大規模 Web 情報分析のための分析対象ページの段階的選択

赤峯享*† 加藤義清* 川田拓也* レオン末松豊インティ*
河原大輔*‡ 乾健太郎*§ 黒橋禎夫*‡

*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, tkawada yutaka}@nict.go.jp　inui@ecei.tohoku.ac.jp　{dk, kuro}@i.kyoto-u.ac.jp

1. はじめに

インターネットは他に比するものがない巨大な情報の宝庫となった．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，日常の体験を綴ったブログなど様々な情報が日々発信されている．人々は，商品の購入，健康の管理，病気の治療，政策の善し悪しの判断などの意思決定を行う際に，これらの Web 情報を参考にするようになってきており，その支援を行う情報分析システムが望まれている．ここで情報分析とは，人の意思決定等を支援するために，クエリに関連するページ集合から発信者や意見などを抽出し，利用者にページ集合の全体像を多角的に提示したり，特徴的な発信者，意見，ページなどを提示したりする処理を指す．

情報分析は計算コストのかかる重い処理なので，関連するページ集合の一部を選択して実行する必要がある．これを分析対象と呼ぶ．Web は様々な種類の文書が混在しており，2 章で述べるように分析対象として適したページと適さないページがある，しかも，分析対象に適さないページが大量に存在する．したがって，情報分析では，分析対象として適したページ集合を選択することが重要な課題となる．

本稿では，10 億ページ規模の大規模 Web ページを収集して，分析対象となる 1 億ページ規模の Web ページ集合をクエリ独立で選択するための方式を提案し，予備調査の結果を報告する．本方式は，以下を特徴とする．(1)Web ページの選択を，商品カタログページやコピーページなどの不適格ページのフィルタリングと，ページランクやテキスト内容の品質等でバイアスをかけた重み付きサンプリングで行う．(2)多段階で選択を行い，計算コストのかかる後段の処理結果を前段にフィードバックする．

2. Web 情報分析の分析対象ページの問題

インターネットは，極めて低いコストで情報発信が可能であるため，品質の低いページが大量に存在する．特に，スパムページや，商品カタログ等のデータベースから自動生成される商品販売ページは，オリジナルの情報のコピーや切り貼りで自動生成が可能であるため，コストをかけずに無尽蔵に作成でき，しかも，情報分析の対象には適さないことが多い．このような情報分析に適さない品質の低いページが分析対象に含まれた場合，以下の問題が発生する．

- 利用者に役に立つ分析結果を提示できない．

  スパムページや商品カタログページばかりを分析対象としてしまうと，意見分析等の分析精度がいくら高くても，利用者にとって有益な分析結果は得られない．
- 計算機リソースを消費する．

  インターネットは，日々新しいページを無尽蔵に供給可能な情報源であり，分析対象ページはその一部分のある時点でのスナップショップと言える．一般に品質の低いページほど大量作成が可能である．そのため，収集対象を選別せず，単純な幅優先探索等でページを収集して，分析対象に加えた場合，大量に存在する低品質ページの収集やインデキシング等の処理に計算機リソースを費やし，分析に適した品質の高いページを分析対象に加えることができなくなる．

これらの問題は，有限の均質な文書の集合である論文や新聞記事を対象とした情報分析では発生しない問題である．また，従来の Web 検索では，利用者が通常アクセスするのは検索結果の上位の数ページであり，検索結果の下位（例えば，ランキングで数百番目のページ以降）に大量に品質の低いページが存在しても，大きな問題にはならない．一方で，Web 情報分析では，検索結果の上位数百～数千ページを対象として情報を抽出し，分析を行うため，低品質な下位ページの存在は，分析精度に深刻な悪影響を与えやすい．

本稿では，筆者らが開発し，運用している Web 情報分析システム WISDOM[1]の環境を例として，この問題を議論する．WISDOM は，ページ収集から情報分析までの全ての処理を 240 ノード(1 ノード当り 4CPU core，メモリ 8GB，ローカルディスク 2TB)のクラスタ計算機と 200TB のファイルサーバを用いて行っている[2]．提案する選択方式は，WISDOM に限定したものでなく，ブログの評判情報分析などを含む一般の Web 情報分析システムでも利用可能であり，WISDOM より小規模／大規模なシステムでも利用可能である．

3. 分析対象ページの選択方法

3.1. 選択の方針

分析対象として適したページの選択は，クエリ依存で行う選択とクエリ独立で行う選択があるが，本

稿では後者のクエリ独立の選択にフォーカスする．WISDOM のような Web のテキスト情報を分析する Web 情報分析システムでは，以下のようなページ／ページ集合が分析対象であることが望ましく，これらのページを選択することを基本方針とした．

品質の高いページ 人気のある新しいページだけでなく，テキスト内容が充実したページが望ましい．なお，本稿では，分析対象としての適合度を元に品質を評価する．そのため，例えば，画像や映像だけでテキストのないページは，いかに完成度が高くとも情報分析に適さないため，低品質ページと表現する．

インターネット上と更新の同期の取れたページ インターネット上でページが更新された場合，更新前の古いページを分析対象から外し，同期の取れた新しいページを分析対象にするのが望ましい．

多様な発信者／サイトを含むページ集合 単一の発信者／サイトのページばかりでなく，様々な発信者／サイトの多様な意見が分析対象であることが望ましい．

3.2. 品質によるページの選択

クエリ独立の分析対象ページの選択では，任意のページを品質の順に並べることは人間でも困難である．しかしながら，商品カタログページや写真集のページなどの特に分析対象に適さない低品質ページは，そのページ単独で，分析対象として不適格であると判断は可能である．それ以外のページも，大まかな品質の評価は可能である．例えば，分析課題のクエリに依存しなくても，Wikipedia の記述の充実したページは，数行の日記のブログページよりも品質が高いという判断は可能である．

そこで，品質によるページの選択は，以下の 2 段階で行うこととした．

フィルタリング 商品カタログ等の情報分析に特に適さないページは，そのページ単独で絶対的に不適格ページと分類し，情報分析ページから外す．また，スパムページや，他のページと同様の内容の Near Duplicate ページもフィルタリングして分析対象から外す．

重み付きサンプリング ページランク，文数，特定の単語の出現数などの属性から品質スコアを求め，スコアの高いページが選択されやすくなるようなバイアスをかけてページをサンプリングする．サンプリング方式は，一般的な重み付きサンプリング[3]を用いる．

3.3. 段階的選択とフィードバック

フィルタリングやサンプリングの重み付けで用いるページの属性としては，URL の階層，更新日時，ページサイズ，ページランクなどのメタ情報，及び，文数，特定の単語・構文の出現数などのテキスト情報が考えられる(表 1)．選択の精度を上げるには，メタ情報だけでなく，テキストの内容を利用する方が有利である．しかしながら，それには，個々の Web ページに対してテキストの抽出，文への分割，形態素・構文解析などの処理を行う必要があり，計算コストが高い．例えば，筆者らの運用環境では，10 ノードの PC で約 1000 万ページ／日のページ収集が可能であるが，html ファイルからテキストを抽出して，文切り，形態素解析・構文解析を行えるのは，70 ノードの PC を利用しても約 100 万ページ／日である．したがって全ての収集ページに対して言語解析を行うことは困難である．

そこで，サイト単位で品質を評価することを考える．一般に同一サイトには、同種のページ（高品質ページ／低品質ページ）が集まりやすい．特に低コストで自動生成される低品質ページは，同一サイトに同種のページが大量に存在しやすい．この性質を利用すれば，サイト内の品質評価済みのページの情報を元に、未評価のページの品質を推定することが可能である。段階的に選択を行い，計算コストのかかる後段の結果から，サイト単位の品質を求め，それを前段にフィードバックすることで．計算コストを下げることが可能である．

3.4. 分析対象ページの多様性

情報分析の目的の一つとして，少数意見を含めて，様々な人の多様な意見を発見することがある．同じサイトのページばかりが大量に分析対象になって，その分析結果が利用者に提示されても，有益な情報は得られない．多様な意見を抽出するためには，品質が高いページだけでなく，多くの発信者やサイトのページを選択することが望ましい．そこで，品質順にランキングした上位ページを決定的に選択するのではなく，選択するページを確率的にサンプリングすることで，特定のサイトのページに偏って選択する危険性が減らし，分析対象ページの多様性を確保する．また，サイト単位の情報を扱うことで，例えば，新聞社のサイトや QA サイトなどの特定のサイトの重みを人手で増す／減らすなどの調整も可能である．

表 1 :ページ選択に利用する属性

属性			ページ単位での利用可能箇所
ページ中のテキスト内容	テキスト量	文の数・長さ・密度	図 1 の(4)
	文体	助動詞，感動詞，終助詞，絵文字 の種別と出現数	
	専門性（名詞）	病名，専門用語の出現数	
	具体性（固有名詞）	組織名，人名の出現数	
	高品質ページに出やすい単語	「検証」，「証明」等	
	低品質ページに出やすい単語	「死ね」，「おまえ」等	
	アダルトページに出やすい単語		
	高品質ページに出やすい構文を作る単語	意見，原因・理由，比較	
	ページの種別	ニュース，ブログ，商品販売，リンク集	
ページ中の情報の有無	広告量	アフィリエイトサイトへのリンク数	
	連絡先	住所，電話番号，メールアドレス の有無	
	プライバシーポリシーの有無		
メタ情報	ページランク		(1)～(4)
	OutLink の数		(2)～(4)
	ページのサイズ		(2)～(4)
	更新日	現在の時間からの差	(2)～(4)
	URL	階層，長さ，クエリ	(1)～(4)

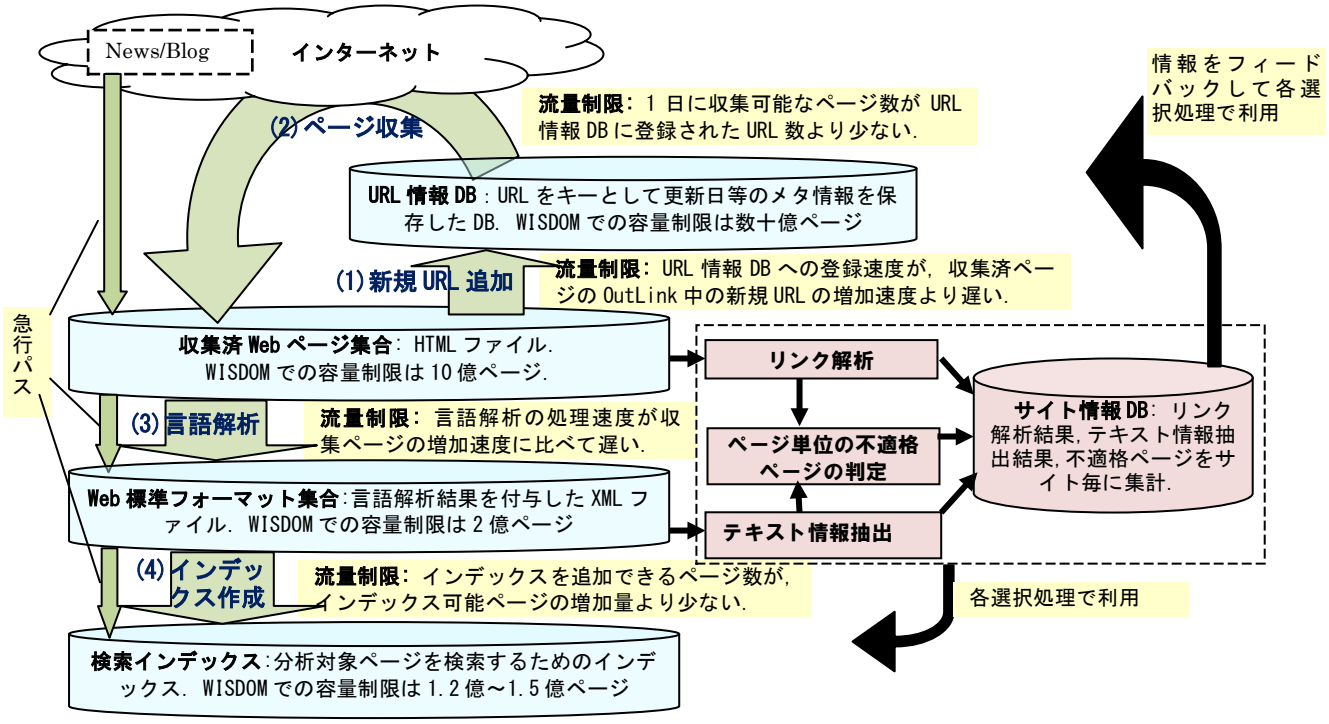

図 1: 分析対象ページの段階的選択の手順

4. ページ収集・検索基盤上での実現

ページ収集から検索インデックス作成までの処理で, 計算機リソースの制限は, 以下の 2 種類がある.

計算速度の制限(流量制限) 入力データ量に対して, 次段階のデータを作成／登録する処理速度が追いつかないため, 流量の制限が必要となる. この制限に対応するために, 入力のデータから不適格ページをフィルタリングし, 残りのページについても重み付きサンプリングを行い, 選択されたページのみを次段階に送る必要がある.

データ量の制限(容量制限) 保持可能なデータ数に限界があるため, 総容量の制限が必要となる. この制限に対応するためには, 新規に追加したデータ数と同数のデータを既存のデータから削除する必要がある.

図 1 に分析対象ページの段階的選択の手順を示す. 計算機リソースの制限のために, 段階的に情報分析に必要なデータを作成しつつ, 分析対象ページの選択を行う. ページ収集から検索インデックス作成までの手順は次の通りである.

(1) 収集済みページの OutLink から抽出した新規 URL を URL 情報 DB に登録する.
(2) URL 情報 DB から収集対象の URL を選択し, インターネット上のページを収集する.
(3) 収集済ページを解析し, 文に分割し, 各文に形態素解析・構文解析等の言語解析を行い, Web 標準フォーマット[4]の形式で保存する.
(4) 言語解析済みページから検索インデックスを作成する.

図 1 に示すように, 上記の全ての箇所で, 流量制限と容量制限への対応が必要となる. また, 各ページに対して求めたリンク解析結果のページランク, テキスト情報抽出結果の文数や単語出現数, 不適格ページか否か等の情報をサイト単位で集計し, 平均値をとり, フィードバックすることで, これらの情報を前段の選択処理に利用することが可能となる.

Web 標準フォーマットの 2 億ページの選択(図 1 の (3))を例として, 選択処理の手順を述べる. 他の選択箇所も同様の考え方で行う.

- 流量制限による追加処理
  - 新規収集された Web ページに対して, サイト情報 DB の不適格サイトを用いて, 不適格ページをフィルタアウトして選択対象外とする.
  - 残りのページに対して, 重み付きサンプリングを行い, 流量制限で決まる最大の件数分のページを選択する. 品質スコアは各属性のスコアにバイアスをかけて総和をとる. 属性は, ページ単位の属性だけでなく, サイト単位の属性も用いる. サイト単位の属性を利用することにより, 後段で抽出されたテキスト内容の属性も利用可能となる. また, ページ更新の同期をとるために, 更新されたページは優先的に選択されるように別途重みを増やす.
- 容量制限による削除処理
  - 収集済み Web ページ集合から Web 標準フォーマット集合へ新たに追加されるページ数と同数のページを Web 標準フォーマット集合から削除する. 削除するページは, 追加処理と同様の情報を用いて, まず, フィルタリングで不適格ページを求め, 次に品質スコアの逆数を用いてサンプリングを行い, 削除ページを

決定する．ただし，全体の整合性をとるため，下流(検索インデックス）で使われているページは削除しない．

WISDOM の実運用では，ニュース記事とブログ記事については，例外的に急行パスを用いて高速に分析対象に追加している．特定ニュースサイトと RSS でフィードされるページは専用のクローラを動かして収集を行い，流量制限を優先的に使い，容量制限をチェックせずに，Web 標準フォーマットの作成，検索インデックスへの登録を行っている．

5. 予備調査

5.1. フィルタリング対象の不適格ページ

フィルタリング対象となる不適格ページが，実際にどの程度存在するかを調査するために，人手で不適格ページを判定した評価用データを作成した．

評価対象の Web ページは，3 種類の Web 検索エンジン(WISDOM で利用している「検索エンジン基盤 TSUBAKI」，「Yahoo!検索 Web API V1」，「Yahoo! 検索 Web API V2」)で，100 クエリの検索結果の上位 1000 ページ(「Yahoo!検索 Web API V2」は取得限界の上位約 300 ページ）を取得し，検索エンジン毎にランダムに約 1000 ページを選択した．なお，検索クエリは，情報分析の入力を想定して，「コーヒーは健康に悪い」，「赤ちゃんポスト」など，WISDOM で評価用に作成したクエリ，及び，WISDOM の運用で実際に入力されたクエリを用いた．

評価者に不適格ページの基準として以下のような情報を与え，不適格ページであるかないかの 2 値で判定した．

- 人や機械の錯誤を目的としたスパムページ
- 商品レビューや商品の解説記事などを含まない商品販売目的ページ
- ナビゲーションのためのリンク集やメニューページ
- 画像集や時刻表などテキスト情報を含まないページ

表 2 に検索エンジン毎の不適格ページの数と割合を示す．それぞれの検索エンジンにおいて，16%から 29%の不適格ページが含まれていた．検索結果の上位のページは，ページランクが高い等，比較的品質の高いページが多いはずであり，検索対象の全ページや収集ページには，この割合以上に不適格ページがあると考えられる．この調査により，分析対象ページの選択に不適格ページのフィルタリングが重要なことが確認できた．

表 2 不適格ページの数と割合

検索エンジン	不適格ページの数	適格ページの数	検索エンジンの取得ページ
TSUBAKI	320(29%)	774(71%)	上位 1000 件
Yahoo V1	204(19%)	858(81%)	上位 1000 件
Yahoo V2	174(16%)	900(84%)	上位 300 件

5.2. ページ品質の可評価性

重み付きサンプリングが有効に働くためには，クエリ独立に，Web ページに対して品質による重みが与えられることが前提となる．その前提を確認するために，人がページの品質を評価できるかの調査を行った．

評価対象のページは，前節の TSUBAKI の評価対象の 1 000 ページから，不適格ページを除き，残りのページからランダムに抽出した 100 ページを用いた．5 名の評価者それぞれが，同じ評価対象ページに以下の 5 段階の品質スコアを付与し，5 名の間でスコアの相関関係があるかを確認した．

5: 分析対象として非常に役立つページ
4: 分析対象としてかなり役立つページ
3: 分析対象として役立つページである
2: 分析対象として多少役立つページ
1: 分析対象として役立つとは言えないページ

5 名から 2 名のペアを取り出し，ペア毎に相関係数をとったところ，最も強い相関係数のあるペアで相関係数が 0.64 で，平均の相関係数が 0.52 であり，かなり強い正の相関があることが確認できた．したがって，計算機で同様の評価を再現できれば，重み付きサンプリングを用いることで，ランダムにサンプリングするよりも，品質の高いページ選択が可能であると考えられる．

6. おわりに

本稿では，10 億ページ規模の大規模 Web ページを収集して，分析対象となる 1 億ページ規模の Web ページ集合をクエリ独立に選択するための方式を提案した．本方式は，計算機リソースの制限を考慮して，フィルタリングと重み付きサンプリングを行うこと，計算コストのかかる後段の処理結果を前段にフィードバックすることを特徴としている．

今後は，今回，調査用に作成したデータを用いて，フィルタリングと重み付きサンプリングの性能評価を行う予定である．また，本選択方式を，WISDOM の収集・検索基盤に組み込み，実際の Web 情報分析システム上で有効性の確認を行う予定である．

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] 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．Web ページの大規模収集・検索基盤の構築と運用，情報処理学会 データベースシステム・情報学基礎 合同研究発表会 DBS-148・FI-95, 2009.

[3] Pavlos S. Efraimidisa and Paul G. Spirakis Weighted random sampling with a reservoir Information Processing Letters Volume 97, Issue 5, 16 March 2006, Pages 181-185.]]]

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