnid: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery, 1997.

[2] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Comput. Netw.*, 31(11–16):1623–1640, may 1999.

[3] Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 96–105, New York, NY, USA, 2001. Association for Computing Machinery.

[4] Robert Meusel, Peter Mika, and Roi Blanco. Focused crawling for structured data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1039–1048, New York, NY, USA, 2014. Association for Computing Machinery.

[5] Miyoung Han, Pierre-Henri Wuillemin, and Pierre Senellart. Focused crawling through reinforcement learning. In *ICWE '18*, 2018.

[6] Jason Rennie and Andrew McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 335–343, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[7] Koby Crammer and Claudio Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Mach. Learn.*, 90(3):347–383, mar 2013.

[8] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 440–447, New York, NY, USA, 2008. Association for Computing Machinery.

[9] Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via lipschitz context multi-armed bandits. In *Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[10] Shijun Wang, Rong Jin, and Hamed Valizadegan. A potential-based framework for online multi-class learning with partial feedback. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 900–907, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

付　録

3 節の提案手法の説明の中で，δ が常に非負の値であることの証明をすることは，読者が全体的な論理展開の流れを把握することを妨げかねないと判断し，式変形や証明を省いて記述している．そのため，この付属の章においてそれを補完する．式 (8) から式 (10) を引いた値 δ が非負であることの証明を行う．

A.1　$\hat{\theta}_{i,t_i} < 0.5$ のとき

式 (10) を変形すると以下のようになる．

$$\begin{aligned}\hat{P}(e_{i,t_i+1}=1) &= \int_0^{\frac{1}{2}} z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} + \int_{\frac{1}{2}}^1 z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} - \\ &\quad \int_{\frac{1}{2}}^1 z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} + \int_{\frac{1}{2}}^1 (1-z_{i,t_i})P(z_{i,t_i})dz_{i,t_i} \\ &= \int_0^{\frac{1}{2}} z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} + \int_{\frac{1}{2}}^1 (1-2z_{i,t_i})P(z_{i,t_i})dz_{i,t_i}\end{aligned}$$

仮定より $E[z_{i,t_i}] = \int_0^1 z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} = E[y_i|\hat{\alpha}_{i,t_i},\hat{\beta}_{i,t_i}] = \hat{\theta}_{i,t_i}$ であるから，

$$\hat{P}(e_{i,t_i+1}=1) = \hat{\theta}_{i,t_i} + \int_{\frac{1}{2}}^1 (1-2z_{i,t_i})P(z_{i,t_i})dz_{i,t_i}$$

$\hat{\theta}_{i,t_i} \leqq 0.5$ のとき，$\hat{P}(e_{i,t_i}=1) = \hat{\theta}_{i,t_i}$ であるから，

$$\hat{P}(e_{i,t_i+1}=1) - \hat{P}(e_{i,t_i}=1) = \int_{\frac{1}{2}}^1 (1-2z_{i,t_i})P(z_{i,t_i})dz_{i,t_i}$$

ここで定積分の定理より，$a \leqq x \leqq b$ において，

$$f(x) \leqq g(x) \Rightarrow \int_a^b f(x)dx \leqq \int_a^b g(x)dx$$

であることより，$f(x) = (1-2x)P(x), g(x) = 0$ としたとき，範囲 $\frac{1}{2} \leqq x \leqq 1$ において，$f(x) \leqq 0$ である．よって，$f(x) \leqq g(x)$ であるから，

$$\int_{\frac{1}{2}}^1 (1-2z_{i,t_i})P(z_{i,t_i})dz_{i,t_i} \leqq 0 (= \int_{\frac{1}{2}}^1 0dx)$$

よってこれが負であると言える．

A.2　$\hat{\theta}_{i,t_i} \geqq 0.5$ のとき

式 (10) は次のようにも変形される．

$$\begin{aligned}\hat{P}(e_{i,t_i+1}=1) &= \int_0^{\frac{1}{2}} z_{i,t_i}P(z_{i,t_i})dz_{i,t_i} - \int_0^{\frac{1}{2}} (1-z_{i,t_i})P(z_{i,t_i})dz_{i,t_i} \\ &\quad + \int_0^1 (1-z_{i,t_i})P(z_{i,t_i})dz_{i,t_i} \\ &= \int_0^{\frac{1}{2}} (2z_{i,t_i}-1)P(z_{i,t_i})dz_{i,t_i} + 1 - \hat{\theta}_{i,t_i}\end{aligned}$$

$\hat{\theta}_{i,t_i} \geqq 0.5$ のとき，$\hat{P}(e_{i,t_i}=1) = 1 - \hat{\theta}_{i,t_i}$ であるから，

$$\hat{P}(e_{i,t_i+1}=1) - \hat{P}(e_{i,t_i}=1) = \int_0^{\frac{1}{2}} (2z_{i,t_i}-1)P(z_{i,t_i})dz_{i,t_i}$$

A.1 のときと同様に定積分の定理より，

$$\int_0^{\frac{1}{2}} (2z_{i,t_i}-1)P(z_{i,t_i})dz_{i,t_i} \leqq 0$$

よって A.1 と A.2 より，

$$\hat{P}(e_{i,t_i+1}=1) \leqq \hat{P}(e_{i,t_i}=1) \tag{A·1}$$

このことから，サイト s_i に対してクロール回数 t_i が増えるごとに推定判定誤り確率は小さくなっていくと考えられる．よって，式 (8) から式 (10) を引いた値 δ は必ず非負であると言える．