Web 情報分析のための大規模 Web ページの収集・選択・検索

赤峯享*† 加藤義清* 河原大輔* レオン末松豊インティ*
新里圭司‡ 乾健太郎*§ 黒橋禎夫*‡ 木俵豊*

*情報通信研究機構 †奈良先端科学技術大学院大学 §東北大学 ‡京都大学
{akamine, ykato, dk, yutaka, inui, kidawara}@nict.go.jp {shinzato, kuro}@nlp.kuee.kyoto-u.ac.jp

1. はじめに

Web は，これまでにない情報／知識の宝庫となった．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，企業情報，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．今や，大規模な Web コンテンツが利用できることは，コーパスを用いた言語処理技術や Web 情報を用いた情報分析技術の研究開発において，必要不可欠なものとなっている．

筆者らは，Web 情報の信頼性評価を支援する Web 情報信頼性システム WISDOM[1]の研究開発を行っている[1]．WISDOM は，様々な個人や組織が発信した Web 情報を多面的に分析し，俯瞰的に表示する（例えば，利用者が入力した検索課題に対して，営利団体・報道機関・個人等の発信者クラス毎に主要な意見と対立する意見を分類して表示する）ことで，利用者がどの情報が信頼できるかを判断することを支援する．

大規模 Web ページを対象として，このような情報分析を行う場合，(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．さらに，商用検索エンジンのランキング方法は公開されていないため，特定のサイトや発信者クラスに偏ったページのみがランキングの上位，つまり分析対象に多く含まれてもそれを検知し，制御する術がない．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[2]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，WISDOM で利用するために，情報通信研究機構(NICT)で構築し，運用している Web ページの収集・検索基盤について，特に検索・分析対象ページの選択と更新の方法について報告する．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．

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

2. 開発方針

検索基盤を利用する上位の情報分析アプリケーションは，一般の実ユーザの利用を元に評価し，それをフィードバックして改良を行うことが重要である．ユーザに情報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(トピック)に対応できることが必要である．また，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任意のトピックで評価でき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．これらは個々のテキスト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．

一方，単一の組織で利用できる計算機資源は限られており，筆者らの環境では，初期の運用実績から見積もって，検索・分析可能なページ数は 1 億ページ程度であった．Web ページの総数は，日本語に限定しても，少なく見積もっても数十億ページは存在しており，全ての Web ページを検索・分析対象とするのは不可能である．また，仮により大規模な計算機資源が利用できるとしても，商品販売サイトや地域情報サイトなどで動的に生成される，文章を含まないページを大量に収集して，全てを検索・分析対象とすることは，情報分析として非常に効率が悪い．したがって，筆者らは以下の方針で，収集・検索基盤を構築，及び，検索・分析対象ページの選択を行った．

- 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，逐次更新しながら，検索・分析対象とする．
- ページ収集は 10 億ページ規模で行い，その中から分析対象として適した質の高ページをバランス良く選択することで，約 1 億ページを検索・分析対象とする．
- HTML ファイルからテキスト情報を抽出して，文に分割し，形態素解析，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．

3. システム構成の概要

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．

収集・登録時は以下の手順で動作する．

1. Web クローラは，インターネット上の Web ページを最大 10 億 URL 収集する．また，収集した Web ページを一定間隔で更新する．
2. ページ選択 1 は，収集した Web ページから，URL，及び，ページランク等のリンク解析結果を元に，Web 標準フォーマット作成対象として，約 2 億ページの高品質ページを選択する．
3. Web ページ標準フォーマット作成部は，まず，HTML ファイルをパーズし，タイトルなどのメタ情報抽出，及び，文区切り解析により文抽出を行う．次に，抽出した文に対して，京都大学で開発された言語解析ツールを用いて，形態素解析(JUMAN)・同義表現解析(KNP)・構文解析(SYNGRAPH)を行い，解析結果を Web ページ標準フォーマット[3]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．
4. ページ選択部 2 は，Web ページ標準フォーマット中の文数やページ種別(ブログ，QA，ニュースなど) などの情報を用いて，分析目的に合った 1 億ページを選択する．
5. 検索インデックス作成部は， Web ページ標準フォーマットを元に，TSUBAKI をインデックスの作成を行う．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．

また，検索／分析時は以下の手順で動作する．

1. 検索部(TSUBAKI)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．現状，WISDOM では上位 1000 件を利用している．
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．

4. Web ページの選択収集と更新

4.1 クローラの種別と運用

筆者らは，10 億ページを収集し，その中から一定の品質を保った，できるだけ最新のページを選択して，検索・分析対象とするという方針で，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・QA クローラ
  主要なニュースサイトや QA 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･QA 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可能である．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，実際のインターネット上のページと同期した最新のページを検索対象とすることが可能である．

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 日あたり 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 40 日で 10 億ページを収集できることになる．

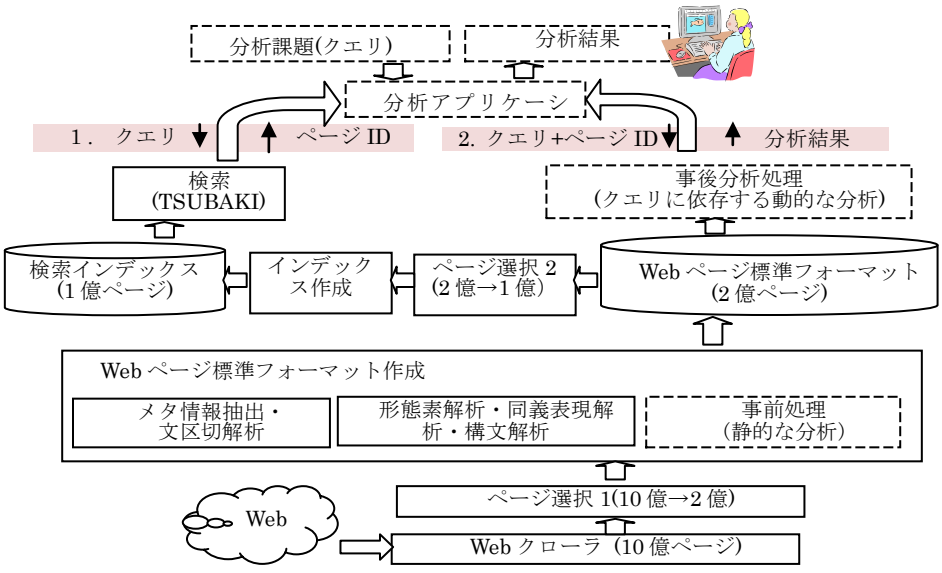

図 1. システム構成

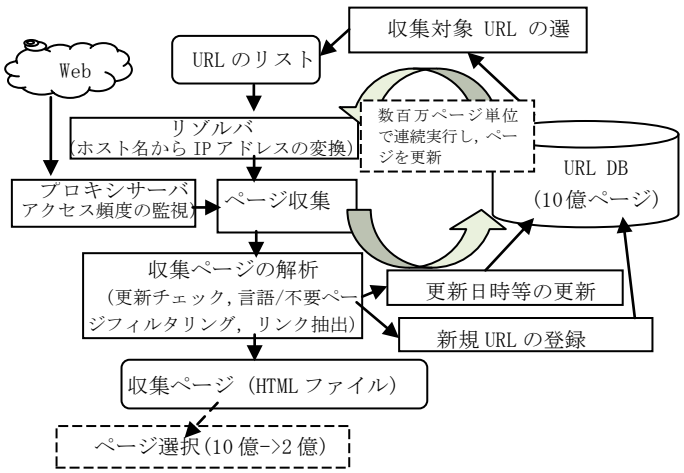

図 2. Web クローラの構成

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 に格納された最大 10 億ページの URL に対して，更新チェック用の URL を数百万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．

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
   - インターネット上の 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．

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．

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

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．

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

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．また，RSS クローラ，及び，ニュース・QA クローラが収集ページも URL DB に登録し，収集した全ての URL に対して，10 桁の一意の ID を付与することで，URL DB 上で全ページの更新情報等を一元管理している．

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も多い．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．

5. 検索・分析対象ページの選択と更新

5.1 検索・分析対象の選択

本基盤では，分析目的に合った質の高いページを選択するために，収集した Web ページを段階的に選択している．第一段階のページ選択 1 では，URL 情報，リンク解析結果，HTML ファイルの MD5 チェックサムの情報を用いることで，10 億ページから 2 億ページの選択を行い，第二段階のページ選択 2 では，テキスト解析結果，及び，サイト／ページ種別の情報を用いて 2 億ページから 1 億ページの選択を行う．

具体的には，第一段階のページ選択 1 では，以下により，10 億ページから 2 億ページの選択を行う．

- リンク解析により，リンクファームなどのスパムページの可能性が高いページ削除する．
- MD5 チェックサムが一致する同一内容のページを，代表ページを残して削除する．
- 残ったページに対して，以下により，ページ重要度のスコア付けを行い，上位ページを選択する．
  - ✓ リンク解析結果のページランクの値，及び，ドメイン単位のページランクの値
  - ✓ URL の階層の深さ(深い程スコアが低い)
  - ✓ URL の cgi の引数の長さ(長い程スコアが低い)
  - ✓ クロール日時(新しいほどスコアが高い)

また，第二段階のページ選択 2 では，以下により，2 億ページから 1 億ページの選択を行う．

- URL DB で，削除／更新済みのページを削除する．
- ページ選択 1 で用いたページ重要度のスコアリングと合わせて，テキストの重要性を元にスコア付けを行う．現状は，一定の文字数以上の文数を元にテキストの重要性をスコア付けしている．
- ページ数の多いサイトについては，サイト毎にページ数の上限を求めて，上限を超えた場合は，スコアの小さいページを削除する．現状は，動的に大量のページが生成される商品販売サイトと地域情報サイトのみに対して人手で上限を設定している．

5.2 検索・分析対象の更新

更新／新規ページをできるだけ早く検索・分析可能にし，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，検索対象の自動更新処理を設計した．更新処理は，毎日，もしくは，数日毎に以下の処理をパイプライン的に実行する．また，更新処理を一定回数繰り返すとインデックス等が肥大化するため，ガベージコレクションを行い，インデックス等から削除済み(更新済み）ページを実際に削除する．

- Web クローラによって新規・更新収集されたページは，ページ選択 1 を用いて最大 2 億ページ以下になるように選択し，Web 標準フォーマットを作成する．
- ページ選択 2 を用いて，Web 標準フォーマットから，検索対象が約 1 億ページになるように検索対象ページを選択する．
- 検索インデックス作成においては，新規・更新されたページは追加登録用ノード上に新規インデックスを作成することで，順次，検索対象にする．この際に，削除・更新ページは既存インデックスからは削除せずに，検索時に削除する．

6. 検索・分析対象ページに対する考察

10 億ページを収集し，その中から分析に適した 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[4]，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[5].

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，分析対象に適したページを選択し，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．

2009 年 12 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」のような単語で約 630 万ページ程度がヒットし，「地球温暖化」で約 41 万ページ，「アガリクス」のような特殊な単語でも約 12 万ページがヒットする．また，ページ総数上位のホスト毎のページ数は表 2 のようになっており，QA サイト，blog サイト，ニュースサイトが上位を占めている．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 ページの総数は約 1800 万ページであり，全体の約 18%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．

表 1 検索エンジンのヒット件数

クエリ	ヒット件数	
	WISDOM/TSUBAKI (1 億ページ)	Google
京都	6,298,228	118,000,000
ダイエット	5,603,659	123,000,000
地球温暖化	408,695	5,670,000
アガリクス	126,373	644,000
ステロイド	87,304	1,330,000
バイオエタノール	23,156	363,000

表 2 サイト毎のページ数

サイト	ページ数
chiebukuro.yahoo.co.jp	2,835,899
okwave.jp	657,797
news.livedoor.com	436,368
blog.livedoor.jp	414,616
plaza.rakuten.co.jp	401,151
ameblo.jp	327,190
news.www.infoseek.co.jp	309,324
d.hatena.ne.jp	245,721
news.goo.ne.jp	236,416
blog.goo.ne.jp	235,546
sankei.jp.msn.com	199,067
www.yomiuri.co.jp	129,155

7. おわりに

本稿では NICT で構築し，運用している，Web ページの収集・検索基盤について報告した．本基盤，及び，本基盤で構築したデータは，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．今後は，ページ種別やページ中の文のオリジナル度などを考慮することで，ページ選択の高精度化を図る予定である．

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・選択・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．

参考文献

[1]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.

[2]K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi: Tsubaki: An open search engine infrastructure for developing new information access methodology, IJCNLP2008, pp. 189–196, 2008.

[3]K. Shinzato, D. Kawahara, C. Hashimoto and S. Kurohashi: A Large-Scale Web Data Collection as a Natural Language Processing Infrastructure, LREC08, 2008.

[4]村岡洋一，山名早人，松井くにお，橋本三奈子，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.

[5]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.