時系列分析によるＷｅｂ文書の情報信頼性判断支援：全体概要

中澤 聡　岡嶋 穣　大西 貴士　河合 剛巨　安藤 真一
NEC 共通基盤ソフトウェア研究所
{s-nakazawa@da, y-okajima@bu, t-onishi@bq, t-kawai@bx, s-ando@cw}.jp.nec.com

1. はじめに

近年、Ｗｅｂ上の文書は爆発的に増大しており、日常生活においても欠くことの出来ない情報源として利用されている。しかし、Ｗｅｂ上には有益な情報だけでなく、誤りや偏った情報、既に無効となった情報が混在しており、従来の情報検索の手法ではこうした信頼性の低い情報を的確に判断することは難しい。こうした問題に対し、Web 情報の信頼性を分析する研究がいくつか行われている[1],[2]。

我々は、横浜国立大学、NAIST と共同で、信頼性を判断したい単文レベルのテキスト情報（言論）と、論理的、または、時間的に関連性の高い言論をＷｅｂ上から抽出・集積し、集めた言論の関係をまとめて提示することで、利用者の信頼性判断を支援する研究プロジェクトを 2008 年にスタートさせた（研究期間は３年間で、2011 年３月末終了予定）。

本稿では、この研究プロジェクトで取り組む情報信頼性判断支援の課題と、それに対するプロジェクト全体のアプローチを２節で簡単に説明する。３節では試験的に設計した情報信頼性判断支援システムのモックアップについて紹介し、ついで４節にて我々が特に注力する時系列分析について想定する効果や計画について述べる。最後に５節で、まとめとする。

2. 課題とアプローチ

まず信頼性判断にどのような課題があるのかについて簡単に考察する。信頼性が低い情報には様々な種類のものが存在する。完全に根も葉もないでたらめから、本来別の対象や領域に対する記述としては正しかったものが、悪意や過度の期待により、他の対象・領域に拡大適用されたもの、定性的には間違っていないが効果の程度など定量的な性質が過大に謳われているもの、伝搬過程で単なる推測が事実に置き換えられたもの、などである。また、それぞれ事実ではあっても、立場や状況、解釈などによって一見相反する記述となる情報もあり得る。例えば、

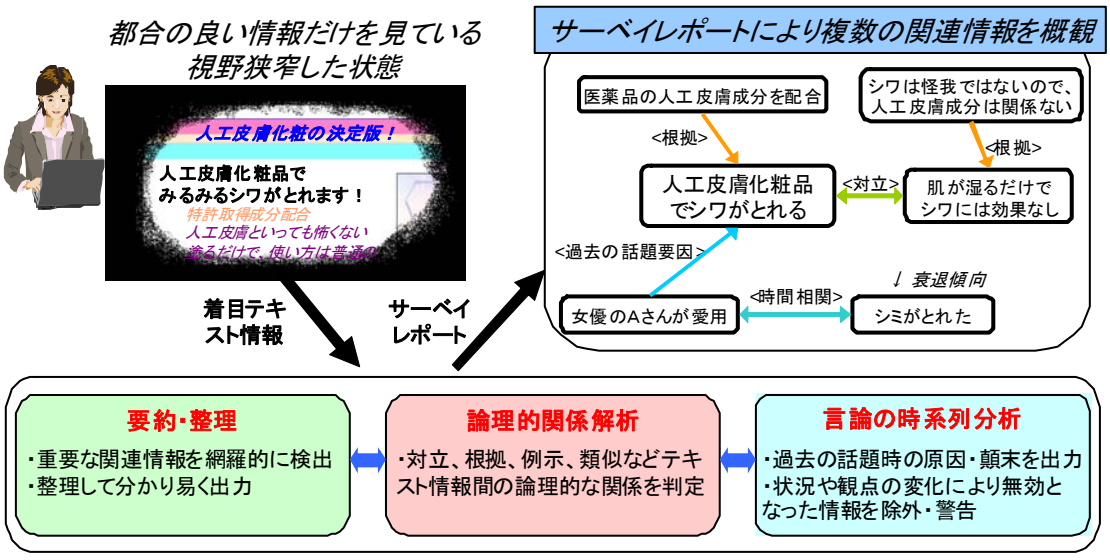

図１．信頼性判断支援イメージ

「抗生物質が必要」であるかどうかを判断するため、それに関する記述を調べたものとする。「手術後には抗生物質が必要」、「抗生物質の使用により、恐ろしい耐性菌が発生するため、軽い風邪などには抗生物質を使うべきでない」「最近、未就学児に強い副作用が現れることが発表されたので、抗生物質ｘの投与は奨めない」などの記述は、それぞれ状況や対象に制約がかされている事実であるが、制約以外の部分だけ見ると、相互に対立している記述に見えてしまう。さらには、インターネット上には、「奇跡の還元水を飲めば、体に悪い菌がつきにくい体質になるので、抗生物質は不要です」のような意見も存在する。

故に「日本の首都は東京である」のような真偽が一意に定まる情報に対しては、人気度や多数決に基づく手法が有効だが、上述したような、発言者の立場、状況、時期に応じて多様な意見や主張が存在する情報に対して、混在する大量の情報の中から信頼性あるいは有用性の高い情報を的確に判断するには、「中立の立場から広い視点で、各情報の主張を自分が採用すべきかどうか、論理的かつ構造的に考え、確認しながら思考を進めていく」クリティカルシンキングの手法が必要となる。

しかし、一般の情報利用者がこうした手法による思考を行うことは難しい。そこで本研究プロジェクト全体では「利用者が着目するテキスト情報に対して、複数の情報源から、対立する情報や過去の経緯など関連する情報を網羅的に取得、各情報の根拠と供にサーベイレポートとして要約・整理して提示」することで、クリティカルシンキングが要求される利用者の適切な情報信頼性判断を支援するアプローチをとる。図１はそうした支援のイメージを示したものである。

図１にあるような支援を実現するためには、まず利用者が着目する信頼性を判断したいテキスト情報に対して、その情報に論理的に関係する(対立、根拠、例示、類似)情報を、Web 上のテキストから判定･抽出する技術(論理的関係解析)が必要となる。そうした論理的に関係する情報を元の着目情報と合わせて関係と供に提示することで、利用者はその真偽や有効性を判定するためのてがかりとして活用できる。また、情報と情報の関係の中には、ある時期・状況下にのみ成立したものもある。さらに情報の真偽は時間と供に変化する。そこで、テキスト情報の時間的な変遷をトラッキングし、時間的に相関性の高い情報を検出し、さらにそれらの関係を分析することで、無効となった情報を判定する時系列分析技術が必要となる。また広い中立の立場から情報の論理的関係を解析するために、そうした論理的関係が記述されている重要なパッセージを大量のＷｅｂテキストから網羅的に検出するとともに、最終的に得られた関連情報を論理的関係に従って分かり易く要約・整理する技術が必要となる。本研究プロジェクトでは、これらの３つの技術を研究開発し、それらを連携させることで図１のような支援を可能とする計画である。なお、要約・整理に関しては渋木ら[4]、論理的関係解析に関しては村上ら[3]の研究成果と連携し、彼らと共同で研究プロジェクトを進めていく。

3. 情報信頼性判断支援システム

本研究プロジェクトの目標の１つに、情報信頼性判断支援技術を実装した信頼性判断支援システムの開発がある。２節で、情報信頼性判断支援のためのアプローチについて述べたが、着目する情報と論理的あるいは時間的に関係する情報を、その根拠と供に提示する、と一口で言っても、関係する情報は大量になる場合があり、どのような情報をどのような手順で利用者に提供することで、効果的に情報信頼性判断支援を実現できるかは、自明ではない。我々はこの問題に対する最初の試みとして、信頼性判断支援システムの動作を示す、モックアップシステムの試作を行った。本節では、このモックアップ支援システムの概要を説明する。

モックアップ支援システムでは、利用者が信頼できるかどうか迷っている単文レベルのテキスト情報(以下では着目言論と呼ぶ)を入力するものとする。この着目言論に対して、まず着目言論の信頼性判断に大きく関わる重要な情報がまとめられている要約画面(図２）を提示する。要約画面には、着目言論を支持する根拠などの

肯定的言論や、反対する対立意見とその根拠などの否定的言論のうち、重要なものを選別して出力する。利用者は両方の言論の内容を比較・検討することで、一方に偏らない信頼性判断を下せるようになる。

図２．要約画面の例

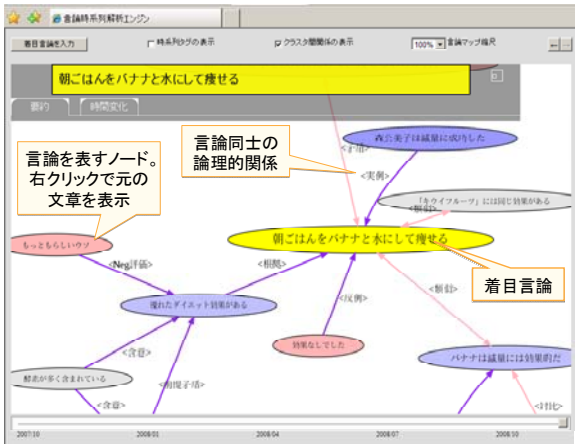

図３ 俯瞰マップ画面の例

一方、着目言論に関係する情報全てが要約画面に出力できるわけではなく、また、特定の情報に関して、その情報に関係する情報をより詳細に調べたい場合があり得る。そこで、要約画面とは別に、着目言論に関係する様々な言論が根拠や対立などの論理的関係で結ばれたネットワーク図で提示する俯瞰マップ画面(図３)を用意する。各言論に賛成・反対する意見を一望できるとともに、その根拠となる情報や対立する情報を辿っていくことができ、利用者が自身の関心に従い、信頼性判断に関わる情報を効率的に探索していくことが可能となる。

さらに着目言論の出現数の時間変化のグラフと、変化の要因となった出来事を提示する時間変化画面や、各言論が含まれている元のテキストを表示する言論詳細画面を用意している。

利用者が、これらの画面を相互に行き来することで、偏った情報に縛られることなく、利用者の目的や状況に応じた信頼性判断を実行可能になると想定している。

4. 時系列分析による信頼性判断支援

我々は図１で示したような信頼性判断支援を目指して、特に言論の時系列分析について取り組んでいる。本節ではその内容と目的について述べる。

言論の時系列分析とは、Web 文書における、単文レベルのテキスト情報(言論)の出現数の時間変化を調べることで、ある言論と時間的に相関して出現する言論をその言論の時間的な関連情報として検出し、さらには時間的に相関して出現する言論の出現数の変遷パターンから、それらの言論間の関係を推定する技術である。Web 文書における特定の言語表現の出現数の変遷を調べて、情報信頼性分析を行う研究としては山本らの研究[5]などが存在するが、本技術では着目する言論の時間変化を調べるだけでなく、その結果から着目言論の時間変化に影響している関連言論を抽出する点にポイントがある。

２節で、クリティカルシンキングに基づく信頼性判断支援を行うため、着目言論に関係する情報を示すことを述べた。一般にある情報に関係する情報には、常にその関係が成立するものの他に、ある時期や状況でのみ成立するものもある。例えば、ある言論の流行り廃り(ここでは、その言論の出現数が大きく増大もしくは減少することを指す)の要因は、その言論の根拠などとは異なり、流行り廃りのときにのみ存在する関係といえるが、そうした要因を変化時期の情報と合わせて提示することは、その言論の信頼性判断において有用である。またある変化以後、元の言論と合わせて語られるようになった言論は、その変化の結果や顛末を示す手がかりとなる言論と考えられ、同じく有用な情報である。

こうした時間的に関係する言論を求めるには、単純には様々な言論の時間変化を調べて、着目する言論の変化と相関して変化する言論を求めればよいが、言論には膨大なバリエーションが存在するため、偶然同タイミングで変化するものもあり、それだけではうまくいかない。そこで、我々はまず着目言論の時間変化から関連する言論の候補を求めて、さらに言論間の共起関係や言語的制約のような他の条件と組み合わせて､関連する言論の候補を絞り込む手法をとる。現在、複数キーワードの組合せによる簡易的な言論表現を用いて、着目言論の変化点付近からその変化要因を求める手法を検証している[6]。

また言論には、単に時間がたって廃れたり状況の変化によって、有効性が失われてしまったものが存在する。インターネット上には、そうした古くて無効となった情報も新しい情報と混在しているため、信頼性判断を適切に支援するためには、そのような無効情報を警告または除外することが望ましい。無効となった情報を警告・除外することで、利用者がたまたま目についた古い情報により間違った判断をしてしまうことを防ぐことができる。

無効な情報の検出には、単純には各情報が作成された時間情報を求めて、古い情報をチェックする手法が考えられる。しかし、インターネット上には古くてもなお有用な情報も多数蓄積されており、また、古くて無効になってしまうタイミングが情報の内容や分野に応じてそれぞれ異なるため、一律に古い情報を除外・警告する手法はうまくいかない。そこで、我々は、着目する言論の出現数の時間変化だけでなく、その言論の時間変化と相関して変化する関連言論の変化も調べて、総合的にある言論が無効となったかどうかを判定する手法をとる。例えばある言論が減少傾向となったときに出現して、以後、負の相関性をもって変化する言論は、時間的に対立する言論ととらえ、その存在によって元の言論が無効となったかどうかを判定する。他にも、ある言論が出現しやすい文書集合や発言者集合の特徴を求めて、そうした特徴の減少などを、複合的にチェックすることを想定している。

5. おわりに

本稿では、情報信頼性判断支援研究プロジェクト全体と、特に時系列分析の観点から、信頼性判断支援における課題とアプローチに関して、我々の取り組みを説明した。今後は４節で述べた研究課題を進めると供に、３節で述べたような情報信頼性判断支援として提供できるサービスイメージについても実データに則して具体化を進め、2010 年度には、情報信頼性判断支援Ｗｅｂサービスの実証実験を公開する予定である。

謝辞

本研究は、独立行政法人情報通信研究機構(NiCT)の委託研究「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」の成果である。

参考文献

[1] S. Kurohashi, S. Akamine, D. Kawahara, Y. Kato, T. Nakagawa, K. Inui, Y. Kidawara, “Information Credibility Analysis of Web Contents”, In Proceedings of the Second International Symposium on Universal Communication, 2008.

[2] 木俵 豊, 他,「Web コンテンツの信頼性分析」, 言語処理学会第 15 回年次大会, 2009.

[3] 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･言語理解とコミュニケーション合同研究会, 信学技報 Vol.108 No.141, 2008-NL-186, pp.55-60, July 2008.

[4] 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, 2009

[5] 山本 祐輔, 手塚 太郎, Adam Jatowt, 田中 克己,「ほんと?サーチ: 検索結果の集約とページ生成時間分布解析による Web 情報の信用度評価」, 日本データベース学会 Letters, Vol.6, No.1, June 2007.

[6] 大西 貴士, 他: 時系列分析によるＷｅｂ文書の情報信頼性判断支援 : 時系列変化からの重要トピックの抽出, 言語処理学会第 15 回年次大会, 2009.