社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (11)
2008－BIO－15 (11)
2008／12／17

Webアーカイブにおける差分収集に用いる更新間隔推定手法の開発とその評価

廣道 尚弓 * 石川 千里 † 高田 雅美 † 城 和貴 †
nao214@ics.nara-wu.ac.jp
* 奈良女子大学 理学部 情報科学科
† 奈良女子大学大学院 人間文化研究科 複合現象科学専攻

概要

本稿では，Webページの更新間隔を推定する手法を開発する．この手法を用いてWebページを差分収集するデータベースを構築する．これは，Webアーカイブにおいて，Webページの更新，変更，削除に対応するために用いられる．開発にあたって，Webページが更新される間隔の分布型を予め仮定する．そのため，ブートストラップ法が適用可能である．

Development of the Update Space Estimate Technique to Use for Collection for a Difference in the Web Archive and the Evaluation

Naomi Hiromichi* Chisato Ishikawa† Masami Takata† Kazuki Joe†
* Department of Information and Computer Sciences, Faculty of Science, Nara Women's University
† Graduate School of Humanities and Sciences , Nara Women's University

Abstract

In this report, We develop a technique to estimate the update period of a Web page. We build a database collecting a Web page to validate this technique. This is used corresponding to the update, the change and the deletion of the Web page in a Web archive. On development, We assume the distribution type of the interval when a Web page is updated beforehand. Therefore We can apply a boot strap method.

1 はじめに

WebアーカイブとはWebページの保存・公開を目的とするアーカイブでインターネット上の図書館としての役割を担っている．最大規模のWebアーカイブ機関は，Web全体のアーカイブ作成を行っているInternet Archive[1]である．ここでは，1996年以降の全世界のWeb情報550億ページを包括的に収集し，2001年からネット上で公開している．. 2007年から高性能収集ロボットHeritrixによる世界のWebサイト200億ページを収集するプロジェクト「around the World in 2Billion Pages」を実施している．Internet Archive以外にも，各国の国立図書館を中心にWebアーカイブの開発が進められている．代表的なWebアーカイビングプロジェクトとしては，米国議会図書館によるMINERVA[2]，英国図書館によるUK Web Archiving Consortium[4]，スウェーデン王立図書館によるKulturarw3[3]，オーストリア国立図書館によるAOLA（Austrian On-Line Archive）[5]，オランダ国立図書館によるe-Depot[6]，オーストラリア国立図書館によるPANDORA[7]，などがある．各プロジェクトは，収集方針として，選択的収集かバルク収集を採用している．選択的収集とは，手作業によるWeb情報の選択的な収集である．この収集方法は特定のテーマや著作権などの許諾を得たWebページを選択的に収集するため，精度の高いアーカイブ構築が可能である．その反面，数多くのページを収集することは困難であるという性質を持つ．一方，バルク収集とは，自動収集ソフトウェアによるWeb情報の一括収集のことである．Web上のデータを国単位，あるいは世界全体など非常に膨大な規模で収集を行うため，大規模なアーカイブ構築が可能である．しかし選択的収集と比較すると低コストという利点があるものの，著作権等の法的問題を内包する上，玉石混交のアーカイブとなってしまうという欠点がある．

日本では，国立国会図書館がWARP（Web ARchiving Project）[8]を行っている．WARPでは選択的収集を採用しており，ウェブサイトと電子雑誌の収集を行っている．ウェブサイトとしては，国の機関，都道府県，政令指定都市，法定合併協議会およびその構成市町村，特殊法人等，大学法人化以前の国立大学，国際的・文化的イベント等のホームページを主に対象としている．電子雑誌とは同一のタイトルのもとに，終期を予定せず，巻次・年月次等の表示を伴って，継続的に発行されるネットワーク電子情報であると定義す

る．現在はインターネット上で無償公開されている電子雑誌を，WARP の収集対象としている．収集の頻度は，ウェブサイトについては年 1 回，電子雑誌については年 4 回である．

Web 情報は日々刻々と変化していくもので，その変化は Web ページ毎に異なる．この変化に WARP は対応できないため，収集できない Web ページ情報がある．そこで，本研究では更新・変更された Web ページに関してリアルタイムで差分収集を行うサーバを構築する．

差分収集を行う理由として，以下の点が挙げられる．まず，WARP の一括収集では保存できない Web ページの更新内容を残すためである．また，収集する度に全ての Web ページを一括収集していてはデータがパンクしてしまうためである．

Web ページを差分収集するにあたり，全てのページを 100%収集することは実質不可能である．そのためには更新する側が何らかの情報を収集する側に提示しなければいけない．また，毎分おきに Web ページを差分収集していては，そのページの管理者に迷惑がかかり，収集が弾かれる恐れがある．そこで，可能な限り少ない収集間隔で多くの情報を得なければならない．そのため，差分収集サーバ構築の際，Web ページが更新される間隔を推定することが必要である．ここで，世界中の Web ページを収集するとその更新間隔は正規分布に近付くことが容易に考えられる．しかし，Web ページの管理者の性質によって個別の Web ページ更新間隔の分布型は異なる．例えば平日の 9 時から 17 時まで，土日に集中，深夜のみ，毎月 1 日のみなどの更新間隔が挙げられる．そこで Web ページ毎に分布を推定する必要がある．その分布は正規分布とは限らず，例えば毎週月曜日と金曜日に更新されればその分布は二項分布になると考えられる．

そこで，ブートストラップ法を用いて更新間隔の推定を行う．ただし，各 Web ページは，予め仮定した分布型に基づいて更新されるとする．

2 章ではブートストラップ法について説明する．3 章では提案手法，4 章で実証分析について述べる．そして最後に，5 章でまとめる．

2 ブートストラップ法

ブートストラップ法とは，統計的推論の手法であり Efron[9] によって提唱されている．この手法はリサンプリング法に分類される．リサンプリング法とは，既に得られたデータからサンプリングを行い推定値のばらつきを評価するパラメータ推定の一般的な手法であり，確率シミュレーションの一種である．大量の計算を必要とするため，計算機の発展と共に実用化されてきている．ブートストラップ法を用いることで，実験データに内在するばらつきの影響により導かれる誤差を防ぐことが可能になる．

ブートストラップ法には，パラメトリック・ブートストラップ法とノンパラメトリック・ブートストラップ法の 2 種類が存在する．

パラメトリック・ブートストラップ法とは，母集団の変動が，ある確率分布に基づいていると仮定する推定法である．平均値の差の検定や分散分析などがこれに該当する．この手法は，標本サイズが小さい場合には分布型が不正確なことが多い．

ノンパラメトリック・ブートストラップ法とは，母集団の分布型に一切の仮定を設けない推定法である．パラメトリック・ブートストラップ法に対し，この手法は常に適用可能である．しかし，標本サイズが大きい場合には計算が煩雑になり正確なデータが得られず，評価しにくい面がある．

本稿では，各 Web ページが更新される間隔を仮定して推定を行う．1 つ目は各 Web ページが一定間隔で更新される場合について，2 つ目は任意の曲線に基づいて更新を行う場合について推定を行う．この場合，Web ページの分布型が仮定できる．そのため，パラメトリックな手法を用いた推定法を提案する．

3 更新間隔の推定法の提案

本章では Web ページの更新間隔をブートストラップ法を用いて推定する方法について説明する．各 Web ページの更新間隔の分布型を正規分布と任意の曲線の 2 つに仮定し推定を行う．

正規分布の密度関数 $f(x)$ は式 (1) で表される．

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{1}$$

ここで，任意の正の整数 n のデータ $x_1, x_2, \ldots, x_n$ において，μ, σ は以下のように定められる．

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{2}$$

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{n}} \tag{3}$$

これを $N(\mu, \sigma^2)$ とすると，正規分布 $N(\mu, \sigma^2)$ の平均 $E(X)$ と分散 $Var(X)$ はそれぞれ

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

となる．また，基準の正規分布の平均値 $E(X)$，分散 $Var(X)$ とブートストラップ法を用いた分析結果の平均値 $E(X^*)$，分散 $Var(X^*)$ の誤差 $SE\_E(X^*)$，$SE\_Var(X^*)$ は以下のように定められる．

$$SE\_E(X^*) = |E(X) - E(X^*)| \tag{4}$$

$$SE\_Var(X^*) = |Var(X) - Var(X^*)| \tag{5}$$

これらの式を用いて，本分析で使用するアルゴリズムは以下のようになる．

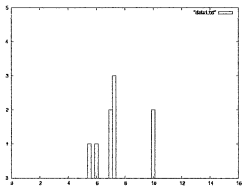

図 1: k=10

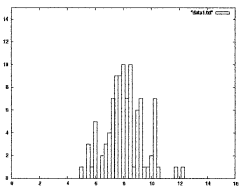

図 2: k=100

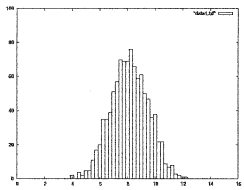

図 3: k=1000

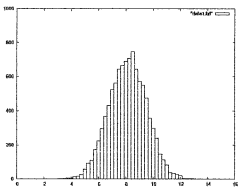

図 4: k=10000

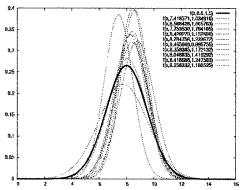

図 5: k=10

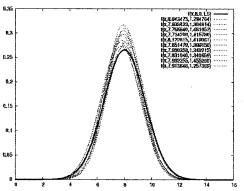

図 6: k=100

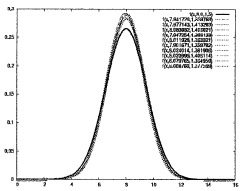

図 7: k=1000

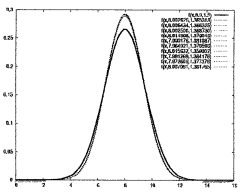

図 8: k=10000

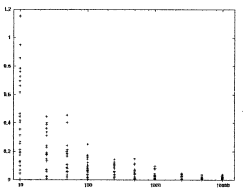

図 9: 平均値の誤差

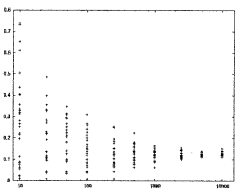

図 10: 分散値の誤差

1. 仮定した分布型から無作為に n 個のデータを作成し，これを母集団 $(x_1, x_2, \ldots, x_n)$ と設定
2. 母集団 $(x_1, x_2, \ldots, x_n)$ からリサンプリングにより，サイズ $m(m < n)$ の標本 $(x_1^*, x_2^*, \ldots, x_m^*)$ を k 組作成
3. k 組の標本からそれぞれ μ, σ^2 を算出
4. 3. で求めた値と，仮定した分布型の値との誤差を式 (4) と式 (5) を用いて分析し，統計的に更新間隔を推定

4 分析

本分析では更新間隔の分布型を仮定した Web ページを対象とする．まず，一定間隔で更新される Web ページの推定を行う．この場合，分布型は正規分布となる．

各変数は，以下のように設定する．

$$E(X) = 8.0, \quad Var(X) = 2.25$$

$$n = 10000, \quad m = 10$$

分析において，母集団の値は式 (1) から算出する．リサンプリング回数である k の値を $k = 10$, $k = 100$, $k = 1000$, $k = 10000$ と変化させる．また，無作為にリサンプリングを行うため，各 k の値に対して 10 回ずつ実験を行う．

次に，Web ページの更新間隔の分布が任意の曲線に基づくと仮定し推定を行う．各変数は，以下のように設定する．

$$n = 800, \quad k = 10000$$

分析において，標本数である m の値を $m = 5$, $m = 10$ と変化させる．また，任意の曲線の母平均は 4 点あり，$\mu 1 = 87, \mu 2 = 347, \mu 3 = 568, \mu 4 = 758$ と設定する．

図 1 から図 4 は標本平均値 μ を算出し，ヒストグラムにした結果である．横軸は標本平均値 μ の値，縦軸は同じ μ の値を持つ Web ページの個数を表している．図 1，図 2 より，標本数が少ないと大幅な誤差を含む推定となっている．一方，標本数の多い図 3 と図 4 の場合，正規分布に従っていることがわかる．ゆえに，標本数を増やすことによって，ヒストグラムは滑らかな曲線に近付くことがわかる．

図 5 から図 8 は最初に仮定した正規分布のグラフとの比較である．太線は最初に仮定した基準となる正規分布のグラフ，薄い線はブートストラップ法を用いたグラフである．それぞれの標本おいて μ, σ^2 を算出し，正規分布の式 (1) に代入する．k の値が大きくなるにつれて基準の正規分布のグラフに似た曲線を描くことが確認できる．

図 9 と図 10 は平均 $E(X)$ と分散 $Var(X)$ の値の具体的な誤差を式 (4) と式 (5) を用いて算出し，モデル化したグラフである．縦軸は誤差の値，横軸はリサンプリング回数 k を表している．このグラフからも，k の値が大きくなるにつれて基準の正規分布との誤差が小さくなることが確認できる．

図 11 と図 12 は，ある Web ページの更新間隔の分布が任意の曲線に基づくと仮定した場合を，ヒストグラ

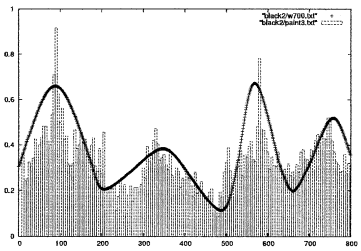

図 11: m=5, k=10000

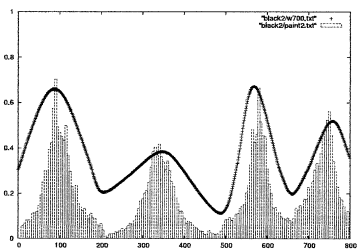

図 12: m=10, k=10000

ムにした結果である．図の棒グラフは実験値を階級区分した密度であり，曲線はその理論分布である任意の曲線を表す．標本平均値は母平均 $\mu 1$ から $\mu 4$ を中心とする範囲に集中しており，標本数 m が大きくなると母平均への集中が強まることが確認できる．また，この集中の程度を分布の広がりの尺度である標準偏差 $\frac{1}{\sqrt{m}}$ で比較すると，標本数 $m=5$ の時は $\frac{1}{\sqrt{5}}$，$m=10$ の時は $\frac{1}{\sqrt{10}}$ となり，標本数が大きいほど，確率変動の幅は縮小される．よって，母平均への集中度合いが強まることがわかる．

分析の目的は，ある分布型に基づいて更新を行うWebページの更新間隔がその分布型に従うと仮定した場合，その分布型から無作為に作成されるデータを用いて，元の分布型に近似することができることを確認することである．図 9 と図 10 より，標本数の増加によって，明らかに誤差が小さくなることが確認できる．また，図 1 ～8 より，リサンプリング回数 k を大きくすればするほど，正規分布に従うことがわかる．そして，図 11 と図 12 より，標本数が大きくなると元の曲線における母平均に集中し，仮定した分布型の平均値との誤差が小さくなることが確認できる．よって，Webページの更新頻度を推定する場合において，予め仮定した分布との誤差は標本数とリサンプリング数を大きくすることで更新間隔をある程度推定できるということがわかる．

5 まとめ

本稿では，ブートストラップ法を用いたWebページの更新間隔を推定する手法を開発し，その評価を行った．分析結果より，任意に仮定した分布型で更新されるWebページに対して，その更新間隔の分布型から抽出したデータにブートストラップ法を適用した場合，基準の分布型との誤差は，標本数とリサンプリング回数に大きく依存することが確認され，その回数を大きくすることで基準の分布型と近似することができることがわかる．

本稿では，任意のWebページを対象に，その更新間隔はある分布型に従うと仮定して推定を行ったが，World Wide Web 上のあらゆる Web ページは様々な間隔で更新が行われる．そこで，今後あらゆる分布型で仮定して Web ページの更新間隔を推定する．また現在，実際にある Web ページを Heritrix というクローラを用いて差分収集している．その結果をブートストラップ法を用いて分析し，更新間隔の推定を行う．

参考文献

[1] Internet Archive:http://www.archive.org/index.php

[2] Library of Congress (USA),MINERVA: http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[3] Swedish National Library, Kulturarw3: http://www.kb.se/english/

[4] The British Library, UK Web Archiving Consortium: http://www.webarchive.org.uk/

[5] National Library of Austria, Institute for Softwaretechnology and Interactive Systems at the Technical University of Austria, AOLA:http://www.ifs.tuwien.ac.at/ aola/

[6] National library of the Netherlands, e-Depot: http://www.kb.nl/dnp/e-depot/e-depoten.html

[7] National Library of Australia, PANDORA: http://pandora.nla.gov.au/

[8] 国立国会図書館　インターネット情報選択的蓄積事業　WARP:http://warp.ndl.go.jp/

[9] Efron, B:Bootstrap methods: another look at the jacknife, Annalsf Statistics, 7, pp.1-26(1979).