2N5-OS-21b-6

大規模 Web クローラの運用事例にみる課題と対策

ハイパフォーマンス・エラスティック・クローリング（第二報）

Challenges and solutions drew from operational cases of a large-scale web crawler

藤井 秀明[*1] Hideaki Fujii　原口 弘志[*1] Hiroshi Haraguchi　田中 康司[*1] Kouji Tanaka　泥谷 誠[*1] Makoto Hijiya　岩瀬 高博[*2] Takahiro Iwase　岩爪 道昭[*1] Michiaki Iwazume

[*1] 独立行政法人 情報通信研究機構 National Institute of Information and Communications Technology

[*2] 株式会社神戸デジタル・ラボ Kobe Digital Labo., Inc.

We have developed and operated a web crawler to construct a web archive of billion pages. Collecting huge data, we often encounter not a few troubles or failures we didn't expected in planning phase. In such a case, it is important to form a feedback loop in which we take advantage of data from operating result.

1. はじめに

独立行政法人 情報通信研究機構（以下，「NICT」とする）では，数十億件規模の Web アーカイブ構築を目指して，大規模 Web クローラを運用・開発している．このようなビッグデータ級のデータ収集においては，設計段階では想定していなかった事態や，データ量の少ない初期段階では顕在化していなかった障害が多々発生し得る．

しかしながら，すべてを事前に予測し対応することは非常に困難であり，可能であったとしても費用対効果の点で適切であるとは限らない．

そこで，運用に際して適切なログデータを収集・分析し，発生した問題に対応するとともに，得られたデータを次期研究開発に活かすフィードバックループを形成することが，ビッグデータ級のデータを扱うシステムにおいては重要となる．

本稿では，我々が実施している運用状況を紹介するとともに，これまでに発生した諸問題のうち，単位時間当たりのクロールページ数であるクロール速度の低下現象を事例として報告する．また，同事例を通して明らかとなった課題と対策について検討し紹介する．

2. クローラの運用状況

現在，NICT で運用している大規模 Web クローラ（以下，「NICT クローラ」とする）の運用状況について以下に示す．

2.1 システム・マシン構成

システム概要については拙著[藤井 2012]にて紹介済みであるため，本稿においては変更のあったマシン構成について表 1 に，またマシンスペックを表 2 に示す．

表 1 マシン構成

種別	用途	ノード数	プロセス数
新規・更新	制御・DB 登録	1	1
	ページ収集	4	4
	ページ解析	4	16

連絡先：藤井秀明，独立行政法人情報通信研究機構，〒619−0289 京都府相楽郡精華町光台 3−5，h−fujii@nict.go.jp

表 2 マシンスペック

項目	内容
CPU	Intel(R) Xeon(R) X5570 2.93GHz (Nehalem 2CPUs/Node)
Memory	8GB
Local Storage	500GB SATA x2 (Hot Swappable/RAID-1)

2.2 クローリング関連データ量

クローリング関連のデータについて，その数量及びサイズを表 3 に示す．

表 3 クローリング関連データ数量・サイズ

項目	数量・サイズ
クロールページ総数	約 69 億
クロールページ総サイズ [1]	約 12TB
URL DB 登録 URL 件数	約 10 億

※2012 年 9 月 30 日時点の値

2.3 監視体制

NICT クローラの監視体制としては，表 4 に示す項目について実施している．

表 4 NICT クローラ監視項目

No.	項目
1	日別のクロールページ数
2	ダウンロードサイズのリアルタイム値
3	ストレージ利用率
4	クローラ実行マシンの稼働状況

以下，各項目の監視内容について概要を説明する．

[1] HTML ファイルに加え，クローリング処理に関連する各種ログファイル等も含む．

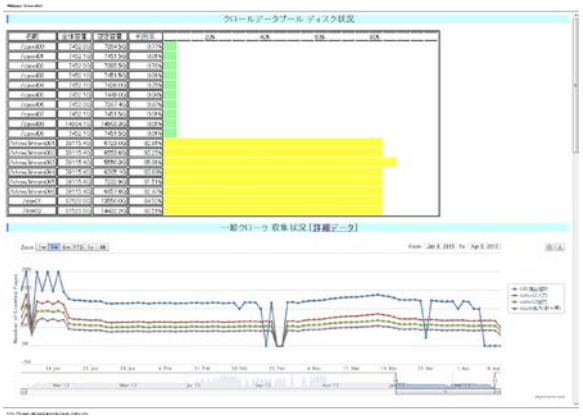

図 1 クローラ監視画面①

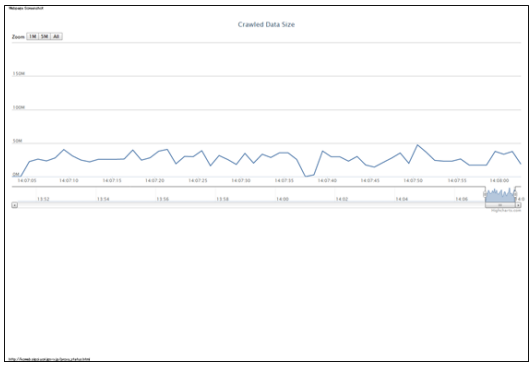

図 2 クローラ監視画面②

(1) 日別のクロールページ数

クロールされた Web ページの数を一日単位で集計し，表形式及び折れ線グラフで可視化するシステムを開発した．図 1 下に折れ線グラフの画面を示す．日別でのクロールページ数の変化を可視化し，極端にページ数の少ない／多い日など通常と異なる事態が発生していないか監視を行っている．

(2) ダウンロードサイズのリアルタイム値

ダウンロード中のデータ転送速度を，Mbps 単位でリアルタイムにグラフ表示するシステムを開発した．図 2 に画面を示す．クローリング処理の実行中／停止中の確認，クローラが接続されているネットワークに対して過度の負荷をかけていないかどうかの監視等に利用している．

(3) ストレージ利用率

クロールした Web ページ(の HTML ファイル)はファイルサーバ上に保存されるが，このストレージの利用率を表示するシステムを開発した．図 1 上に画面を示す．利用率はストレージを構成するノード単位に棒グラフにて表示され，また利用率に応じて色分けされるため，どのノードのストレージが容量不足かどうか認識しやすくなっている．

(4) クローラ実行マシンの稼働状況

クローラが稼働しているマシンの状態(CPU 稼働率，メモリ使用率，ネットワークトラフィック等)を監視する．画面を図 3 に示す．なお同項目の監視には，オープンソースのクラスタ監視ツールである Ganglia[Ganglia]及び Cacti[Cacti]を利用している．

図 3 クローラ監視画面③

3. クローラ運用に関する事例

NICT クローラを運用監視する過程で発生した問題のうち，クロール速度低下現象を事例として紹介する．

3.1 クロール速度低下の現象

NICT クローラはクローリングのアーキテクチャとして，バッチクローリング[Olston 2010]を採用している．バッチクローリングでは，重複を含まない URL のリストを用いてクロールを行い，すべての URL をダウンロードし終えることで 1 回のクロール処理が終了する．NICT クローラを含む一部のクローラは，Web ページの更新や新規に獲得した URL のクロールのために，定期的にこのクロール処理を繰り返す仕組みとなっている．

2.3 にて説明した監視項目「(2)ダウンロードサイズのリアルタイム値」より，この 1 回分のクロール処理において，クロールを開始してから時間が経過するとともに，クロール速度が徐々に低下する現象が発見された．これを受けて「(1)日別のクロールページ数」の集計前データをグラフ化し，1 サイクル(約 180 分間)のクロール速度の推移を示したものが図 4 である．クロール開始直後は約 10,000～15,000 ページ／分であったものが，60 分を過ぎた辺りから徐々に減速し，160 分以降は 100 ページ／分にまで低減している様が観測できる．

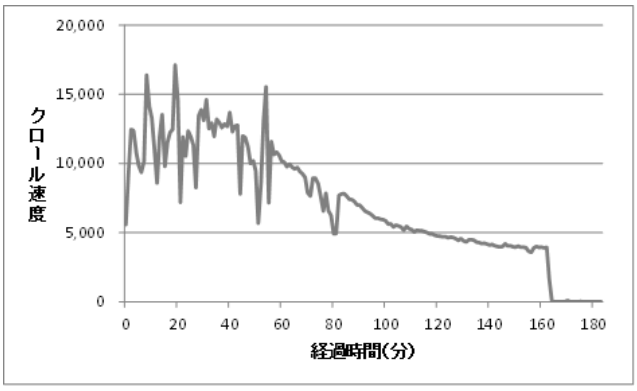

図 4 クロール速度の推移

3.2 パラメータの調整と結果

クロール速度低下に対する運用レベルの対策として，クロール速度が低下したら一旦クロールを中断し，新しい URL リストで再開する運用を試みた．

表 5 は，1 回のクロールにおける最長実行時間を設定した上で，1 週間クロールを継続した際の 1 日の平均クロールページ数を示している．

表 5 クロール時間とページ数

クロール時間	3 時間	2 時間	1 時間
平均ページ数	8,161,000	12,524,000	10,095,000

3 時間と 2 時間の値を比較すると，クロール時間を短縮し，低クロール速度の時間も短縮することで収集効率が向上する傾向が分かる．一方で，2 時間と 1 時間の値より，単にクロール時間を短縮すれば良いわけではなく，適切な時間設定が必要であることが分かる．これは，1 回のクロール時間を短縮する，すなわち 1 日あたりのクロール回数が増加すると，クロールを実行するための事前処理の回数も増えてしまうことが原因と推察される．

4. 今後の課題と対策

運用事例を踏まえて，今後の課題と対策を検討した．

4.1 フィードバックループに基づく次期研究開発

3.2 における対応は，運用レベルで対応可能な範囲のものであるため，システム改修を視野に入れた根本的解決方法を検討する必要がある．

推測されるクロール速度低下の原因は，紳士的クロールの仕組みである．これは相手サーバへの負荷軽減のために，同一ホストへのアクセスごとに一定のインターバルを設ける仕組みである．インターバルの間には，別ホストのサーバにアクセスを行うことで無用なアイドル時間が発生しないよう制御しているが，クロールが進行するに伴って URL 件数の少ないホストは早々にクロールが完了してしまい，少数の URL 件数の多いホストが残ってしまうためクロール速度が徐々に低下するものと思われる．

したがって，ホスト数を一定数以上に維持するため，適宜新しい URL の追加が可能な仕組みを開発する必要がある．

4.2 監視環境の整備

運用データに基づき次の研究開発を進めるフィードバックループを効果的に形成するためには，何よりも適切な運用データを収集可能な監視環境が必要である．

現時点においても本稿で紹介した監視環境により，必要最低限の運用データは収集可能であるが，より高効率なデータ活用を実現するために，下記に示すような機能を持つ監視環境が必要であると考える．

- 複数のデータを関連付ける表示機能
  例えば，時刻情報を手掛かりに結び付けられた同一時点で発生した複数のデータや，通信中の 2 台のマシンのデータ等を一覧にて表示する．

- 現象発生時の状況再現機能
  障害等が発生した前後の時間帯における各運用データの変動をリプレイすることで，データ間の関連性を見つけ易くする．

- 監視と操作の統合環境機能
  運用状態の監視中に注意すべき現象が発生した場合等，同一画面上でシステムに対する操作を行い，そのフィードバッグを監視しつつ運用を続ける．

5. おわりに

本稿では，大規模クローラの運用に際して，我々が直面した課題とその対策を事例として紹介した．紹介事例以外にも，クロールした Web ページ保存先容量の逼迫に起因する不具合や URL DB の処理遅延など，対象データ数が増大化して初めて顕在化する課題は少なくない．今後も引き続き，運用データをより効果的なチューニングや機能向上の改修に役立て，その成果を用いて運用を継続していくというフィードバックループの形成を進めていきたい．

参考文献

[藤井 2012] 藤井他: ハイパフォーマンス・エラスティック・クローリング，2012 年度人工知能学会全国大会，1A1-OS-17a-2, 2012.

[Cacti] Cacti - The Complete RRDTool-based Graphing Solution, http://www.cacti.net/.

[Ganglia] Ganglia Monitoring System, http://ganglia.sourceforge.net/.

[Olston 2010] Christopher Olston and Marc Najork : Web Crawling, Foundations and Trends in Information Retrieval, Now Publishers Inc., 2010.