Web コンテンツの信頼性分析

木俵 豊[†]　赤峯 享[†]　河原 大輔[†]　加藤 義清[†]　中川 哲治[†]
黒橋 禎夫[†,‡]　中澤 聡[††]　乾 健太郎[†,‡‡]　森 辰則[†††]

[†]情報通信研究機構　[‡]京都大学
[††]日本電気株式会社　[‡‡]奈良先端科学技術大学院大学　[†††]横浜国立大学
[†]{kidawara,akamine,dk,ykato,tnaka}@nict.go.jp
[‡]kuro@i.kyoto-u.ac.jp　[††]s-nakazawa@da.jp.nec.com
[‡‡] inui@is.naist.jp　[†††]mori@forest.eis.ynu.ac.jp

1 はじめに

現在，インターネットは実社会における情報流通に必要不可欠なインフラとなっている．そのネットワーク上で流通する情報は膨大な量となっているだけでなく，加えて情報の内容・品質が多種多様となり玉石混淆のデジタルコンテンツが流通している．その結果，どのような情報がインターネット上に存在するのかを把握する事すら困難となっている．

我々はインターネットから必要な情報を探し出す際には，検索エンジンを利用するが，一般的なキーワードを入力した場合には，その検索結果が大量に出力されることが多い．その結果，専門知識を持たない一般的なユーザは検索結果の上位数件を閲覧した後に同様の内容が書かれていれば，その Web ページの情報を信用することが多く，情報の発信元やその内容の分布に充分な注意を払っていないのが現状である．

しかし，SEO（Search Engine Optimization）と呼ばれる特定の Web ページのランキングを上位にする操作によって，検索エンジンによる検索結果は恣意的に操作されている可能性も高く検索エンジンのランク順は問い合わせに対する内容の適切度を表しているとは言えなくなっている．

このような情報流通環境の変化の中で，玉石混淆となっている Web コンテンツから信頼出来る情報を見つけ出す手がかりを明らかにして，適切なものを見つけ出すための技術開発が求められている．

2 Web コンテンツの信頼性分析プロジェクト

インターネット上の玉石混淆の Web コンテンツから信頼性の高い情報を見つけるためには Web ページに記述されている内容に踏み込んだ分析が必要不可欠である．しかし，一般に使われている現在の検索エンジンでは内容まで踏み込んだ分析を行い提示するものがない．そこで情報通信研究機構 (NICT) では Web コンテンツの信頼性分析に焦点を絞った研究開発プロジェクトを 2006 年 4 月からスタートさせた．また，2007 年度から総務省においても同様の研究課題に着目した「電気通信サービスにおける情報信憑性検証技術に関する研究開発プロジェクト」がスタートした．

このような Web コンテンツの信頼性分析技術の研究開発という非常に挑戦的な課題に対して，我が国では NICT と総務省が中心となって研究開発が推進されているが，信頼性を判断するためにはどのような情報を抽出すべきであるか，また，その抽出された情報をどのように信頼性評価に結びつけるのかを明らかにすることが，両プロジェクトの大きな課題であり，また，2011 年 3 月末までの限られた研究開発期間で実際に使える技術を社会に産み出すことが求められている．

これらのプロジェクトでは，情報信頼性についての分析課題を表現する単語や文章を分析対象語あるいは文として入力すると，その課題について分析する情報信頼性分析システムの実現を目指している．具体的には従来の検索エンジンと同様に分析対象語 (文) を入力すればインターネット上から関連情報を検索するだけでなく，それぞれの内容を多面的に分析して，信頼性評価に役立つ意見，外観，内容，評判情報などを分類・提示するシステムを研究・開発している．この研究・開発では，システムが出力する分析結果によって，ユーザの信頼性判断を支援することを目的とする．つまり，開発システムが自動的に信頼性を判断するのではなく最終的な判断は人

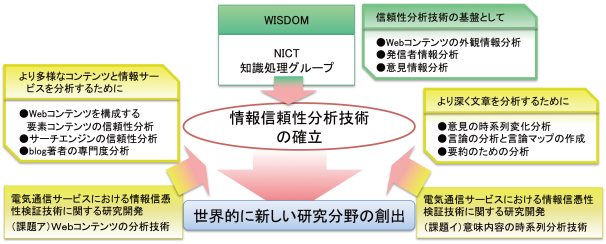

図 1: プロジェクトの相互関係

間にゆだねる事としている．

このようなユーザが情報の信頼性を判断する手がかりとなる情報を抽出するための情報分析技術においては，テキストデータとして記述された Web ページから，意味内容を抽出することが重要であり言語処理技術がコア技術として用いられる．その一方で，画像データを含む Web コンテンツの解析においてはデータ構造分析や画像処理技術なども必要不可欠であり，多様な技術を駆使する必要がある．そのため，総務省のプロジェクトにおいては，テキスト情報を対象とした言語処理技術による情報分析技術だけでなく，データ工学的・画像処理的なアプローチを駆使した分析技術なども研究開発している．図 1 にプロジェクト相互の関係について示す．

本稿では，NICT と総務省の両プロジェクトにおいて言語処理技術が果たす役割について述べる．

3 Web コンテンツの情報信頼性分析における言語処理技術

検索エンジンの結果から信頼出来る情報を簡単に見つけられない大きな原因として，情報の発信者が容易にわからないことや検索結果にどのような内容が含まれているのかを分類・整理した形で見られない事が挙げられる．例えば，ユーザが閲覧している Web コンテンツや分析したい課題に対して，どれくらいの人間が同意しているのかまた，それぞれの情報が誰の手によって記述され発信されているのかがわかりやすく提示されれば，そのユーザは情報の信頼性について評価することが容易となる．

このような情報を Web コンテンツから抽出するためには，Web のリンク構造などからランキングするような手法とは全く異なった記述内容を深く解析する手法が必要となる．つまり，Web コンテンツの信頼性をユーザが判断するためには，対象としている Web コンテンツやユーザによって与えられた分析課題に対して実社会の人々による「評判情報」や，その情報の「発信者情報」，また，与えられた課題に対する「主要意見」や「対立意見」が必要不可欠と考える．そしてそれらを文章から取り出すためには，高度な言語処理技術を駆使する必要が

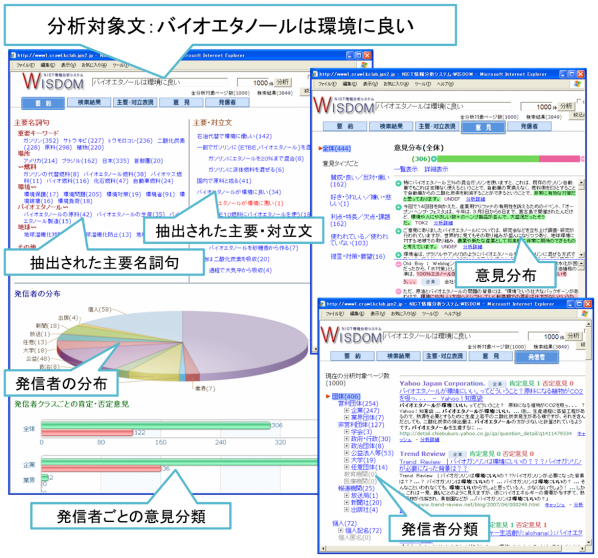

図 2: WISDOM の画面例

ある．次章にて我々が開発している情報信頼性分析エンジン WISDOM とそこで使われてい言語処理技術を用いた情報分析技術について述べる．

4 情報信頼性分析エンジン WISDOM

NICT で開発が進められている情報信頼性分析エンジン WISDOM(Web Information Sensibly and Discreetly Ordered and Marshaled)[3] は，従来の検索エンジンと同様に分析したいキーワードやキーセンテンスを入力することで，TSUBAKI[2] によって検索された Web コンテンツを対象として上記の項目について分析を行う．分析内容はタブ形式で分類されたインタフェース上に提示され検索内容とその分析結果がシームレスに表示できる．WISDOM の分析機能について以下に記述すると共に，一例として，バイオエタノールについての分析課題を「バイオエタノールは環境に良い」という分析対象文として入力した画面例を図 2 に示す．

4.1 主要・対立表現の抽出

主要表現とは，与えられた課題に対する Web ページ集合に対して，高頻度に出現する言語表現のことであり，名詞句と述語項構造（文）の 2 つのタイプを解析する．これに対して対立表現とは述語項構造の主要表現に対立，矛盾する言語表現のことを示す．WISDOM では，与えられた課題に対してこれらの表現の分析を行う [5]．具体的には WISDOM は，TSUBAKI の検索結果 1000 件に対して以下のように処理を行っている．

1. 各ページからの主要表現を抽出
   各ページから与えられた分析課題に関する 15 文程度

の重要文を選択し，複合名詞，括弧で囲まれた表現，述語項構造などを主要表現の候補として抽出する．

2. 主要表現の集約
   高頻度の表現を抽出するだけではなく，形態素解析による表記揺れの吸収を行い，国語辞典や Web 等から自動獲得した同義表現のマージ，さらには，部分全体関係にある表現のマージなどを段階的に行うことによって高精度に主要表現を集約する．
3. 対立表現の抽出
   主要表現として抽出した述語項構造について，述部がその否定となっているもの，及び反義語に置き換わっているものがあれば対立表現として抽出する．

4.2 評価情報の抽出

文書として記述された内容から意見情報を抽出し分類する技術は商品を評価する等の幅広い応用が想定されており，盛んに研究が行われている [6]．従来の研究では，商品に関する主観的な表現によって著者の意見が表明されている評価情報を対象としたものが多い．その一方で，「この食品は発がん作用を促進する.」,「買って 3 日後に壊れた」などの客観的な表現で記述されているものも数多く存在するため，従来手法では不十分であった．そこで，WISDOM では，多様な評価情報の抽出を実現するために，評価情報を付与したコーパス [8] を作成した後に，6 種類の評価表現に分類して分析を行っている．これらの分類とコーパス構築と並行して，自動分類機構の開発を進めている [1]．これは，まず文中の単語の形態素情報や評価表現辞書に登録されている単語の極性情報を素性として SVM による機械学習で評価極性の判断を行う手法である．現在までの実験によって 83%の精度が得られている．

4.3 Web ページの発信者分類と情報の発信の分類

Web ページの情報発信者とは，ページに含まれる情報の内容や，情報の公開に責任を持つ人物や個人であると考える．そのため，サイト運営者のみならず，ページ著者を考慮した分類モデルが必要となる．本研究では，Web ページ情報発信の形態を情報発信構成と名付けて，以下の５つの基本構成を定義している [4].

(a) 単一タイプ
個人管理サイトで個人が発信しているもの，組織のサイトで組織として発信しているもの等

(b) 所属タイプ
企業のサイトで経営者や社員が情報発信しているもの等

(c) 掲載タイプ
新聞社や学会などのサイトに専門家の寄稿記事が掲載されているもの等

(d) 引用タイプ
blog などで，他のサイトの情報などを引用している場合等で，引用者の責任で引用しているもの等

(e) サービスタイプ
掲示板や blog のコメント欄などのように著者やサイト運営者以外のもの等

このような定義の下で，HTML 文書構造を用いてサイト発信者情報が記述されている可能性の高い場所と固有名解析を行うことで人名，組織名の可能性のあるものを選択する．さらに選択された候補について，分析ページ中に現れる頻度や関連ページ全体に現れる頻度，出現するページの種類（分析対象ページ，トップページなど），固有名解析結果，構成形態素の情報等を用いて RankingSVM による学習を行いランキングする．

5 意味内容の時系列分析技術の開発

前章で述べたとおり WISDOM の開発を通じて主要・対立表現及び，評価表現の抽出や発信者情報の抽出・分類等の機能を開発しているが,「情報の信頼性を判断する手がかりとなる情報を提供する」といった目的にはまだまだ不十分である．たとえば，Web コンテンツに含まれる意見情報について対立情報だけでなく，根拠となる情報や上位・下位概念の情報などが信頼性を判断する大きな手がかりになる．また，分類・表示された情報が古いものであれば，役に立たない事も多い．さらに最終的に信頼性の判断をする場合には，その記述がなされている Web コンテンツの内容を把握する必要があるが，全てを読むことは大きな労力となるため，内容を要約して提示することも重要である．総務省の「電気通信サービスにおける情報信憑性検証技術に関する研究開発プロジェクト」においては，2 つの課題について研究プロジェクトを進めているが，そのひとつである「意味内容の時系列分析技術の開発プロジェクト」では，このような課題について自然言語処理技術を核とする研究開発を進めている.「意味内容の時系列分析技術の開発プロジェクト」が目指す分析システムの概念を図 3 に示し，それらの現状について以下に述べる．

5.1 論理的関係解析技術の開発

ユーザが分析課題として指定する分析対象文に対して，それと根拠・例示・詳細・矛盾などの論理的関係をなす文を Web 文書集合の中から抽出し,「言論マップ」と呼ばれる言論間の関係グラフを自動生成する技術の研究開発を進めている [9]．これまでに，約 130 万対の上位下位関係や 4.6 万対の事象間関係などの知識ベースを整備

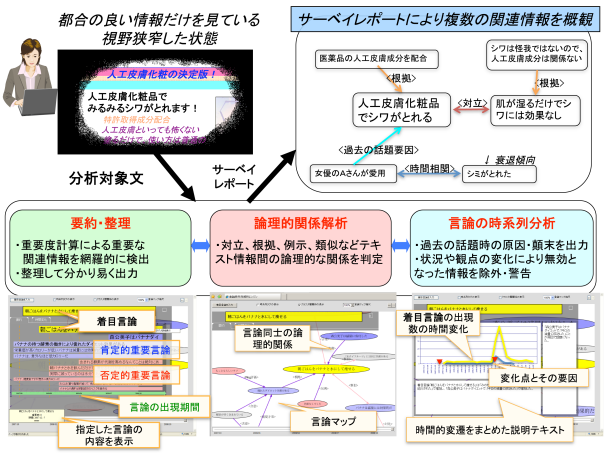

図 3: 「意味内容の時系列分析技術の開発プロジェクト」の研究・開発項目

するとともに，モダリティや否定情報を付与した約 4000 文のコーパスを作成し，また上述のような論理的関係を付与したコーパスの構築にも着手している．

5.2 Web コンテンツの要約・整理技術の開発

Web コンテンツに含まれる文書情報から抽出した重要な言論やその論理関係を含んだパッセージを骨組みとして，TextRank による要約を行う技術を開発している．現在は「裁判員制度」等の 3 トピックにおいて 80%程度以上の抽出精度を実現している [7].

5.3 時系列分析技術の開発

着目した言論が含まれる Web コンテンツ数が時間と共にどのように変化するのかをトラッキングすることで，着目した言論の変化点やその出現数の時間変化に影響している関連言論を抽出すると共に，求めた関連言論と言論マップで抽出された言論の時間的推移からそれらの伝搬フェーズを推定し，有効期限等を判定する技術の開発を進めている．現在は，複数のキーワードを用いて簡易的に表現された言論に対して，この手法の有効性の確認を行っている [10].

6 おわりに

総務省と NICT では，情報信頼性分析技術の開発という新しい分野の技術開発に焦点を絞り，情報爆発時代における玉石混淆の情報資源の中から信頼性と価値の高い情報を見つけるための情報分析技術について研究・開発を進めている．紙面の都合上，本稿では自然言語言語処理技術を用いた NICT の研究開発プロジェクトと総務省プロジェクトについて記述したが，検索エンジンや Web サービスの信頼性，また，言語処理技術だけではカバーできない画像を含む Web ページの信頼性分析などの技術開発についても，「電気通信サービスにおける情報信憑性検証技術に関する研究開発プロジェクト」の中で「Web コンテンツの分析技術プロジェクト」として行っている．

Web コンテンツの信頼性を分析するための情報分析技術は，多方面にわたる技術を用いる必要があるが，テキストとして記述されている内容を解析するためには言語処理技術が果たす役割が大きく，新たな技術開発が必要である．今後も産学官連携体制の下で研究・開発を進めていく予定である．

参考文献

[1] Tetsuji Nakagawa, Takuya Kawada, Kentaro Inui, and SadaoKurohashi. Extracting subjective and objective evaluative expressions from the web. In *Proceedings of the 2nd International Symposium on Universal Communication*, 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. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.

[3] S.Kurohashi. Information credibility criteria project. In *Proceedings of the First International Symposium on Universal Communication*, pp. 49–52, 2007.

[4] Y.Kato, D.Kawahara, K.Inui, S.Kurohashi, and T.Shibata. Extracting the author of web pages. In *in Proceedings of Second Workshop on Information Credibility on the Web (WICOW08)*, 2008.

[5] 河原大輔, 黒橋禎夫, 乾健太郎. 主要・対立表現の俯瞰的把握 - ウェブの情報信頼性分析に向けて. 情報処理学会自然言語処理研究会 NL-186, pp. 49–54, 2008.

[6] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[7] 渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 鈴木貴子, 森辰則. 情報信憑性判断のための要約に関する基礎的検討. 言語処理学会第 15 回年次大会,p.1-4, 2009.

[8] 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 テキストにおける評価情報の整理・分類およびタグ付きコーパスの構築. 言語処理学会第 14 回年次大会, pp. 524–527, 2008.

[9] 村上浩司, 松吉俊, 隅田飛鳥, 森田啓, 佐尾ちとせ, 増田祥子, 松本裕治, 乾健太郎. 言論マップ生成課題：言説間の類似・対立の構造を捉えるために. 情報処理学会研究報告, 自然言語処理研究会, 2008-NL-186, pp. 55–60, 2008.

[10] 中澤聡, 岡嶋穣, 大西貴士, 河合剛巨, 安藤真一. 時系列分析による web 文書の情報信頼性判断支援. 言語処理学会第 15 回年次大会, 2009.