3G7-4

ブログサイトのクローリング戦略の最適化に関する分析

A Study of Optimization of Crawling Strategies for Weblog Sites

副島啓一[*1]
Keiichi Soejima

福原知宏[*2]
Tomohiro Fukuhara

[*1] 東京大学工学部システム創成学科
Department of System Innovation, Faculty of Engineering, The University of Tokyo

[*2] 東京大学人工物工学研究センター
RACE (Research into Artifacts, Center for Engineering) The University of Tokyo

大向一輝[*3]
Ikki Ohmukai

武田英明[*2*4]
Hideaki Takeda

[*3] 国立情報学研究所コンテンツ科学研究系
Digital Content and Media Sciences Research Division, National Institute of Informatics

[*4] 国立情報学研究所実証研究センター
Principles of Informatics Research Division, National Institute of Informatics

We report analysis results of crawling strategies of Weblog sites. Design policy of Weblog crawlers is different from traditional Web crawlers because Weblog crawlers should (1) check RSS or Atom feeds that contain recent articles of a Weblog site periodically, and (2) consider loads of Weblog servers for avoiding unnecessary accesses. We describe several analysis results obtained from crawling simulations using postings data of 10,000 blog sites.

1. はじめに

今日, インターネット上には膨大な数のブログサイトが存在する. ブログサイトを対象とした網羅的な記事検索システムを構築するには, (1)クロール先サーバの負荷を考慮しつつ, (2)各ブログサイトからの記事の取りこぼしを少なくするクローリング戦略が必要である. 本研究では実際のブログサイトにおける投稿データと計算機シミュレーションを用いて行ったクローリング戦略の最適化に関する実験結果について述べる.

本論文の構成は次の通りである. 2.ではブログサイトを対象としたクローリングの問題について述べる. 3.では実験方法について述べ, 4.では実験結果について述べる. 5.では関連研究と今後の課題について述べ, 6.で本論文の議論をまとめる.

2. ブログサイトを対象としたクローリングの問題

ブログサイトからのクローリングでは次の 2 つの問題が存在する:(1)記事収集漏れの問題, (2)サーバへの負荷の問題である.

2.1 記事収集漏れの問題

ブログサイトでは通常, 新着記事を RSS(RDF Site Summary) や Atom などの新着記事配信フォーマット(フィード)として提供している. フィード中に含まれる記事数は 5 件から 20 件程度に限定されており, フィードを定期的に確認するクローラにとっては, 新着記事がフィードから消えることは記事の取りこぼしにつながる. このため, クローラは新着記事がフィード内に存在している間にフィードを確認する必要がある.

2.2 サーバに対する負荷の問題

新着記事が存在しないにも係らずクローラがサーバにアクセスすることでサーバの負荷を増大させる問題がある. このため, クローラはサーバに対して必要最小限のアクセスを行う必要がある.

連絡先:福原知宏, 東京大学人工物工学研究センター価値創成イニシアティブ(住友商事)寄付研究部門, 千葉県柏市柏の葉 5-1-5, E-mail: fukuhara@race.u-tokyo.ac.jp

3. 実験方法

(1)実験データ, (2)クローリングにおける評価尺度, (3)シミュレーションにおける変数について述べる.

3.1 実験データについて

分析対象となるデータには筆者らが収集したブログサイトの記事投稿データの内, ランダムサンプリングした 1 万サイトの記事投稿データを用いた. データの期間は 2006 年 1 月 1 日から 2006 年 7 月 19 日までの 200 日間であり, 1 万サイトに含まれる総記事数は 223,569 件(1 日あたり 1117.85 件), 1 サイトあたりの平均記事数は 22.36 記事(最大 1081 記事, 最小 0 記事)であった. この 1 万サイトの内, 58.6%のサイト(5,858 サイト)が期間中に 1 件以上の記事を発信していた.

Fig. 1に期間中の累積記事数に対するブログサイト数を示す. 記事数が 0 の更新されていないサイトが含まれている.

Fig. 2に期間中の記事数の推移を示す. X 軸は 2006 年 1 月 1 日からの経過日数を示し, Y 軸はそれぞれの 1 日における観測された記事数である. 図中, 200 日付近で記事数が下がっているのは, 全てのサイトをクローリングするまでに 10 日前後の日数を要するためである.

3.2 評価尺度について

クローリングの評価尺度には次の 2 つの尺度を用いた.

(1) **損失記事数**(*Lost Article; LA*)
クローリングにおいて記事を取りこぼした件数である.

(2) **重複アクセス数**(*Failure Access; FA*)
クローリングにおいて新着記事が存在しないにもかかわらずサーバにアクセスした回数である.

LA はクローラにとっての問題であり, FA はサーバにとっての問題である.

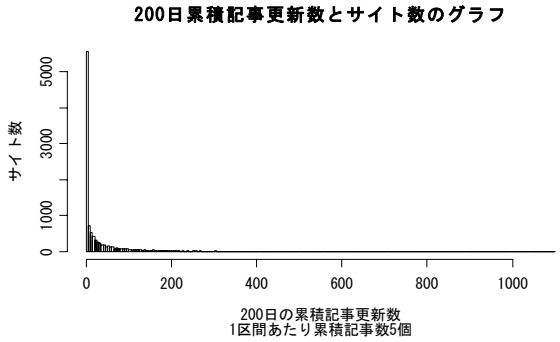

Fig. 1　期間中の累積記事数の分布

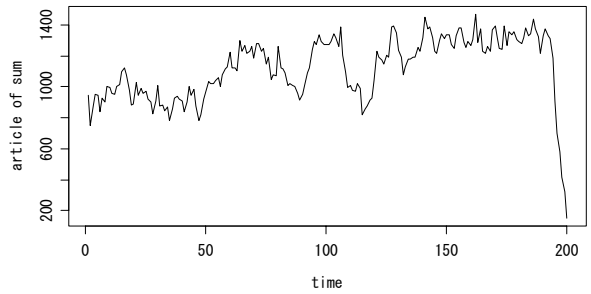

Fig. 2　期間中の記事数の推移（2006/1/1 – 2006/7/19）

本研究では LA と FA がパレート最適となるパレートフロントを見つけることを目標とする. なお本研究では 1 サイト 1 サーバと考える[1].

3.3 シミュレーションにおける変数について

シミュレーションにおける変数として以下を用いる.

(1) **新着記事表示件数**:各ブログサイトが保持する新着記事表示件数である. 通常, RSS/Atom フィードは 5 件から 20 件程度の新着記事を提供しており, それより過去の記事はフィード中には表示されない.

(2) **アクセス間隔**:クローラがブログサイトにアクセスする間隔(単位:日)である. なおクローラは 1 日に 1 回のアクセスを上限とし, 1 日に複数回のアクセスは行わないものとする.

(3) **クローラ台数**:クローラの台数である.

(4) **クローラセット**:クローラ台数とアクセス間隔の組である. 例:クローラセット(1,1,2)は 1 日に 1 回, 受け持ちのブログサイトにアクセスするクローラが 2 台, 2 日に 1 回アクセスするクローラが 1 台あることを表す.

4. 実験結果

実験結果について述べる. クローリングに当たっては 1 日にアクセスできるサイト数に上限を設ける場合(制約あり)と上限を設けない場合(制約なし)の場合を用意した.

以下, (1)単独クローラ(制約なし), (2)複数クローラ(制約なし), (3)複数クローラ(制約あり), (4)新着記事表示件数と LA の関係, (5)複数クローラ(制約あり, 動的グループ更新)の各条件での結果について述べる.

[1] 実際のブログサービス企業では 1 台のサーバで複数のブログサイトを運用しているが, ここでは単純化して 1 サイト 1 サーバと考える.

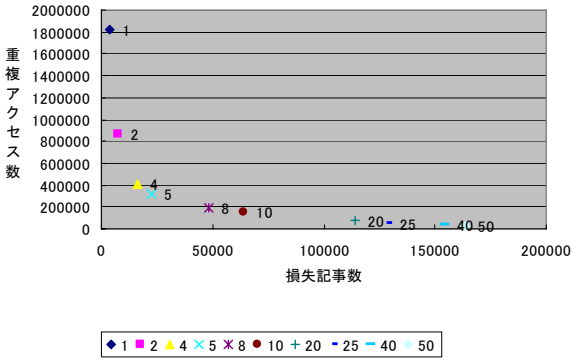

Fig. 3　1 クローラの場合での LA と FA（新着記事表示件数が 5 件の場合）

Table 1　各クローラへの担当サイト配分（N=2）

	クローラ I	クローラ II
サイト数等分(a)	5,000	5,000
最大記事数(b)	2	9,998
累積記事数(c)	830	9,170

Table 2　各クローラへのサイト配分（N=3）

	クローラ I	クローラ II	クローラ III	クローラ IV
a	2,500	2,500	2,500	2,500
b	2	0	10	9,988
c	309	521	909	8,261

Table 3　各クローラへのサイト配分（N=5）

	I	II	III	IV	V
a	2,000	2,000	2,000	2,000	2,000
b	1	1	1	33	9,964
c	231	361	532	899	7,977

4.1 単独クローラ(制約なし)

ここではベースラインとして, 1 台のクローラでクローリングする場合を考える. d 日(d=1, 2, 4, 5, …, 50)に 1 度, 全ブログサイト(10,000 サイト)にアクセスし, 新着記事を収集することを想定する. Fig. 3に新着記事表示件数 5(一律)の場合における結果を示す. X 軸は LA, Y 軸は FA である. アクセス間隔が 4 日および 5 日あたりで効率良くクロールできることが分かる.

4.2 複数クローラ(制約なし)

次に 2 台以上のクローラでブログサイトにアクセスした場合の LA と FA の変化を調べる. ここでの仮説は, クローラ数を増やせば 1 台の場合に比べて LA を改善できるという予測である.

ここでは複数のクローラ(クローラ台数 N=2, 4, 5)に対して次の 3 通りのサイト割り当て方法を用いた.

(a)サイト数等分法
　各クローラにサイト数を等しく分ける条件である. 累積記事数の少ないサイトから順に割り当てる.

(b)最大記事数等分法
　実験データの中で最も記事数の多かったサイトの記事数(最大記事数)をクローラ台数で割り, 各サイトの期間中の記事数(例えば総記事数 0 から 200, 201 から 400 など)に応じて割り振る条件である.

(c)累積記事数等分法
　グループ内の累積記事数が等しくなるよう分割した条件である.

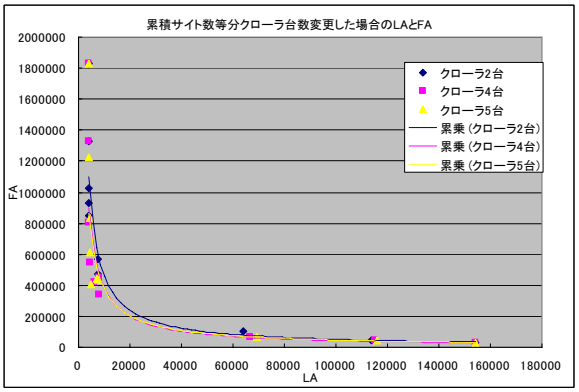

Fig. 4　サイト数等分条件(a)におけるパレートフロント

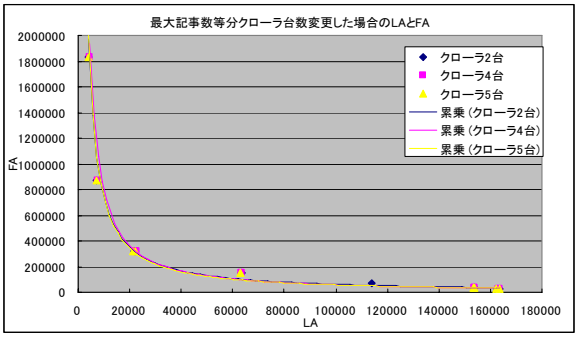

Fig. 5　最大記事数等分条件(b)におけるパレートフロント

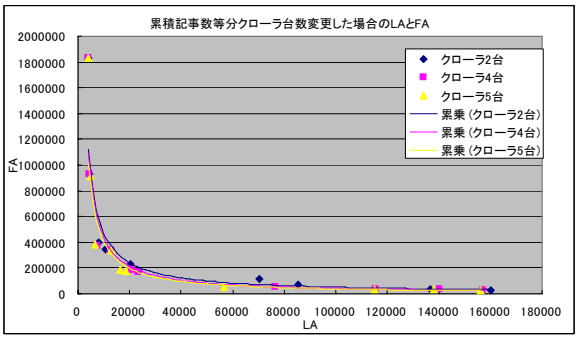

Fig. 6　累積記事数等分条件(c)におけるパレートフロント

それぞれの割り当て方法における各クローラの受け持ちサイト数をTable 1からTable 3に示す.

Fig. 4からFig. 6に各割り当て方法におけるパレートフロント(近似曲線)を示す. なお, ブログサイトの新着記事表示件数は一律 5 とした. 条件(a)におけるパレートフロントをFig. 4に示す. $N=5$ のとき, 3 条件の中で最も原点に近付いていることが分かる. このことから収集効率の改善にはクローラ数を増やせば良いことが分かる. 同様に条件(b)におけるパレートフロントをFig. 5に, 条件(c)におけるパレートフロントをFig. 6に示す. 3 条件の中では条件(c)が最も原点に近付いていることが分かる. またいずれの条件においてもクローラ数が多いと原点に近付くことが分かる.

4.3 複数クローラ(制約あり)

前節まではクローリングを行う際の制約を無視して実験を行った. しかし現実にはサーバへの過剰なアクセスは許されないことから, クロール回数は制限される必要がある. ここではクローリングを行う際に制約を導入した場合の戦略について述べる.

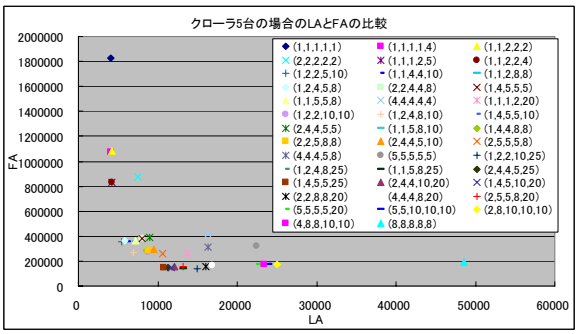

Fig. 7　複数のクローラセットに対する LA と FA (新着記事表示数=5 のとき)

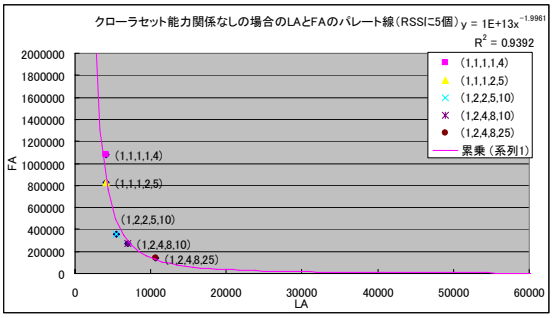

Fig. 8　パレートフロントに含まれるクローラセット

ここでは異なるアクセス間隔を持つ複数のクローラを組み合わせてクローラの組(クローラセット)でクローリングした場合のパレートフロントについて調査する. クローラセットを作る際の制約は次の通りである.

- クローラ数は 5 台で考える.
- アクセス間隔は 200 を割り切る数から選択する. 選択可能なアクセス間隔は 1, 2, 4, 5, 8, 10, 20, 25, 40, 50, 100(日)である.
- あるクローラセットにおいて各クローラが 1 日あたりに処理するブログサイト数は一定とする.
- 各クローラのアクセス間隔はそれぞれ異なるものとする. 例えば 1 日に 1,000 件のサイトを処理できるクローラがある場合, (1)1 回のアクセスで 5,000 件のフィードを取得し, 5 日かけて処理するクローラと, (2)1 回のアクセスで 1,000 件のフィードを取得し, 1 日で処理するクローラは, 1 日あたりに処理するサイト数は同じであるが, サイトにアクセスするタイミングは異なる.
- 処理に要する日数と 1 日あたりに処理されるブログサイト数の積は定数(10,000)とする. 例えば 1 日処理のクローラが 1, 2 日処理のクローラが 2 ある場合, 処理日数は 1+2+2=5 と数え, 5 処理日で 10,000 サイトを処理するものと考え, 1 日あたり 2,000 サイトを処理するようクローラセットを用意する.
- クローラセットの台数は $N=5$ とする. ここでは 39 通りのクローラセットを用意し調査した.

Fig. 7にクローラセット能力の高い順における LA と FA の散布図を示す. 例えば, 全てのクローラが毎日受け持ちのブログサイトにアクセスするクローラセット(1,1,1,1,1)では, 損失記事数は少ないがサーバへの負荷が極端に大きくなることが分かる.

Fig. 8にFig. 7の中でパレートフロントに含まれるクローラセットを取り出した図を示す. 原点に近いクローラセットを見ると(1,2,2,5,10), (1,2,4,8,10), (1,2,4,8,25)という組である. 一方, 残りの 2 組(1,1,1,1,4), (1,1,1,2,5)はアクセス間隔が 1 のクローラが多い.

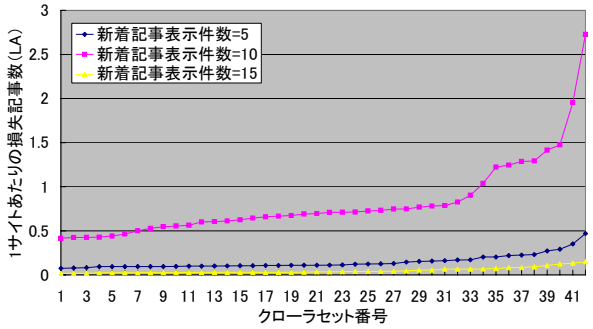

Fig. 9　新着記事保持数と 1 サイトあたりの損失記事数

以上から LA と FA を改善するクローラセットはアクセス間隔に適度なばらつきを持たせると良いと言える.

4.4 新着記事表示件数と損失記事数の関係

各ブログサイトの新着記事表示件数とクローラの損失記事数の関係について考える. クローラ側の観点からは LA を少なく抑えることが望まれる. ここでは新着記事表示件数がどれだけあれば LA を抑えられるかについて調査した.

Fig. 9に結果を示す. 新着記事表示件数は 5 件, 10 件, 15 件の 3 種類であるが, 表示件数が多いほど 1 サイトあたりの LA は少なくなることが分かる. このことから多くの記事を検索対象としたいブログサイトは新着記事表示件数を多くすれば良いと言える. 一方, クローラ側の観点からは FA を少なくするようアクセス間隔を大きくとって記事収集する必要がある. どの程度アクセス間隔を取るかについては今後調査を行う必要がある.

4.5 複数クローラ(制約あり)動的戦略

複数のクローラで制約を設け, かつ動的に受け持ちサイトを更新(グループ更新)することで収集効率の改善を行う場合を考える. ここではブログサイトを各クローラ(グループ)に割り振り, 一定期間ごとにグループ内のサイトを入れ替えるものとする. ここでは 200 日の期間のうち, 前半の 40 日を初期グループ作成用データとして用い, 後半 160 日のデータで評価する.

クローラセットには A(1,1,1,2,5), B(1,1,2,2,4), C(1,4,5,10,20), D(1,2,2,5,10), E(1,1,4,4,10)の 5 セットを用いた.

各クローラへのサイトの割り当て方法は, グループ更新時点から過去 n 日間の各サイトの累積記事数を求め, 記事数の多いサイトから順にアクセス間隔の短いクローラに割り振ることとした.

Fig. 10に動的戦略における更新間隔とクローラセット間の比較を示す. 図中, **Baseline** とは前半 40 日のデータを用いてグループ分けをし, 後半 160 日のデータを用いて 1 サイトあたりの LA を評価した結果である. **暫定正解**とは, 後半 160 日のデータを用いてグループ分けを行い, 評価した結果である. 可能性として他にも LA の少ない条件は存在し得るが, 暫定的な正解データとした. Baseline と暫定正解ではグループ更新は行っていない.

図中, 20 日, 40 日, 80 日ごとにグループ更新を行った結果を示しているが, 全てのクローラセットにおいて 20 日ごとにグループ更新するのが暫定正解に近付くことが分かった. この場合, 暫定正解への到達率(最善値／暫定正解値)は A:134%, B:120%, C:157%, D:177%, E:162%であり, クローラセット B において 20 日おきにグループ更新した場合, 最も暫定正解に近付くことが分かった.

5. 関連研究と今後の課題

クローリングに関する関連研究と今後の課題について述べる.

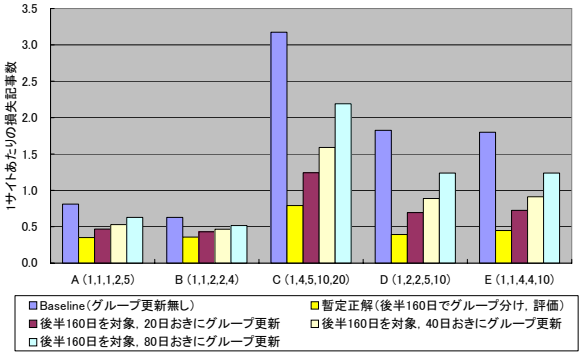

Fig. 10 動的戦略における更新日数・クローラセット間比較

5.1 関連研究

山名らは Web ページを対象とした分散型 WWW クローリングに関する実証研究を行っている[山名 2000]. 本研究ではブログサイトを対象としたシミュレーションを行ったが, 今後シミュレーション結果に基づく実装評価についても検討する.

Cheung らは RSS リーダを想定したクローリング手法を提案している[Cheung2007]. Cheung らが個人向け RSS リーダにおけるクローリングを想定しているのに対し, 本研究ではブログ記事検索エンジンを想定した組織的クローリングを想定している. 組織的クローリングではサーバに対する負荷も考慮せねばならず, LA と FA を考慮したパレート最適のクローリング戦略を検討する必要がある.

5.2 今後の課題

- 他のデータセットでの評価

本研究ではランダムサンプリングした 1 万件のデータを用いたが, 他のブログサイトのデータに対しても本研究の結果が成立するか確認する.

- 非定常的な記事数変動への適応

本研究では 1 日に投稿される記事数について定常性を前提としたが, ブログ空間では重大な事件や事故に対して突発的に記事数が増える場合がある. こうした非定常的な記事数の増減に追随できるクローリング戦略・手法の開発が必要である.

- 他の最適化・適応手法の検討

線型計画法やニューラルネットワーク, 遺伝的アルゴリズム, そのほかの機械学習を用いた最適化や適応について検討する.

6. まとめ

本研究ではブログサイトを対象としたクローリング戦略最適化のための基礎的実験を行った. ブログサイトを対象としたクローリングでは損失記事数と重複アクセスの双方を満足させる戦略が必要である. 本実験の結果, 複数のクローラを用いたクローリング戦略に関する基礎的データを得た. 今後の研究を通じて, クローラ設計者が損失記事数と重複アクセスのトレードオフを検討する際の指針となるクローリング品質保証(Quality of Crawling; QoC)の基礎データの蓄積を続けたい.

参考文献

[山名 2000] 山名早人, 分散型 WWW ロボット実験:Web ページ数増加との戦い, Bit, 共立出版, 2000.

[Chung2007] Ka Cheung Sia, Junghoo Cho, Koji Hino, Yun Chi, Shenghuo Zhu and Belle Tseng Monitoring RSS Feeds Based on User Browsing Pattern, Int'l Conf on Weblogs and Social Media (ICWSM), pp.161-168, 2007.