5-3　情報分析システム WISDOM の開発

5-3 Development of the Information Analysis System WISDOM

木俵　豊
KIDAWARA Yutaka

要旨

第 2 期中期計画において知識創成コミュニケーション研究センター知識処理グループの研究成果として開発された情報分析システム WISDOM は、大規模な Web 情報から信頼性の高い情報を発見するためのシステムである。この WISDOM は自然言語処理技術、情報検索技術、機械学習技術、大規模データベース技術、並列計算機技術などが高度に融合して実現しており、Web 情報の発信者や評価表現をはじめとする内容を分析・分類して提示するシステムである。本稿では、WISDOM の機能などについて概要を記述する。

NICT Knowledge Clustered Group researched and developed the information analysis system “WISDOM” as a research result of the second medium-term plan. WISDOM has functions that users find high-credible information from huge amount of Web pages. WISDOM is the comprehensive and integrated system based on Natural Language Processing (NLP), Information Retrieval (IR), Machine Learning (ML), Database (DB) and High Performance Computing (HPC) Technology. The system has processing capability of Web information analysis, publisher detection, reputation information extraction, display all the processing result within proper category. The paper describes overview of WISDOM.

[キーワード]
自然言語処理，情報分析，情報検索，大規模情報管理，ビッグデータ
Natural language processing, Information analysis, Information retrieval, Huge data management, Big data

1　まえがき

インターネットによる情報流通によって、多様な情報が我々の生活に大きな影響を与えるようになっている。しかし、インターネットが構築され、ブロードバンド化が進み、パソコンだけでない多様な端末で利用されるユビキタス時代を経て、インターネットに流れる情報は大きく変わってきた。当初、ある程度専門的な知識がなければ情報発信できず、一般的なユーザは情報を閲覧する情報の消費者であったが、インターネットのブロードバンド化やユビキタス化が進むにつれて、専門的な知識がないユーザでも簡単に多様な情報が発信できるようになっている。第 2 期中期計画の 2006 年から 2011 年を振り返ると、CGM（Consumer Generating Media）と呼ばれるそれまで情報の消費者であった一般ユーザがパソコンや携帯電話で気軽に情報発信ができるようになった時代であった。このような環境の変化は Web 2.0 と呼ばれ、インターネット上に蓄積され、流通する情報の量が爆発的に増加していた情報爆発時代の始まりであった。また、今では一般的に使われている「クラウドコンピューティング」という言葉が生まれ、ネットワークで大量のデータを処理する時代の始まりであり、インターネットのブロードバンド化とモバイル端末などを用いたユビキタス化の結果始まる新たな時代の幕開けであったといえる。

このような変化は、情報の「質」に対しても大きな変化をもたらした。一般ユーザが手軽に発信できる環境は、情報の多様化につながったが、必ずしも良い面ばかりではなく信頼性の不確かな情

報も大量に生み出され、質の高い情報を見つけ出すことがきわめて困難になった。通常の検索エンジンにおいても検索結果が数百万件を超えることも珍しいことではなく、ユーザが情報の全体を把握することは到底不可能であるにも関わらず、情報の質を判断するのはユーザの責任であるため、しばしば誤った内容の情報で混乱することも発生するようになっていた。

我々はこのような時代を第1期中期計画の最終年度である2005年に予見しており、「情報の信頼性」という課題にどのように取り組むべきなのかをメディアインタラクショングループで議論しており、第2期中期計画時の重要テーマとして設定された。その課題は第2期中期計画の知識処理グループに引き継がれ、情報分析エンジンWISDOMの開発に取り組んだ。

本稿では、**2**において情報分析エンジンWISDOMの構成について記述し、**3**においてWISDOMを支える技術について紹介する。**4**において、WISDOMの利用例を示すと共に、**5**では関連技術について紹介する。**6**はまとめである。

2 情報分析エンジンWISDOM

2.1 情報信頼性分析支援

Webに蓄積された情報の信頼性は、ユーザの視点によって大きくことなることがあり、自動的に判断することは容易ではない。WISDOMでは、信頼性の判断はユーザに委ねることとして、その判断をサポートするために分析対象とする課題についての背景的知識、事実、論点・対立点、意見分布などを的確に提示することを目的としている。そのためには、文や文章の構造を分析し、その性質や関係を抽出した上で同じ意味の別表現や表現の多義性などを分析・表示する必要がある。さらには、信頼性を判断するためには重要な手がかりとなる情報の発信者やその発信者が所属する組織の専門性なども表示する必要があり、人名や組織名などの固有表現認識に加えて文書の総合的解析が必要不可欠となる。図1にWISDOMによって実現を目指した情報信頼性支援による意思決定手順を示す。

このような意思決定支援を実現するために、WISDOMの開発においては、コア技術として自然言語処理技術を位置づけると共に、リンク解析技術なども含めてユーザの信頼性支援を目的として、以下の評価軸を設定した。

1. 情報内容の信頼性
2. 情報発信者の信頼性
3. 情報外観の信頼性

WISDOMは、これらの観点において情報を分析・提示するように設計されている。これを実現

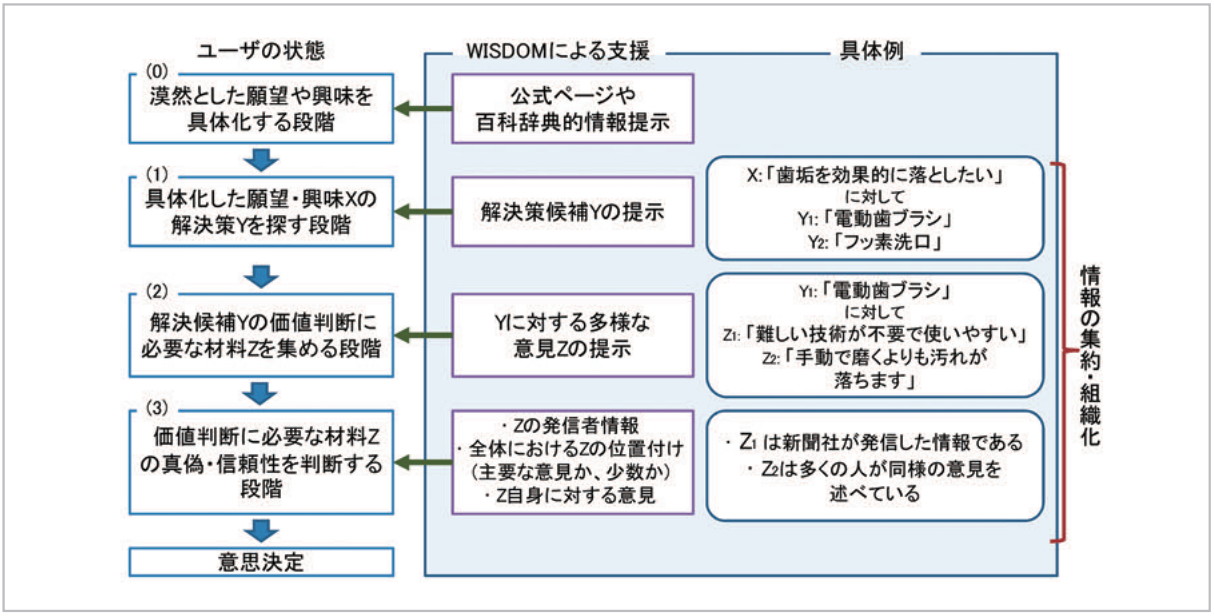

図1 意思決定の過程とWISDOMによる支援

するためには、大量の Web 情報の収集・集積・管理、Web 情報に含まれる文情報、構造情報の分析、Web のリンク情報の分析などにおいて最先端の技術が必要不可欠となる。

2.2 WISDOM の構成

情報分析エンジン WISDOM は、図 2 に示すとおり大きく分けて情報分析基盤部、情報分析エンジン部、フロントエンド部の 3 つに分類される。これらの詳細については **3** にて紹介する。

3 WISDOM を支える技術

3.1 情報分析基盤を構成する技術

情報分析基盤においては、大規模な情報を適切かつ高速に収集・集積した上で高速にアクセスするための管理機構を実現している。

3.1.1 クローラ

クローラは Web 情報を取得するだけであり、技術的な要素はないと思われる傾向があるが、実際には大規模な情報源に対して過大な付加を与えないように適切に取得する必要がある。また、相手先によって情報の更新頻度は異なるためスケジューリングなどにおいても十分に考慮しなければならない。WISDOM のクローラは、一般的なクローラ以外に特定の URL を基点として同一ドメインのリンクをたどりながら未取得のページを収集する深度クローラ、RSS フィードを取得して、フィードされている未習得の URL についてページを収集する RSS フィードクローラから構成される。このクローラによって、WISDOM は一日あたり約 1,000 万 Web ページを収集している。この取得ページの比率は、更新された Web ページが約 72 %、新規ページが約 27 % であり、残り 1% の Web ページは、深度クローラと RSS フィードによって収集される。運用にあたっては、利用可能な帯域（100 Mbps）を考慮して、4 並列でページ収集を行っている。

3.1.2 データプール

クロールデータプール

クロールされた Web 文書に対しては、URL 文字列フィルタや robot.txt フィルタ、content-type フィルタ、言語フィルタ、辞書フィルタ等で処理を行い、次回のクローリングや後処理としての分析を行うための各種情報を出力する。さらに、このような各種情報やページデータを圧縮した圧縮ページファイル、リンクを圧縮した圧縮情報ファイルをクロールデータプールに登録する。これらのデータは、次回のクローリング時の情報として用いられたり、情報分析を行うための元データとして利用される。

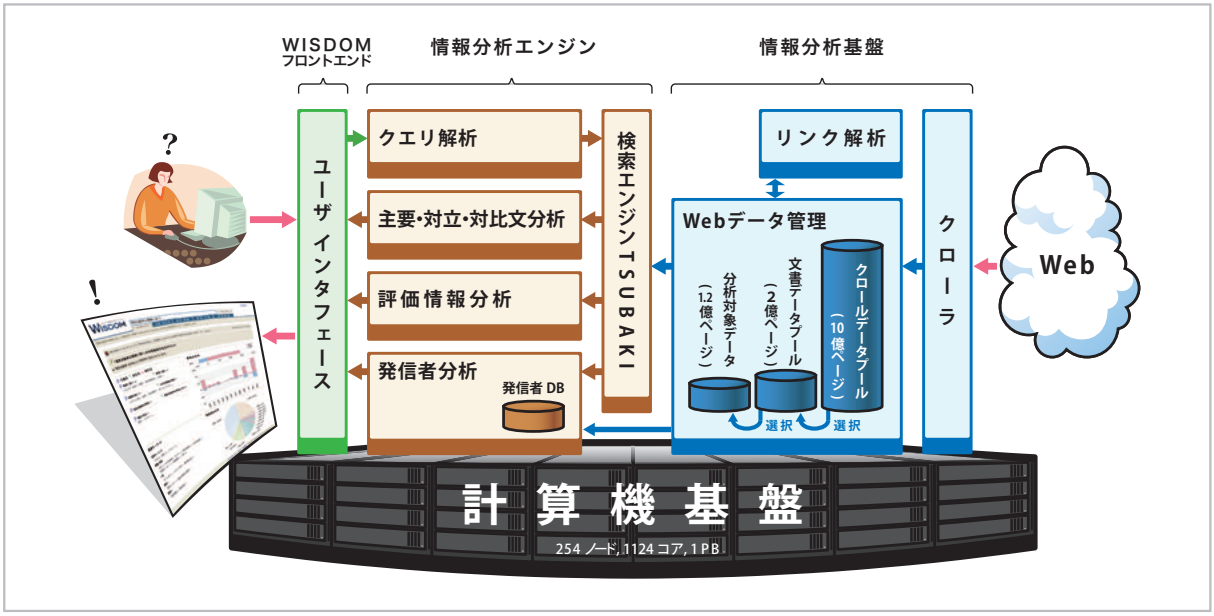

図 2 情報分析システム WISDOM の全体構成

文書データプール

クロールデータプールに格納された取得したWebページから、テキスト情報の内容やリンク情報などの解析によってSPAMページの可能性の高いものなどを排除し、さらに外観分析情報等の解析を経て、分析対象にするページを2億ページ選別している。さらに、そのページを対象として解析を行い、そのページの特徴を表すXMLで記述されたWeb標準フォーマットを作成する。Web標準フォーマットにおいては、対象ページのリンク情報や、文書ID情報、文として抽出されたテキスト列に対して構文解析などを行った結果が記述されており、後述の多様な情報分析手法や検索エンジンの情報として利用される。

3.1.3 オフライン情報分析

リンク解析

分析対象となるWebページを選択する際には、Webスパムと呼ばれる無意味なページを発見して排除することが重要である。Webスパムは、コンテンツスパム、リンクファーム、なりすましという3種類に分類される。コンテンツスパムは、隠しテキストや超微細テキスト、単語の羅列、タイトルと内容の異なるものなどWebページに無意味な情報を潜ませて検索エンジンのランキングに影響を与えるものである。検索エンジンのランキングにはリンク情報が活用されていることを利用したリンクファームというものがある。これは、リンクページを大量に生成し、リンクの価値を高めようとするものである。なりすましとは、クローラとWebブラウザのエージェントに応じて異なるコンテンツを提供するものであり、クローリングしたキャッシュページと実ページが異なるものである。WISDOMにおいては、リンク構造に基づいたWebスパム抽出を行っている。これは、Webのリンク構造を大規模なグラフとして表現し、強連結成分を抽出して図3のように蝶ネクタイの構造を得るものである。このアルゴリズムを用いて、高密なサブグラフを抽出した後にSVMによるサブグラフのスパム判定を行って、ホスト単位の推定結果を集約する。さらにホストグラフでのトラストとアンチトラストを連鎖する偏向ページランクによってスパムを発見するアルゴリズムを開発している。

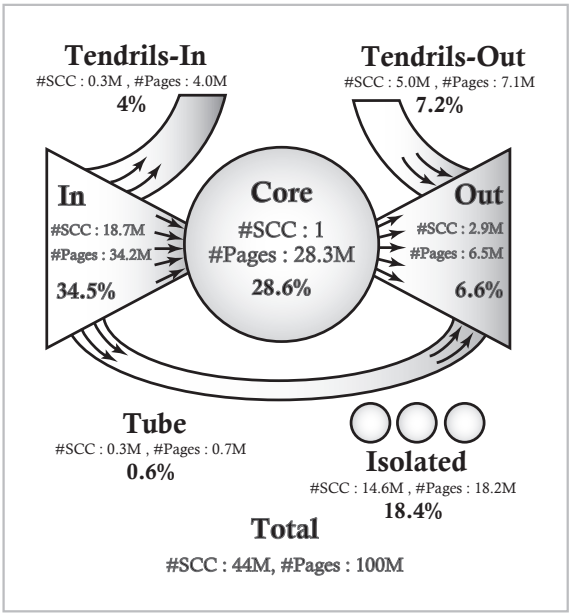

図3 Webリンクの蝶ネクタイ構造

外観分析

Webページには構造があり、その知識を持つユーザが作成したWebページでは発信する情報がその構造に基づいて整理されて記述されている。一方、スパムページなどにおいては、その構造が持つ意味と中身のコンテンツが一致していないこともあるため、外観的特徴として不整合がある場合も多い。さらには企業などWebページにおいてサイトポリシーや連絡先などの必要不可欠であるページについては、情報の中身についても精査されているか疑問が生じる場合もある。このようにページの外観からあるべき情報や構造と内容の一致度などが、そのWebページの信頼度に大きく関わってくるため、WISDOMではWebページの構造解析の後に、どのような情報が記述されているを解析し、あるべき情報や記述すべき場所などの分析を行っている。

発信者分析

情報の信頼度を判断するためには、その情報の発信者の情報が非常に重要な要素となる。専門家が発信している情報と明らかに素人が発信している情報では、その情報の根拠が大きく異なる。Web情報は、その情報の中身を理解すれば発信者が明確であるのか匿名であるのか等がわかる場合が多い。従って、発信者や情報の著者を同定す

ることは、情報抽出のタスクであると見なすことができる。WISDOMの開発においては、サイトに依存しない性質として、Webページの主要部と情報発信者名の出現位置の関係に着目し、それを利用した発信者同定の手法を開発している。

WISDOMにおいては、Webページの情報の内容および、その公開について責任を有する人物や団体などを含む実態を発信者と定義して、サイト運営者と著者に分類している。さらに情報発信者クラスとして6種類に分類して、各Webページの発信者情報の分析結果を整理している。

発信者の同定については発信者の情報が含まれていると思われる抽出対象ページ領域の選択を行った後、抽出対象文の選択を行い、情報発信者候補の抽出を行う。これらの作業を実現するためにWebの構造を解析した後にサイト運営者の情報については、情報発信者がよく現れるページ領域としてページの上部のバナーや下部の著作権表示の中にサイト運営者の情報が現れやすいとして重点的に解析を行う。また、本文中にも発信者情報が含まれているため、発信者を表す文に含まれる助詞の中で「の」以外の助詞が使われている割合が低い。また、形態素解析の結果、人名や組織名、組織名末尾、未定義語が含まれる可能性が高いなどを考慮して対象文として可能性の高いものを抽出する。このようにして抽出した文について、1）情報源全体における出現頻度、2）候補が出現するページの頻度、3）候補が出現する文書の種類、4）構成語の品詞属性、5）先頭形態素・末尾形態素、6）形態素数、7）ページ内位置、8）著作権表示由来か否か、等を組成として機械学習によって分類している。

3.2 情報分析エンジン

クエリ解析

各情報分析機能は、WISDOMに入力されたクエリ解析によって必要な情報が渡される。クエリは名詞列や単語列もしくは自然文によって入力されることを想定しており、その解析によって何に対してWISDOMの分析を行うかを決めるため、WISDOMにおいて重要な処理の1つとなっている。入力されたクエリはトピックとサブトピックに分類され、評価表現分析にはトピックとサブトピックが渡され、主要対立・対比分析にはトピックが渡される。このトピックとサブトピックの抽出には構文解析器KNP*が使用されている。

主要対立・対比分析

クエリ解析によって抽出されたトピックに関する関連キーワードおよび主要・対立・対比文を対象となるWebページ集合から抽出する。関連キーワードおよび主要文とは、対象のWebページ集合上で高頻度に出現する言語表現のことであり、それぞれ名詞句と述語項構造（文）が対象となる。対立文とは、主要文に対立・矛盾する文であり、対比文とは主要文に対して対比されている文を示す。これらの分析・抽出を実現するために述語項構造を抽出する。述語項構造とは、述語1つと、それに係る1つ以上の自立語列を抽出した項からなるものである。このような述語項構造を抽出した後、同義の述語項構造や包含関係を分析した後に集約する。さらに主要・対立・対比の分類を行うために否定フラグの反転や述語の反意語への反転などを行ったうえで、述語項構造集合を再検索して発見する。

評価情報分析

クエリ解析によって得られたトピックに関する肯定的・否定的な意見や評価をWeb文書から自動的に抽出・分類して出力するものである。WISDOMでは「感情」「批評」「メリット」「採否」「でき事」「当為」「要望」の7種類の評価情報に分類して分析を行っている。これらの分類を行うために100個のトピックを選択し、収集したWeb情報から1トピックあたり200文について評価情報を人が評価した上でタグ情報として付与し、2,000文の評価情報タグ付きコーパスを作成し、機械学習のための教師データとした。そして、そのコーパスを用いて、pairwise法を用いて多値分類に拡張したSVMによる分類を行う。まず、SVMを用いて与えられた評価表現がトピックに関連するかしないかを判定する2値分類器を学習させた後に、得られた評価表現の事例をその分類器で分類し、その際の分離平面からの距離を関連度として出力して、関連度の高いものを評価情報として出力する。

＊ http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

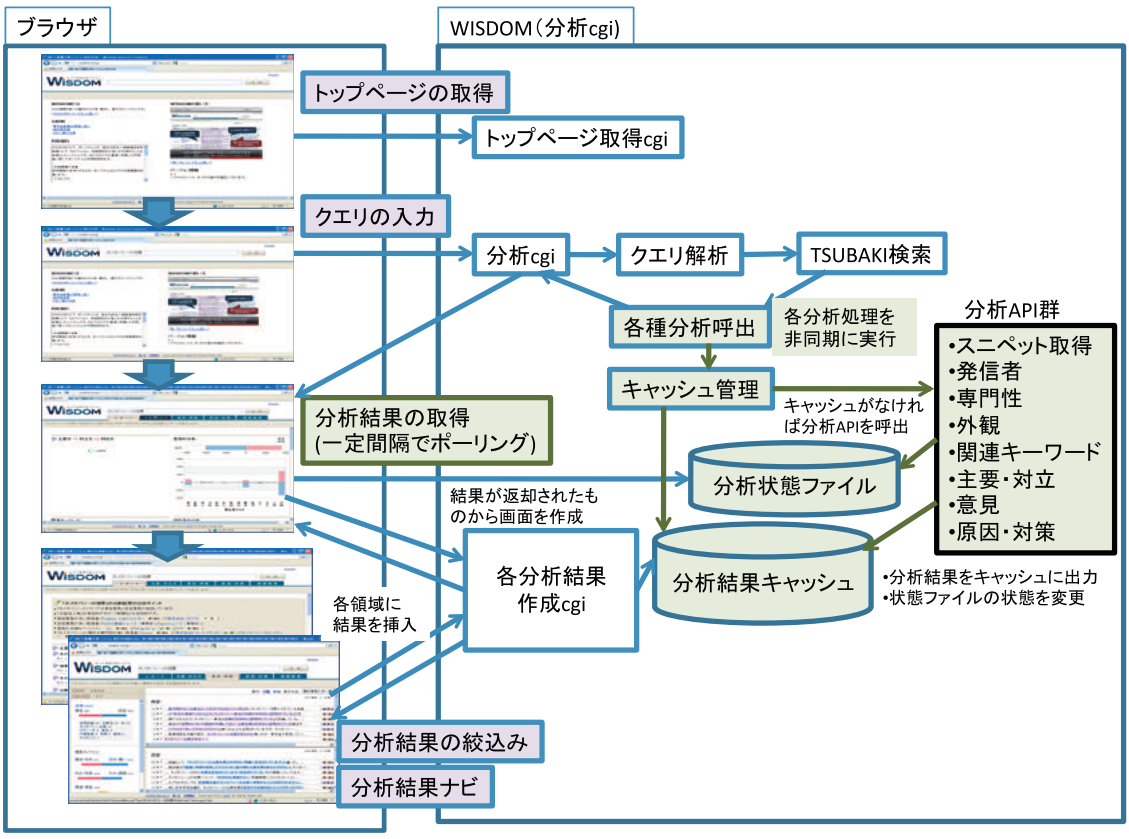

図 4 WISDOM の処理フロー

3.3 WISDOM フロントエンド

ユーザインタフェース

WISDOM はブラウザを介して利用される。大規模な情報の集約や分析はサーバによって行われる。クエリの入力や分析機能の切り替えはブラウザ上のクリックやタブの切り替えなどによって実現される。図 4 に処理フローを示す。

4 WISDOM による分析

WISDOM においては、Web インタフェース上のタブで各機能を切り替えてこれまでに述べた各種の分析結果を表示することで結果を評価することができる。WISDOM の最も特徴的なインタフェースがレポート画面である。図 5 に「電動歯ブラシは歯に良い」という分析対象文に対する出力結果（レポート）を示す。このページには、分析結果の注目ポイントや関連キーワード、発信者分布などがまとめて表示されており、分析対象について概観できる。

5 関連研究

本研究は、情報の信頼性分析という非常に難しいテーマに取り組むことを目的としていた。計算機の出力をユーザがどのように受け取るかということに関しては、Fogg ら [1] の研究によって、仮定された信頼性（presumed credibility）、評判に基づく信頼性（reputed credibility）、表面的な信頼性（surface credibility）、経験に基づく信頼性（experience credibility）の 4 つに分類できるとしている。Fogg ら [2] では、これらの概念をさらに整理した上で、情報の信頼性は主として“trustworthness”と“expertise”を元に判断されるとしている。Rieh ら [3] は大学生を対象に信頼性判断の認知的なプロセスと戦略について考察を行っている。これによると、人間が情報の信頼性を判断するプロセスには予測的判断

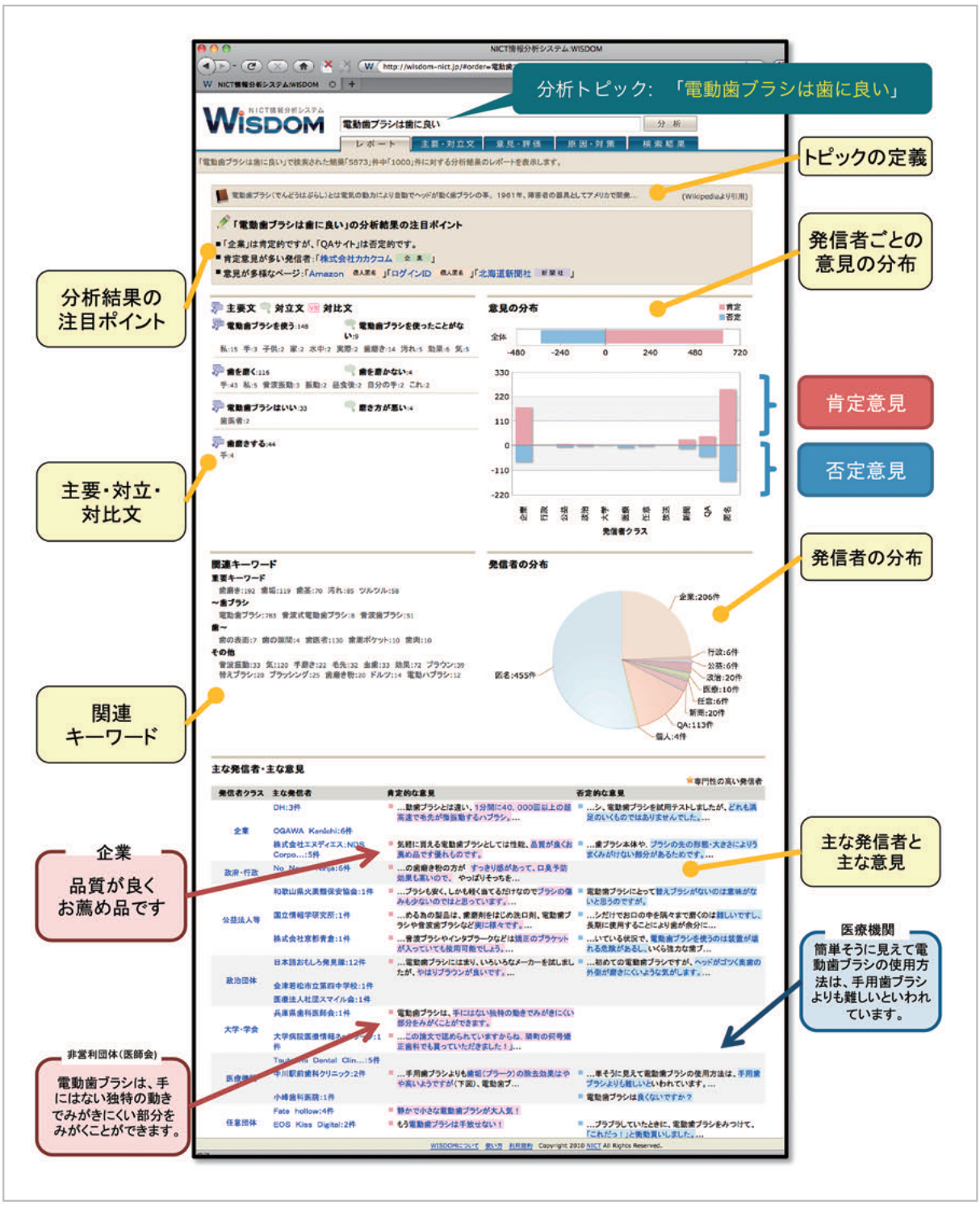

図５　WISDOM の利用例

(predictive judgement) と評価的判断 (evaluative judgement) の２種類があるとしている。そして信頼性判断は予測的判断と評価的判断を繰り返して判断する過程であるとしており、情報の信頼性判断は批判的思考も含めて複雑な認知的営みであることを指摘している。こうした調査や分析のいずれもが示唆することは、情報の信頼性がさまざまな要因の組み合わせからなる複合的な問題であるということであり、情報の信頼性は情報の真偽や正確さと等価ではないということがいえ

る。WISDOMでは、このような報告を元に（1）発信者に基づく信頼性、（2）情報の外観的特徴に基づく信頼性、（3）情報の評判に基づく信頼性、（4）情報の意味内容に基づく信頼性という視点で分析を行えるように設計が行われている。Webを対象とした専門家検索としては、CastilloらのWikipediaを利用した専門家検索の提案[6]やセマンティックWeb的なアプローチの提案[7]などがある。また、中島ら[8]は、特定の分野の熟知度に基づいてブログをランキングする手法を提案しているが、本研究の手法は一般のWebページを対象としており、辞書を用意する必要がない点で優れている。評判情報等の抽出においては、小林ら[9]や宮崎ら[10]がレビュー記事やブログ記事から商品に関する評価情報を抽出する手法を提案している。本研究での手法は「商品Xは購入後三日後に壊れた。」等の客観的な記述に含意される評価情報の抽出を視野に入れているところが異なる。

6 まとめ

本稿では、WISDOMの概要について述べた。WISDOMは、自然言語処理技術や情報検索技術、機械学習技術、さらには大規模情報管理技術、並列処理技術など非常に高度な情報処理技術が融合的に機能している。我が国の大学や研究機関において、これほど大規模かつ定常的に運用できるシステムは他になく、また、情報の信頼性という視点に注目していち早く研究を開始したという点においても世界的にユニークなものである。第2期中計画が開始された2006年に比べると、現在はさらに爆発的に情報が増加しており、さらに価値のある情報を見つけることが重要となっている。

2011年3月11日に発生した東日本大震災においては、FacebookやTwitter等のSNSの情報の価値が広く認識された。このような情報の多様化や大規模化がますます進む中で、第1期中期計画のメディアインタラクショングループから始まり、第2期中計画に知識処理グループにおいて、「情報の信頼性評価に関する基盤技術の研究開発」プロジェクト（通称情報信頼性プロジェクト）として研究が実施され、第3期中期計画においてはユニバーサルコミュニケーション研究所に情報分析研究室と情報利活用基盤研究室が設立され、本分野の研究を加速させながら多くの成果が生み出されつつある。このような体制の中で、第3期中期計画の終了年度においては、機能を一新したWISDOM2015を公開すべく研究開発に取り組んでいる。なお、紙面の都合上、本稿はWISDOMのごく一部の機能紹介をするにとどまった。詳細については文献[11]を参照いただきたい。

謝辞

WISDOMの開発において、知識処理グループの客員研究員としてプロジェクトに参加いただいた京都大学　黒橋禎夫教授、東北大学　乾健太郎教授、情報信頼性プロジェクトメンバ、旧知識処理グループの皆様に感謝します。

参考文献

1 Fogg, B. J. and Tseng, H.,“The Elements of Computer Credibility,”Proceedings of the SIGCHI conference on human factors in computing systems, ACM Press, pp. 80–87, 1999.

2 Fogg, B., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. et al.,“What makes Web sites credible?: a report on a large quantitative study,”Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 61–68, 2001.

3 Rieh, S. and Hilligoss, B.,“College Students' Credibility Judgments in the Information-Seeking Process,”The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, pp. 49–71, 2007.

4 Demartini, G.,“Finding Experts Using Wikipedia,”Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07), pp. 33–41, 2007.

5 Jung, H., Lee, M., Kang, I.-S., Lee, S.-W. and Sung, W.-K.,“Finding Topic-centric Identified Experts based on Full Text Analysis,”Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07), 2007.
6 C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna,“A reference collection for web spam,”SIGIR Forum, 40(2): pp. 11–24, December 2006.
7 C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri,“Know your neighbors: web spam detection using the web topology,”In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 423–430, 2007.
8 中島伸介，稲垣陽一，草野奉章，“高信頼性情報の提示を目指した熟知度に基づくブログランキング方式の提案，”日本データベース学会論文誌，Vol. 7, No. 1, pp. 257–262, 2008.
9 小林のぞみ，乾健太郎，松本裕治，“意見情報の抽出／構造化のタスク仕様に関する考察，情報処理学会研究報告，”自然言語処理研究会報告，Vol. 2006, No. 1, pp. 111–118, 2006.
10 宮崎林太郎，森辰則，“製品レビュー文に基づく評判情報コーパスの作成とその特徴の分析，”情報処理学会研究報告 2008-NL-187, Vol. 15, pp. 99–106, 2008.
11 独立行政法人情報通信研究機構知識処理グループ情報信頼性プロジェクト，“情報分析システム WISDOM—Web の健全な利活用を目指して—，”ISBN 978-4-904020-01-2

（平成 24 年 6 月 14 日 採録）

木俵　豊（きだわら　ゆたか）
ユニバーサルコミュニケーション研究所
研究所長
博士（工学）
デジタルコンテンツ管理、ユビキタスコンピューティング、情報検索、情報分析
kidawara@nict.go.jp