Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

Web ページの大規模収集・検索基盤の構築と運用

赤峯享[†]　加藤義清[†]　河原大輔[†]　新里圭司[‡]
乾健太郎[†,§]　黒橋禎夫[†,‡]　木俵豊[†]

本稿では，情報通信研究機構で構築し，運用している，大規模なWeb ページの収集・検索基盤について報告する．本収集・検索基盤は，(1)10 億ページの web ページを更新収集すること，(2)収集ページから選択した約 1 億ページの検索対象ページに形態素解析，同義表現解析，構文解析結果を付与したデータを作成し，上位の分析アプリケーションで利用可能にしていること，(3)選択した検索対象ページに対して検索エンジン基盤 TSUBAKI を用いた検索が常時可能なこと，(4)クラスタ計算機上に分散配置したインデックスやデータを利用することでスケーラブルな構成になっていることを特徴としている．

Development of a Large-scale Web Crawler and Search Engine Infrastructure

Susumu Akamine[†]，Yoshikiyo Kato[†]，Daisuke Kawahara[†]，
Keiji Shinzato[‡]，Kentaro Inui[†,§]，Sadao Kurohashi[†,‡]
and Yutaka Kidawara[†]

This paper reports on the ongoing development of a large-scale Web crawler and search engine infrastructure at National Institute of Information and Communications Technology. The system has several strong features: (1) It collects several hundred million Web pages while maintaining them to be up-to-date, (2) Among the collected pages, selected 100 million pages are converted into the standard data format to store all the results of morphological analysis, dependency parsing, and synonym augmentation. (3) The selected set of pages is regularly searchable and accessible to the users, and (4) The scalability of the system is achieved by distributed data processing over a large-scale cluster machine.

† 情報通信研究機構， National Institute of Information and Communications Technology
‡ 京都大学， Kyoto University
§ 奈良先端科学技術大学院大学， Nara Institute of Science and Technology

1. はじめに

Web の利用は日常生活に浸透し，インターネット上には，政府広報，ニュース，製品情報，製品に対する評価・評判情報，Q&A，日常体験を綴ったブログなど様々な大量の情報が流通している．これらの生きた大規模なテキスト情報は，単語の共起頻度の計数などの基本的な言語処理から，製品の評価・評判情報を利用したマーケティング支援／購入支援などの高度な情報分析まで，数多くの有益な利用が可能である．実際，Web を言語資源として利用した研究は盛んであり，コーパスとしての Web の利用に関するワークショップが継続して開催されている[1]．また，blog などの口コミ情報を対象とした評判情報検索は複数の試行サービスが行われている．逆の見方をすれば，コーパスを用いた言語処理技術の研究開発や，Web 情報を用いた情報分析技術の研究開発のためには，大規模な Web コンテンツが利用できることが必要不可欠なものとなっている．

情報通信研究機構(NICT)では，Web コンテンツの信頼性分析の研究開発プロジェクトを 2006 年度から推進しており[2,3]，その研究開発・評価のために大規模な Web コンテンツが必要となった．大規模な Web コンテンツを利用する場合，(a)Google や Yahoo のような商用の Web 検索エンジンを利用して，検索結果の上位ページをアクセスする，もしくは，(b)自力で Web ページを収集し，それを検索するための基盤を構築する，の何れかの方法が考えられる．(a)は，(b)と比べて開発・運用コストが大幅に少なくて済むが，一方で，商用検索エンジンの API は使用回数の制限があり，かつ，Web ページに直接アクセスできないため，Web ページに対して内容分析などの深い分析を行うのが困難である．したがって，筆者らは(b)を選択し，計算機基盤，Web ページ収集基盤を整備し，さらに検索エンジン基盤 TSUBAKI[4]と連結することで，自前で Web ページを収集・検索するための基盤を構築した．

本稿では，Web 情報信頼性分析システム WISDOM**で利用するために，NICT で構築し，運用している 1 億ページ規模の Web ページの収集・検索基盤について報告する．実際に収集・検索基盤を整備し、運用してみると、当初の予想以上の大きな労力を費やしており，情報分析を行う全ての組織が独立して，これらの基盤整備を一から行うことは非常に効率が悪いと実感している．本報告は，今後の組織を越えた連携協力や Web コーパスの共有のための参考という意味も込め，基盤システムの構築と実運用の情報を報告する

2. 開発・運用の方針

検索基盤を利用する上位の分析アプリケーションは，実ユーザの利用からフィード

** WISDOM は次の URL で試験公開を行っている．
http://wisdom-nict.jp/

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

バックを得て，改良を行うことが重要である．一般ユーザに分析アプリケーションを利用してもらうためには，最近話題になった出来事など様々な分析課題に対応でき，かつ，常時サービスが利用できることが必要である．一方，情報分析技術の研究開発者側の観点としても，特定の閉じたドメインだけでなく多くのトピックで評価することは重要である．

また，テキスト分析において，表記が異なるが意味が同じ表現を同一視したい，分析精度を高めるために単語間の係り受け関係を使いたいなどの要望がある．これらは個々の分析アプリケーションに依存しない共通のものであり，基盤の部分で吸収するのが効率的である．

したがって，筆者らは以下の方針で収集・検索基盤を構築した．

- 特定の話題だけでなく任意の話題を扱える規模の Web ページを，最近のものも含めて，検索対象とする．
- 検索対象の Web ページを追加する際にも，検索サービスを止めずに，常時利用できる．
- HTML ファイルにアクセスできるだけでなく, HTML ファイルからテキスト情報を抽出して，文に分割し，形態素，構文解析，同義表現解析を行い，その解析結果もアクセス可能にする．

日本語に限定したとしても，インターネット上には少なくとも数十億規模の Web ページが存在する．上記の方針を満たし，これらの全てのページを対象とするには，あまりに莫大な計算機資源が必要となり，現実的ではない．大規模収集・検索基盤の課題は，利用可能な計算機資源とのページ規模のバランスをとることである．特に，計算コストが高く，しかも，大量のディスクアクセスが発生する言語解析やインデックス作成部が問題となる．また，最初にページ規模を確定して，それに合わせて必要な計算機資源を準備することは困難である．したがって，更に以下を方針とした．

- ページ収集は検索対象の数倍から 10 倍程度の規模で行い，その中から質の高いページを選択して，言語解析，検索インデックスの作成し，検索対象とする．
- ページ規模の拡大を計算機の追加によって実現できるようにスケーラブルな構成にする．

3. システム構成の概要

現在, NICT の信頼性分析研究開発プロジェクトでは, 220 ノード(1 ノード当り 4cpu core，メモリ 8GB，ローカルディスク 1.5TB～2TB)のクラスタ計算機と 100TB のファイルサーバを用いて研究開発を実施している．

この計算機基盤上に構築した Web 情報分析システムの構成を図 1 に示す. 図 1 において，実線の枠で示された部分が Web ページの収集・検索基盤であり，点線の枠で示された部分は各分析アプリケーションに依存する処理である．

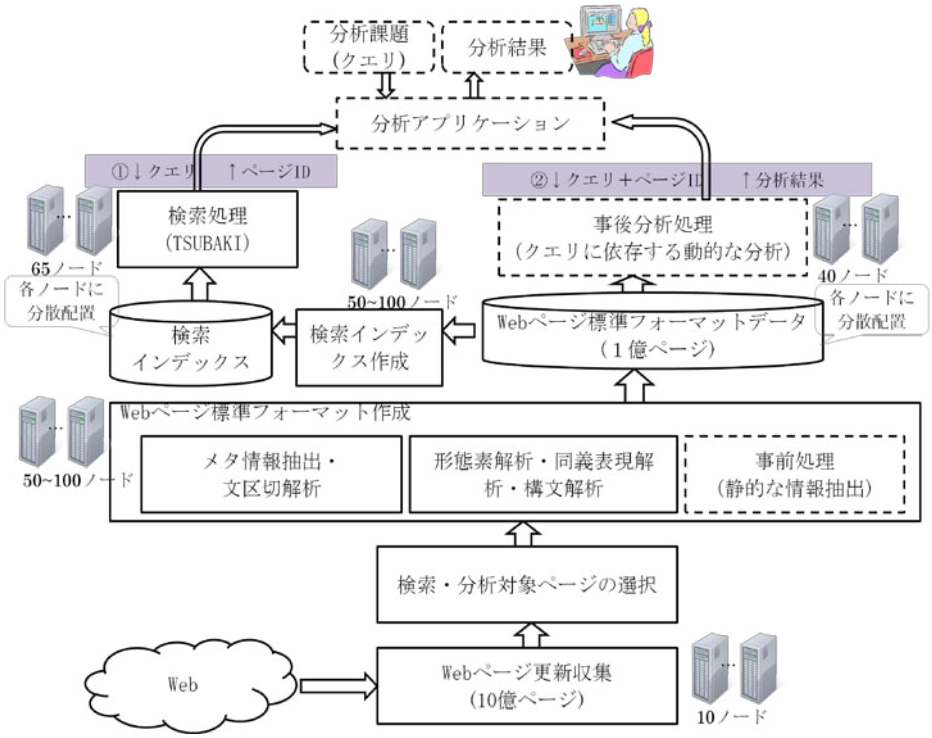

図 1 システム構成

収集・登録時は以下の手順で動作する．

1. Web ページ更新収集部はインターネット上の Web ページを収集し, HTML ファイルを出力する．また，収集した Web ページを一定間隔で更新する．10 ノードを利用．
2. 検索・分析対象ページの選択部は，収集した Web ページから，分析目的に合った質の高いページを検索・分析対象として，選択する．
3. Web ページ標準フォーマット作成部は，メタ情報抽出・文区切り解析，形態素解析・同義表現解析・構文解析を行い，解析結果を Web ページ標準フォーマッ

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

ト(5.1 節で詳細を説明)の形式で出力する．この際に，事前処理として，クエリに依存しない静的な情報抽出を行うことも可能である. 例えば, WISDOM では，情報発信者の抽出や広告や連絡先などの外観情報の抽出を行っている．50～100 ノードを利用.

4. 検索インデックス作成部は，Web ページ標準フォーマットから検索インデックスを作成する．50～100 ノードを利用.

また，検索／分析時は以下の手順で動作する.

1. 検索処理部は，分析課題のクエリが与えられると，インデックスを検索し，上位N件のページ ID の集合を返却する．65 ノードを利用.
2. 事後分析処理エンジンは，検索結果のページ ID から Web ページ標準フォーマットを取得し分析を行い，分析結果を返却する．例えば，WISDOM では，分析課題のクエリに依存する主要対立表現の分析，評価評判情報の分析などの分析処理を同時並列に行っている．50 ノードを利用.

上記のようなシステム構成で，ページ規模の拡大をノードの増強で対応できるようにするために，検索インデックス，及び，Web ページ標準フォーマットデータは全てローカルのディスクに数百万ページ単位で分散して配置している．登録処理や検索／分析処理の際に，これらのデータに対して，小さな単位でランダムに大量のディスクアクセスが発生する．そのため，ファイルサーバ上のデータに，通常の NFS を用いて利用すると，単一のノードからのアクセスでも極端にアクセス速度が低下する．さらに，複数ノードから同時アクセスが発生すると，致命的にアクセス速度が低下する．したがって，スケーラブルな構成にするためには，ページ規模の拡大に伴って，ファイルサーバへのアクセスが拡大しない，上記のような分散配置が必要である.

上記の計算機基盤で，実際に運用を行ってきた経験からすると，随時インデックスやデータの更新行いながら，常時検索／分析サービスを停止せずに運用し，かつ，各エンジンの独自性能評価などを同時に行う場合，この規模の計算機基盤を用いても，1 億ページ程度が限界であった．一方，収集に限っては 10 億ページ程度の収集は可能である.

4. Web ページの収集基盤

4.1 クローラの構築

前章までで述べたように，現状の計算機資源で分析可能な Web ページ数を約 1 億ページである．筆者らは，10 億程度の Web ページを収集し，その中から一定の品質を保った，できるだけ最新の 1 億ページを選択して検索・分析対象とするという方針の下で，以下の 3 種類のクローラを構築した.

- 更新クローラ
  メインのクローラで，最大 10 億ページの URL を管理し，その URL のページの更新・削除を定期的にチェックし，更新ページを収集する．また，更新ページから新規 URL を抽出し，収集対象に追加する.
- RSS クローラ
  RSS フィードの情報を元に，毎日，新規作成された blog 記事などを収集する.
- ニュースクローラ
  主要なニュースサイトに対して，トップページを起点として，毎日，新規ページを収集する.

RSS クローラとニュースクローラを用いることで, 一般ユーザが興味を持ちやすい，最近の話題を優先して収集できる．また，更新クローラによって，更新，削除された古いページを検索対象から排除することが可能になり，定期的に実際のインターネット上のページと検索対象の同期がとれた最新のページが収集できる.

4.2 更新クローラ

メインのクローラである更新クローラのシステム概要を図 2 に示す. URL DB に格納された 1 億～10 億ページの URL に対して，更新チェック用の URL を 200 万ページ単位で選択し，以下のリゾルバ，ページ収集，収集ページ解析を順に行うことで，更新／新規ページを収集し，URL DB を更新する． ページ収集の開始直後は，多くのサイトに並列収集可能でページ収集速度（単位時間当たりの収集ページ数）が速いが，時間が経過するにつれて，4.3 節に述べる同一サイトに対する収集間隔の制限により，ページ数の多い特定サイトのみが収集対象として残り，ページ収集速度が低下する．この収集速度の変化に対処するために，200 万ページ単位で収集を行い，一定時間が経過して収集速度が極端に低下した段階で収集を中止する.

1. リゾルバ
   - host 名から IP アドレスを非同期で取得する.
2. ページ収集
   - 4.3 節の運用方針に基づき Web ページを収集する.
3. 収集ページ解析
   - 収集ページの更新状況をチェックし，更新／削除の情報を URLDB に登録する.
   - 日本語のページのみを選択し，さらに，アダルトページなどの情報分析対象として不必要なページは簡易フィルターを用いて収集対象から排除する.
   - 収集ページから抽出した OutLink を新規 URL として，URLDB に登録する.

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

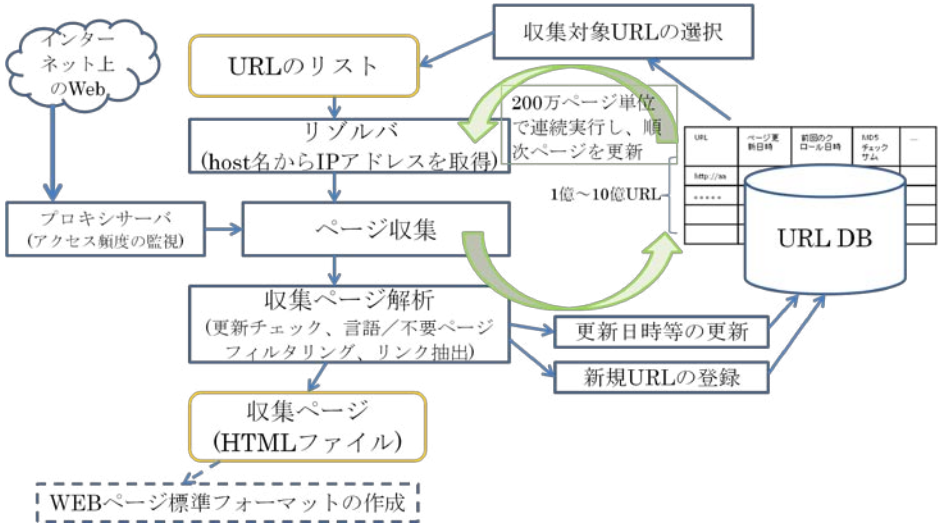

図 2 更新クローラの構成

大規模なページ集合を短期間でチェックし，最新のページを収集するためには，重要で頻繁に更新されるページについては更新チェック間隔を狭め，重要でなくあまり更新されないページについては更新チェック間隔を広げてチェックすることが効率的である．そのために，以下のモデルの(1)式を用いて更新収集の対象ページを選択する．下記のモデルは，重要度が高く，頻繁に更新されるページを高頻度で更新収集の対象として選択する．

更新収集ページの選択モデル

ある時刻 t においてページ p を，更新収集の対象として選択する確率 $\gamma(p,t)$ を以下とする．

$$\gamma(p,t) = \alpha(p) + (1 - \alpha(p))\,\beta(p,t) \quad (1)$$

ここで，$\alpha: p \rightarrow [0,1]$ はページの重要度に基づき与えられるページ選択確率である．β は時刻 t においてページ p が更新されている確率である．ページの更新が指数分布に従うと仮定すれば，ページ p が時刻 t において最終アクセス時間から時刻 t までの間に更新される確率は以下で定義できる．

$$\beta(p,t) = 1 - \exp\{-(t - t_{crawl}(p))/\tau(p)\}, \quad (t \geq t_{crawl}(p))$$

ここで，$t_{crawl}(p)$は，ページ p の最終アクセス時間であり，最後に更新確認のためにそのページ p をアクセスした時間で，更新されなかった場合も含む．$\tau(p)$は，ページ p の平均更新間隔である．

なお，ページの更新回数が 0(新規に追加された URL)もしくは少ない場合に，上記の平均更新間隔を推測するために，Last-Modified の情報などを用いている．ページが更新されたかどうかの確認は，HTTP のステータス・コードだけでなく，HTML ファイルの MD5 チェックサムを利用することでも行っている．

実際の運用で確認してみると，頻繁に更新されるページは，「お知らせ」や広告などの本文とは無関係なごく一部が更新されていることが多い．今後，本文部分のみのページ更新をチェックするなどの対応が必要である．また，更新ページを毎日収集しても，現状の運用では，検索対象にするのに 2 週間程度の遅延が発生する．ページの更新間隔だけなく，[5]のように検索対象の変更間隔も考慮した更新ページの収集が必要である．

4.3 クローラの運用

クローラは，各収集先サイトの robot.txt に記述されたアクセス制限やアクセス間隔，及び，meta タグの情報(<meta name="robots" content="nofollow">)を遵守して運用を行っている††．また，収集先のサイトに過大な負荷をかけないように，同一サイト／ＩＰアドレスに対して約 10 秒の間隔をおいてアクセスを行っている．さらに，万一クローラにバグや設定ミスが発生した場合でも，過度なアクセスが起こらないように，クローラは，プロキシサーバ経由でアクセスを行っている．クローラとは独立したプロキシサーバのアクセス制御プログラムで同一ＩＰアドレスへの過剰なアクセスが発生した場合，そのＩＰアドレスへのアクセスを遮断する対策を行っている．

上記の更新クローラのような収集対象の URL が既知の場合，実測値で，1 日当たり 500 万ページ／ノードの収集が可能である．つまり，単純な収集に限れば，1 ノードでも 20 日で 1 億ページの収集が可能であり，5 ノードを使えば 4 日で 1 億ページを収集できることになる．また，ネットワーク帯域としては，1 日当たり 500 万ページ／ノードの場合，80Mbps 程度が必要となる．現状は，特別な並列分散化は行わずに，更新クローラ，RSS クローラ，ニュースクローラに各 1 ノード，URLDB に 1 ノード，アクセス制御を行うプロキシサーバに 3 ノード，その他のテスト用に 3 ノードの合計 10 ノードで運用している．

クローラの運用に関する課題としては，外部から見た場合，一つの IP アドレスに見える大規模サイトに対して，大量のページの収集が困難なことがあげられる．同一 IP

†† クローラの運用基準は次の URL で公開している．
http://kc.nict.go.jp/project1/crawl-ja.HTML

アドレスには一律に 10 秒間隔でアクセスする制限をかけているため，1 日当り 8,640 ページが収集の限界となる．このため，例えば，収集先サイト内部では複数サーバで構成されており，複数サーバで分散して数百万ページをもつサイトに対しては，全てのページを取得することが困難になっている．また，速度的な問題としては，ページ収集速度よりも，URLDB への新規 URL の登録速度が収集処理全体のボトルネックとなっている．

4.4 クローラの運用に対する問合せ

運用中のクローラに対する外部からの問合せの半数以上は，存在しないページにアクセスしていることに対する問合せである．これは現更新クロールの性質上，新規サイトへのアクセスが比較的少なく，ページ収集が過負荷となる小規模なサイトへのアクセスの絶対数が少ないこと，及び，以前収集した URL を再度収集するため，ディレクトリの構成変更などがあった場合，連続して存在しないページにアクセスするためだと考えられる．この問合せの割合は，[6]に記された問合せの割合(「過大負荷の訴え・アクセス停止要求」が最も多く全体の 6 割以上)と大きく異なっている．また，少数ではあるが，「本来，公開すべきでない情報を公開したため，削除してほしい」などの要求もあり，既に検索対象となっているインデックスや Web ページ標準フォーマットからの削除などの対処を行っている．

4.5 検索・分析対象ページの選択

本基盤では，テキスト情報，URL 情報，リンク情報などを参考として，分析目的に合った質の高いページを選択する．現状の 1 億ページの選択は，収集ページから，以下のページを削除するという，比較的単純な方法で行っている．

- 内容が重複するコピーページ
- 外部からリンクがなくテキストサイズが小さなページ
- 極端にテキストサイズが大きいページ
- cgi の引数が含まれる等 URL が長いページ
- URL のディレクトリの階層が深いページ

5. 検索基盤

5.1 Web ページ標準フォーマットの作成

検索対象となる全ての Web ページは，筆者らが Web ページ標準フォーマット[8]と呼ぶ XML 形式のデータに変換している．収集した HTML ファイルから Web ページ標準フォーマットへの変換は以下の手順で行っている．

1. HTML ファイルをパーズし，タイトル等のヘッダ情報，OutLink の情報(リンク先の URL とアンカーテキスト)，文単位に区切ったテキスト情報を抽出する．

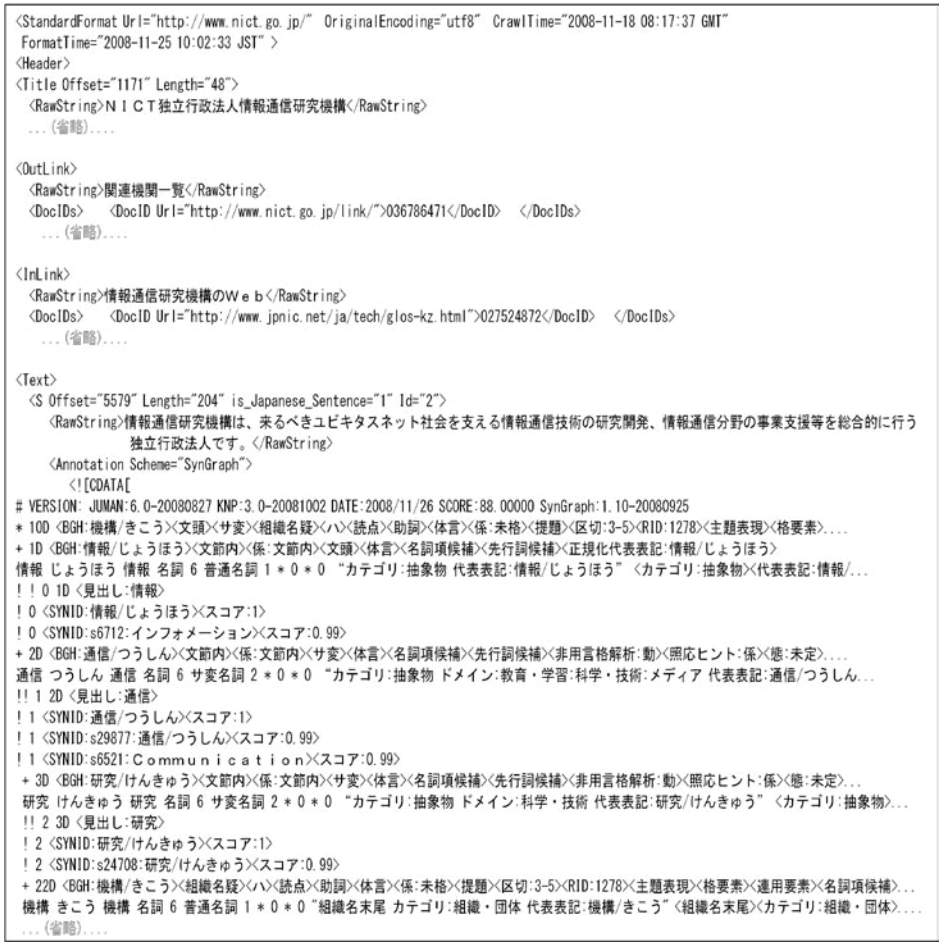

図 3 Web ページ標準フォーマットの一例

2. 抽出した各文に対して，京都大学で開発された以下の言語処理ツールを用いて形態素解析，構文解析，同義表現解析を行い，解析結果を標準フォーマットに出力する．
   - 形態素解析：JUMAN(http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

- 構文解析：KNP (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html)
- 同義表現解析：SynGraph[9]

3. 収集済みの Web ページ集合の OutLink 情報を元に，InLink 情報を作成し，InLink 情報も標準フォーマット上に出力する．

図３に NICT のトップページ(http://www.nict.go.jp/)に対する Web ページ標準フォーマットを示す．図３に示すように，各 Web ページ標準フォーマットには，URL，クロール日時，タイトル，InLink，OutLink，テキスト等の情報が XML 形式で，1 ページ 1 ファイルで格納されている．特にテキスト情報に関しては，形態素解析，構文解析，同義表現解析の結果が付与され，上位のアプリケーションで利用可能となっている．

１億ページの HTML ファイルの容量は圧縮後で約 0.6TB，標準フォーマットの容量は圧縮後で約 6TB である．また，Web ページ標準フォーマットの作成のための処理時間は，1 ノード当たり約 4 万ページ／日である．標準フォーマットの作成は，HTML ファイル単位で独立して実行できるため，単純にノード数を増やすことで，処理の高速化が可能である．現在は，100 ノードをこれに割り当て，約 25 日で 1 億ページの標準フォーマット作成が可能な体制で運用している．

5.2 検索対象の更新

更新／新規ページをできるだけ早く検索可能にし，かつ，インデックスの更新の際にも，常時検索可能な状態にすることを目的として，検索対象の更新手順を設計した．

検索インデックスと Web ページ標準フォーマットは，登録用と検索用の２セットを用意し，更新／新規ページの登録は登録用セットのみに行い，登録したインデックスを検索用セットにコピーする．これにより，検索用セットは常時検索可能である．

通常の更新は，新規登録専用のページ ID を用いて新規専用インデックスを作成し，削除ページは実際にインデックスから削除せずに，検索時に検索結果のページ ID 集合から削除ページを削除することで対応する．これにより，更新は，新規専用インデックスの追加のみで済むため，検索対象の更新が高速にできる．しかしながら，このような運用を繰り返すと，インデックス中に仮削除されて不要になったゴミが増え，インデックスの効率が悪くなる．このため，一定間隔でガベージコレクションを行う．

今後，更新クロールと連携し，以下の手順を自動化して，検索対象の更新を行う予定である．

1. 登録用セットに対して，新規登録専用のページ ID を用いて新規追加ページのインデックスを作成する．また，削除されたページや更新された元のページは削除 ID リストに追加する．
2. 新規追加ページのインデックスのみを，検索用セットにコピーし，検索対象に追加する．また，検索は，削除 ID リストを元に削除ページを検索結果から削除して，出力する．
3. 1.と 2.を繰り返し，新規追加ページ，もしくは，削除ページが一定割合以上になった場合，登録用セットに対して，新規追加ページのページ ID を削除 ID リストの ID に割り当てることで，ページ ID のガベージコレクションを行う．ガベージコレクション後の全てのインデックスと標準フォーマットを検索用セットにコピーし，検索対象をコピーしたインデックスに切り替える．

上記の手順中，検索用セットは常時運用が可能であり，検索サービスが停止するのは，2 のインデックスの追加，及び，3 のインデックスの切り替えの数分間のみである．また，この運用で，ページを収集後 2 週間程度で検索対象に追加することが可能になる予定である．

5.3 検索エンジン基盤 TSUBAKI

1 億ページの Web ページ標準フォーマットを元に，検索エンジン基盤 TSUBAKI[4]を用いてインデキシングおよび検索を行っている．検索インデックスは標準フォーマットに埋め込まれた単語，同義語，およびその係り受け関係について，HTML テキストの本体用とアンカーテキスト用の 2 種類を構築している．1 億ページのインデックス容量は約 3TB である．また，インデックス作成にかかる時間は，1 ノード当たりで約 1 万ページ／時間で，100 ノードを利用することで，約 5 日で 1 億ページのインデックスの作成が可能である．現状，100 万ページを単位として，インデックスの管理，検索を行っており，1 ノード当たり 200 万ページを 50～60 ノードに分散して配置することで，約 1 億ページのインデックスを運用している．

6. 検索対象の Web ページに対する考察

10 億ページを収集し，その内 1 億ページを選択して検索対象とするという本収集・検索基盤の規模は，数百～数千億ページ存在すると言われる Web ページの総数や Google の検索対象数と比較すれば，大きなものではない．日本国内においても，村岡らは日本語のみでなく世界中の Web ページを 100 億ページ規模で収集しており[6]，喜連川らは 9 年間継続して日本語 Web ページを収集することで，累計 100 億ページの日本語 Web ページを収集している[9]．

しかしながら，少なくとも日本語 Web ページについては，1 億ページ規模で，形態素解析，構文解析結果を付与したデータを整備し，随時ページ収集を行いながら，常時検索可能な状態を維持している大規模収集･検索基盤の報告は，筆者らの知る限り，なされていない．

2008 年 4 月に検索対象となっていた 1 億ページおいて，表 1 に示すように，「京都」，「ダイエット」のような単語で 500 万ページ程度がヒットし，「地球温暖化」で 33 万ページ，「アガリクス」のような特殊な単語でも 10 万ページがヒットする．ページ総数上位のドメイン毎のホスト数とページ数は表 2 のようになっており，半数程度の

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

ホスト／ページが jp ドメイン以外のドメインであった．また，ホスト当たりのページ数は最大で 294,810 ページで，今回，RSS クローラやニュースクローラで，新規のブログやニュースページを積極的に収集していることもあり，ホスト当たりのページ数の多いサイトの多くは，blog サイトとニュースサイトであった．さらに，URL やページの特徴(「トラックバック」や「コメント」を表す記述／リンクがある等）から推測した blog ページの総数は 17,730,074 件であり，全体の約 17%が blog ページであった．上記の 1 億ページは，任意の話題が扱える，実データを用いた情報分析の研究開発・評価を行うことが十分可能な規模であると考える．

筆者らは，本収集・検索基盤を整備するために，情報信頼性プロジェクトのかなり多くの労力を費やしている．情報分析を行う全ての組織が独立して，これらの基盤整備を行うことは非効率であり，今後の組織を超えた連携協力等のための参考として，基盤システムの構築と実運用の情報を報告することは有用であると考える．

7. おわりに

本稿では，情報通信研究機構(NICT)で構築し，運用している，Web ページの大規模収集・検索基盤について報告した．構築した収集・検索基盤は，情報信頼性分析システム WISDOM で実運用されており，他の Web 情報分析目的でも利用可能なものとなっている．現在，検索対象の自動更新の枠組みを実装中であり，実装完了後には，ページ収集から検索対象のページの更新までが，ほぼ自動処理で実現する予定である．

本収集・検索基盤を利用している Web 情報分析システム分析結果を確認すると，アフィリエイト目的にコピー／引用したと思われる同一表現が，ノイズになってしまうことがある．今後は，リンクファームなどに対応したページランクや，外観的な特徴から商品販売ページや検索結果ページなどを判定するページタイプ分類を利用することで，より質の高い 1 億ページの集合を検索対象として選択できるようにする予定である．

著作権法の改正により，日本国内でも，研究目的の利用であれば，比較的自由に Web データを共有できる環境が整いつつある．今後は，NICT で収集・整備したデータの公開や，他組織で整備したデータとの連携などに取り組んでいく予定である．

謝辞 Web ページの収集・検索基盤の開発にあたり，クローラのコアエンジンを開発・改良し，提供して頂いた東京大学の田浦健次朗准教授に深く感謝します．また，収集・検索基盤のソフトウェアの開発，及び，日々の運用を行って下さっている原口弘志氏，森井忠史氏，西村晃氏に感謝します．

表 1 検索エンジンのヒット件数

クエリ	ヒット件数	
	TSUBAKI(1 億ページ)	Google
京都	5,031,237	119,000,000
神戸	2,231,908	16,300,000
金閣寺	36,841	758,000
ダイエット	4,938,926	108,000,000
地球温暖化	333,268	5,350,000
アガリクス	100,667	850,000
ステロイド	72,323	469,000

表 2 ドメイン毎のホスト数／ページ数

ドメイン	ホスト数	ページ数
jp	647,621	54,406,895
com	573,729	31,991,283
net	41,374	10,583,960
info	27,251	2,191,098
biz	18,283	844,092
to	7,634	565,875
...		
合計	1,500,090	105,360,017

参考文献

1) Web as Corpus Workshop, http://webascorpus.sf.net/WAC4/
2) 黒橋禎夫: 情報の信頼性評価に関する基盤技術の研究開発,人工知能学会誌，Vol.23，No.6，pp.783-790, 2008.
3) S. Kurohashi, S. Akamine, D. Kawahara, Y. Kato, T. Nakagawa, K. Inui and Y. Kidawara: Information Credibility Analysis of Web Contents, In Proceedings of the Second International Symposium on Universal Communication, pp.146-153, 2008.
4) K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi: Tsubaki: An open search engine infrastructure for developing new information access methodology, In Proceedings

Vol.2009-DBS-148 No.14
Vol.2009-FI-95 No.14
2009/7/28

of the Third International Joint Conference on Natural Language Processing (IJCNLP2008), pp. 189–196, 2008.

5) C. Olston and S. Pandey: Recrawl Scheduling Based on Information Longevity, Proceeding of the 17th international conference on World Wide Web(WWW2008), pp. 437-446,2008.

6) 村岡洋一，山名早人，松井くにお，橘本三奈子，赤羽匡子，萩原純一：100 億規模の Web ページ収集・分析への挑戦，情報処理，Vol.49， No.11， pp.1277-1283，2008.

7) K. Shinzato, D. Kawahara, C. Hashimoto and S. Kurohashi: A Large-Scale Web Data Collection as a Natural Language Processing Infrastructure, in Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC08), 2008.

8) T. Shibata, M. Odani, J. Harashima, T. Oonishi and S. Kurohashi: SYNGRAPH: A Flexible Matching Method based on Synonymous Expression Extraction from an Ordinary Dictionary and a Web Corpus, In Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008, poster), pp.787-792, 2008.

9) 喜連川優，豊田正史，田村孝之，鍛治伸裕，今村誠，高山泰博，藤原聡子：過去 9 年に及ぶ Web アーカイブからの社会の動きを読む， 情報処理， Vol.49， No.11， pp.1290-1296，2008.