Web情報の信頼性検証のための情報分析システムWISDOM

赤峯享* 宮森恒* 加藤義清* 中川哲治* 乾健太郎* 黒橋禎夫*† 木俵豊*
*情報通信研究機構 †京都大学
{akamine, miya, ykato, tnaka, inui}@nict.go.jp, kuro@i.kyoto-u.ac.jp, kidawara@nict.go.jp

1. はじめに

企業の製品情報，政府の広報，個人の体験を記したブログ等，様々な発信者の多種多様な情報がインターネット上に流通するようになった．その結果，ショッピング，観光，健康管理等の日常生活から，製品開発の立案，政策の判断に至る様々な場面において，有益な情報がインターネット上に大量に存在している．今後，一般の人がこれらのWeb情報を用いて様々な重要な意思決定を行うようになると考えられる．

しかしながら，一方でWeb情報には，個人の誤解に基づく誤った情報や，特定の観点からの偏った意見等の信頼性の低い情報も大量に流通している．さらに,Web情報の信頼性を検証する手段が十分に整備されてない．そのため，現状，一般の人がWeb情報を参照して意思決定するにはリスクが伴い,Web情報を安心して利用できない状況にある．

筆者らは，利用者が関心をもつトピックに対して，Web上の様々な発信者の情報を多面的に分析し，Web情報の全体像を様々な観点から集約して提示することで，利用者がWeb情報の信頼性を検証することを支援するシステムの開発を目指している[1,2]．これまでに，アガリクスやマイナスイオン等の利用物，捕鯨問題やCO2問題等の社会問題のトピックを分析対象として，各トピックに対して100ページ程度を分析対象とした評価用データを作成し，評価用データで動作するプロトタイプシステムを開発した[2]．

本稿では，プロトタイプシステムを元に，検索エンジン基盤TSUBAKI[3]と結合することで，任意のキーワードに対して，数億ページ規模のWebページを対象とした分析を行う情報分析システムWISDOM(Web Information Sensibly and Discreetly Ordered and Marshaled)の開発方針とシステム構成について報告する．

2. 開発方針／特徴

筆者らは，以下の開発方針で情報分析システムWISDOMの開発を行っている．

- 様々なWeb情報を多面的に分析する．
- 自然言語処理技術を活用することで，高精度の情報抽出・分析エンジンを開発する．
- 実用規模の評価検証システムを構築する．

2.1 様々なWeb情報を多面的に分析

現状，Web情報をアクセスする場合，多くの人は，Web検索エンジンを利用している．Web検索エンジンは，ユーザが入力したトピック（検索クエリ）に適合する人気ページを一次元でランキングする．そのため，様々なページでそのトピックがどう評価されているかを知ろうとした場合，検索結果の上位から順に全てのページを参照する必要がある．例えば，特定の製品名で検索した場合，検索結果の上位のページの多くが，企業による製品紹介ページだとすると，その製品を実際に使用したユーザが記したページにアクセスするには非常に手間が掛ってしまう．一般にWeb検索エンジン利用者の多くは，検索結果の上位数ページにしかアクセスしないため,Web上に製品購入の判断のための有益な情報(例えば，その製品の欠陥情報)が大量にあっても検索結果の上位になければ見逃してしまうことになる．

一方，WISDOMでは，トピックに関連する数百～千ページ規模のWebページから，以下の観点でWebページから情報を抽出し，多面的に分類することで，トピックに関するWeb情報の全体像を把握することを支援する．

- 情報発信者
  情報を発信した著者，サイト運営者，及び，著者とサイト運営者の関係を表わす情報発信タイプをWebページから抽出する．情報発信者抽出の詳細については参考文献[6]に記す．
- トピックに関する評価情報(意見)
  Webページ中のテキストからトピックに関する評価情報を（評価保持者，評価対象，評価表現，極性）の４項目で抽出する．評価情報は，客観的な事実を表わす「製品Aは3日で壊れた」,「健康食品Bで風邪が治った」等の従来の主観表現よりも広い表現を対象とする．評価情報抽出の詳細については参考文献[7,8]に記す．
- ページ外観
  Webページ中に，電話番号や住所等の実世界の

連絡先が含まれるか，アフィリエート等の広告が含まれるか等，Web ページの信頼性に関連するページの外観的な特徴を抽出する．

上記により，「誰（どういう属性をもった人・組織・ページ）がどういう情報（肯定評価・否定評価）を発信しているのか」を抽出して，発信者クラス別，肯定否定別，広告量別等に Web ページを分類して提示し、実際のページや評価情報に簡単にアクセスすることを可能とする．さらに，利用者がその情報を信頼性するかどうかの手掛かりとなる，以下のような分析を可能とする．

- 製品 A は，全体では肯定評価と否定評価の数は同じくらいだが，肯定意見は企業がサイト運営者のページか，広告が多いページばかりである．
- 行政・新聞社・学会がサイト運営者のページでは，製品 A に関する否定評価のみである．
- 個人ページでは，製品 A の肯定評価と否定評価の両方があるが，否定評価が多い．

2.2 自然言語処理技術の活用

大規模コーパスの利用による統計言語処理技術の発達や，機械翻訳等のシステムの開発による言語資源の蓄積により，固有表現抽出や述語項構造解析については高精処理のためのツールや辞書等の環境が整いつつある．これらの研究成果を基盤に，さらに高度な言語処理技術を研究開発することで，様々な Web 情報に対して実用精度で動作する情報抽出・分析エンジンを開発する．

2.3 実用規模の評価検証システムを実現

様々なトピックに対して，実用規模での評価検証を行うために，数億ページ規模の日本語 Web ページを対象とした分析システムを構築する．そのため，本システムは，PC クラスタを用いた大規模並列計算機環境上に構築する．また，ノード（PC）を追加することで，ページ規模の拡大や処理速度の高速化が可能なスケーラブルなシステムを構築する．

3. 分析画面イメージ

WISDOM の分析画面の実現イメージを図 1 に示す．画面の上部は分析クエリの入力欄であり，その下の「検索結果」，「ページ分類」，「意見分析」，「絞込条件」は検索結果や分析方法を選択するためのタブである．検索結果タブは TSUBAKI の結果を表示し，ページ分類タブはテキスト内容によって，検索結果のページをクラスタリングして表示する[5]．

図 1.a は，発信者分類の画面イメージであり，左メニューは「個人」，「企業」，「行政」等の発信者クラスであり，選択した発信者クラスのページ情報が右側に表示される．

図 1.b は，意見分析の画面イメージであり，左メニューは評価情報をクラスタリングした分類であり，右画面は，肯定否定の割合，具体的な肯定・否定の評価文と，評価文が含まれるページ情報を表示している．また，絞込条件で，発信者クラスを「企業」等に限定することで，「企業」のみを対象とした意見分析の結果も表示可能である．

図 1.c は，各 Web ページの「ページ情報」の画面イメージであり，ページから抽出した「情報発信者」，「ページ外観」，「評価表現」の情報を表示する．

図 1.a 発信者分類の画面イメージ

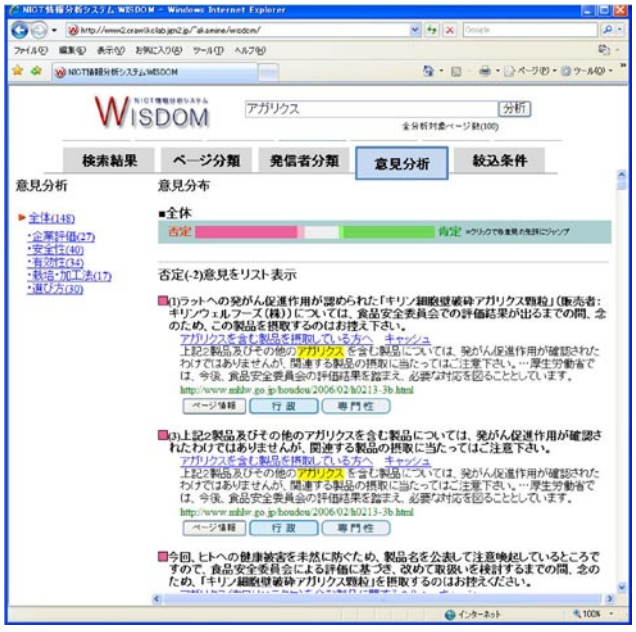

図 1.b 意見分析の画面イメージ

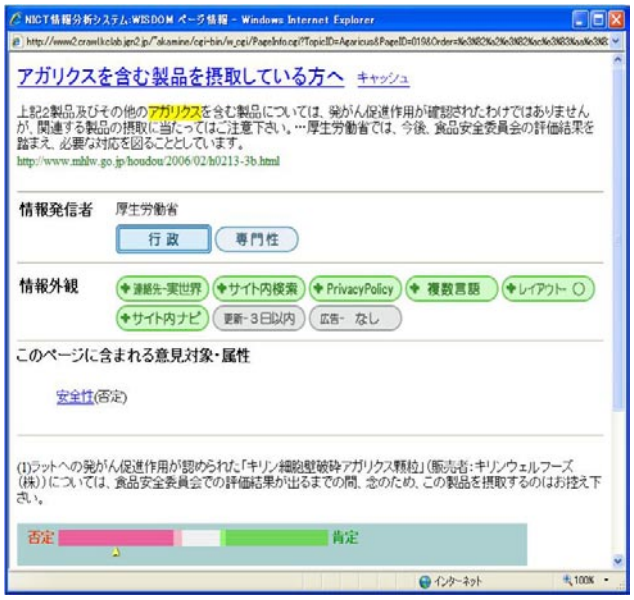

図 1.c ページ情報の画面イメージ

4. システム構成

WISDOM のシステム構成を図 2 に示す．基本的な構成は，Web 検索エンジンに各種の情報抽出・分析エンジンを追加する形となっている．また，Web 検索エンジンと同様にサーバ側で全ての処理を行い，利用者は自分の PC のブラウザ上で，分析クエリを入力し，分析結果を表示する Web アプリケーションで実現している．

4.1 ページ収集・登録部

4.1.1 クローラ

複数ノードで並列動作可能な Web クローラにより，Web ページを収集し，Web ページをローカルなディスクに保存する．現状で約 1 億ページの収集しており，更新頻度が高いニュースサイトやブログは随時収集を行っている．なお，収集の際に，今回の分析対象に含まれず，比較的量が多いアダルトページについては，キーワードによるフィルタリングを行うことで収集対象からは除外する．

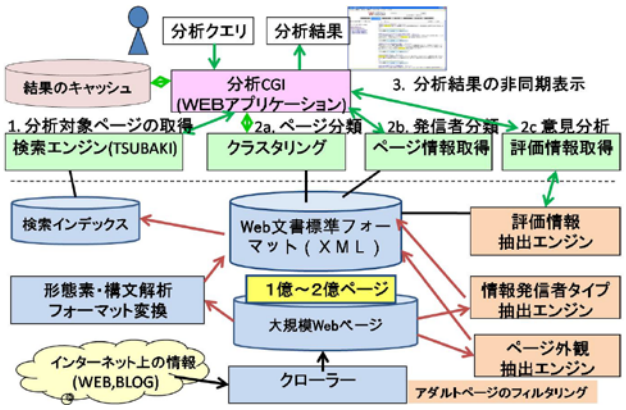

図 2. システム構成図

4.1.2 Web 文書標準フォーマットと TSUBAKI

保存した全ての Web ページのテキストに Juman による形態素解析，及び，KNP による構文解析を実施し，全ての解析結果を XML 形式の Web 文書標準フォーマット[4]に変換して保存する．また，Web 文書標準フォーマットの情報を用いて，全文検索用に検索エンジン基盤 TSUBAKI のインデックスを作成する．

4.1.3 情報抽出エンジン

分析クエリに依存せずに抽出可能な情報発信者，及び，ページ外観については，事前に Web ページから情報を抽出し，Web 標準フォーマットに格納すると共にインデックスを作成する．

4.2 分析処理部

Web 情報の分析は以下の手順で動作する．

1. 利用者が入力した分析クエリ(トピック)を TSUBAKI を用いて検索し，検索結果の上位数百ページの文書 ID 集合を取得する．
2. 数百ページの文書 ID 集合を対象として，ページ情報取得 API，クラスタリング API，評価情報取得 API を用いて，発信者分類・ページ外観用の情報、ページ分類用の情報、意見分析用の情報を並列非同期に取得する．なお，意見分析用の評価情報はトピックに依存するため，Web 文書標準フォーマットから動的に抽出する．
3. 「検索結果」，「ページ分類」，「発信者分類」，「意見分析」の各タブの情報は、非同期に作成され，表示可能になった時点で順次ブラウザでの閲覧が可能になる．つまり，利用者は，まず，検索結果を閲覧し，処理が終わった分析結果から順に閲覧することが可能である．

なお，検索インデックスや Web 文書標準フォーマットをページ数単位で分割し，複数ノード(PC)で分散して処理することで，スケーラブルなシステムを実現している．また，取得結果をキャッシュに保存することで，同一トピックに対する 2 度目の分析からは，高速な処理が可能である．

4.3 大規模並列計算機環境

WISDOM は，PC クラスタシステムを用いた大規模並列計算機環境に構築している．計算ノードは 200 台(400CPU，800CPU Core)，共用ストレージ容量は 120TB であり，最大 20Gbps の研究開発テストベッドネットワーク JGN2 に接続して運用している．

5. 議論／今後の課題

これまでに，前章のシステム構成で１億ページ強の Web ページに対して,Web 文書標準フォーマットと検索インデックスを作成した。また、検索エンジン TSUBAKI との結合，ページ分類，分析 cgi については開発を完了し，動作確認を行った．さらに，情報発信者，意見，ページ外観の各抽出・分析エンジンについては，初期版の動作を確認中である．

以下に現状システムの開発で，情報分析システムの課題として残っている点について記述する．

- 分析結果の提示方法

  現状システムでは，発信者分類と意見分析がタブとして分れており，独立に分析結果を提示する仕様となっている．絞込条件の指定によって特定の発信者のみの評価情報を分析することはできるが，例え発信者クラス(例えば,「企業」と「個人」)で肯定・否定の分布が大きく異なっていても，利用者が絞込条件を指定しなければ，その異常を発見することができない．今後，発信者と評価情報を融合させることで，異常な分布であることをシステムが利用者に提示することで利用者に「気付き」を与えるための枠組みを検討する予定である．

- 分析結果の要約方法

  現状システムを複数文書の要約と捉えた場合，同様の情報をマージする，対立・矛盾する情報を強調するという観点での設計が不十分である．類似する Web ページ／文でも，肯定否定が異なる情報や，発信者クラスが異なる情報はマージできない等の基準を作成して，実装していく予定である．

- 分析のナビゲーション方法

  利用者は，最初は分析対象として思いついたキーワードをトピックとして入力し，次により具体的なサブトピックを対象として分析するような使い方をすることが考えられる．例えば，最初は「アガリクス」というトピックで分析し，次は「アガリクスの成分」,「アガリクスが癌に効く」というアガリクスに関するサブトピックで分析するというような使い方が考えられる．この場合，利用者にサブトピックの候補を提示する等して，分析対象をトピックからサブトピックへスムーズにナビゲートする必要がある．

- 分析対象の更新

  現状，ページ収集からインデックス作成までの処理が自動化されておらず，分析対象の更新は人手で行う必要がある．今後，サイトの更新頻度に応じてサイト毎に更新頻度を可変にし，自動的に分析対象に新規 Web ページを追加する枠組を構築する予定である．

6. おわりに

本稿では,Web 情報の信頼性を検証するために,Web 情報を発信者，意見(評価情報)，ページ外観等から多面的に分析する情報分析システム WISDOM の開発方針とシステム構成について報告した．現在，各種の情報抽出・分析エンジンを改良中であり，今後，改良された情報抽出・分析エンジンを WISDOM に組み込み，本システムの有用性を評価検証する予定である．

謝辞

WISDOM の開発にあたり，クローラを開発・提供して頂いた東京大学の田浦健次朗准教授に感謝します．また，検索エンジン TSUBAKI，クラスタリング等を開発・提供して頂いた京都大学黒橋研究室の新里圭司特任助教，柴田知秀特任助教，馬場康夫氏に感謝します．また，WISDOM の開発を行った原口弘志氏，大槻直子氏，西村晃氏，Nanda Kumar 氏に感謝します．

参考文献

[1] Sadao Kurohashi: Information Credibility Criteria Project, Proceedings of the First International Symposium on Universal Communication, pp. 49-52, 2007.

[2] Hisashi Miyamori, Susumu Akamine, Yoshikiyo Kato, Ken Kaneiwa, Kaoru Sumi, Kentaro Inui, Sadao Kurohashi: Evaluation Data and Prototype System WISDOM for Information Credibility Analysis, Proc. of the 1st Workshop on Information Credibility on the Web (WICOW), pp.25-32, 2007.

[3] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, Sadao Kurohashi: TSUBAKI: An open search engine infrastructure for developing new information access methodology, Proc. of IJCNLP2008, 2008.

[4] 新里圭司，橋本力，河原大輔，黒橋禎夫，自然言語処理基盤としてのウェブ文書標準フォーマットの提案，言語処理学会第 13 回年次大会論文集，2007

[5] 馬場康夫，新里圭司，黒橋禎夫:検索エンジン基盤 TSUBAKI を用いた大規模ウェブ情報クラスタリングシステムの構築，情報処理学会研究報告 2008-NL-183, 2008

[6] 加藤義清，乾健太郎，黒橋禎夫:Web ページの情報発信者の同定とその関係の抽出， 言語処理学会第 14 回年次大会*[本論文集]*，2008

[7] 川田拓也，中川哲治，森井律子，宮森恒，赤峯享，乾健太郎，黒橋禎夫，木俵豊:Web ページの情報発信者の同定とその関係の抽出， 言語処理学会第 14 回年次大会*[本論文集]*，2008

[8] 中川哲治，宮森恒，赤峯享，乾健太郎，黒橋禎夫:Web 上の客観的記述からの評価情報抽出に関する技術的検討，言語処理学会第 14 回年次大会*[本論文集]*，2008