

平成19年度 新規委託研究
「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」
研究計画書



1. 研究テーマ

「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」

2. 研究開発の目的

2006年の日本国内におけるインターネット利用人口は約8,754万人、人口普及率は68.5%に達している。また、インターネット上では様々なサービスが提供されており、国民の社会経済生活を支える重要なインフラともなっている。

インターネット上の情報流通量もブロードバンド化の進展やブログやSNS（Social Network Service）等の普及に伴い、膨大となっている。これらの情報の中には出所が明らかでないものや有害なもの、正しくないものが混在し、信頼できる情報を得ることが難しくなっている。これらの状況は、フィッシング等の犯罪等にも発展し、インターネットの社会的影響におけるマイナス面となっている。

現在、ネットワーク上の膨大な情報から、利用者が必要な情報を得る手段として、検索エンジンが広く利用されている。しかしながら、検索エンジンは必ずしも情報の信憑性に基づいて出力をしているわけではない。情報の信憑性は専ら利用者の判断に任されているのが現状であり、有効な対応技術が無いのが実情である。

しかしながら、情報の内容の真偽や正確性を検証することは困難である上に、情報が意見を述べるものである場合には、利用者により考え方や受け止め方は異なることから、その真偽や正確性を検証することはさらに困難である。そのため、情報の信憑性は、最終的には個々の情報利用者が判断しなければならない。

そこで、ネットワーク上の情報について、言語解析技術や知識処理技術を活用して分析し、利用者による信憑性の判断を支援する技術の実現が求められている。

本研究開発では、Webコンテンツを構成するテキスト、画像・音声・映像情報等の信頼性をデータ解析的な手法に基づき分析・可視化する技術、Web上に膨大に流通している文字情報について、言語処理技術や知識処理技術等を適用することによって、情報間の論理的関係や時間的変遷等を分析・可視化する技術を実現する。

以上の技術開発を通じて、本研究開発ではネットワーク上の文字、音声、画像・映像情報について、情報の信頼性を分析する技術を確立し、信頼できる情報を提供することで、誰でもが思いのまま、簡単に、信頼して、コンテンツを取り扱い、高度に活用できる環境を実現することを目的とするものである。

なお、本研究開発は、総合科学技術会議 科学技術連携施策群「情報の巨大集積化と利活用基盤技術開発」の中で、経済産業省「情報大航海プロジェクト」、文部科学省「革新的実行原理に基づく超高性能データベース基盤ソフトウェアの開発」と連携を図りながら進めることに留意する。

3. 研究開発期間及び予算

研究開発期間：平成 19 年度から平成 22 年度末までの 4 年間。

予算：平成 19 年度は総額 248 百万円程度を上限とする。また、課題毎の内訳の上限額は以下のとおりとする。

課題ア： 年間 120 百万円程度

課題イ： 年間 128 百万円程度

なお、平成 20 年度以降の予算については未定ではあるが、提案を行う際には、課題毎に平成 19 年度提案額と同額或いは未満の金額で提案を行うこと。

また、平成 20 年度以降の予算については、全体の進捗管理を行う情報通信研究機構プログラムディレクターからの助言に基づき、金額が変更となる（場合によってはゼロ円、すなわち研究停止となる）可能性もある。

4. 研究開発課題

以下の通りとする。なお、応募にあたっては、課題ア、課題イのそれぞれを単位として、研究提案を行うこと。

課題 ア Web コンテンツの分析技術

課題アー 1： 画像・音声・映像情報の分析技術

課題アー 2： テキスト情報の分析技術

課題 イ 意味内容の時系列分析技術

課題イー 1： 意味理解評価技術

課題イー 2： 時系列分析技術

課題ア：Webコンテンツの分析技術

画像・音声・映像情報が Web コンテンツの一部としてインターネット上に大量に流通しているが、一般の人々が通常の PC によって容易に映像や音声情報を編集できるようになったこともあり、価値の低い情報や信頼性の低い情報が急増している。さらには、オリジナルのコンテンツを、悪意を持って改ざんした Web コンテンツも多く存在するようになってきている。画像や映像、音声情報を含んだ Web コンテンツはテキストだけのものよりも多くの情報量があり、一旦誤った情報がインターネット上に流通すると、多くの情報利用者が誤った内容を信じてしまうことが危惧される。

その大量の Web コンテンツを構成するテキスト、画像、音声、映像情報等の信頼性をデータ解析的な手法に基づき分析する技術を開発する。

課題アー 1：画像・音声・映像情報の分析技術

Web コンテンツ上の画像・音声・映像を含む情報の信憑性を評価する手がかりを

情報利用者に提示することを目的として、以下の条件を満たす画像・音声・映像情報の分析技術を開発する。

- ・画像・音声・映像情報とそれを説明するテキスト情報や周辺のコンテンツから画像・音声・映像情報の内容を推定し、インターネット上から同等のテキスト情報とそれと関連づけられた画像・音声・映像情報を取得し、対象とする情報と取得された情報を分類するなどして、画像・音声・映像情報とそれを説明するテキスト情報の不整合度を検証するクロスメディア（メディア情報横断型）情報分析エンジンを開発する。

課題アー２：テキスト情報の分析技術

情報利用者が Web コンテンツの信憑性を評価する手がかりとして、Web コンテンツを構成するテキスト表現の表層的特徴（文体、表現、語彙等）やテキスト情報の発信者の専門性等が寄与することが知られている。

このため、情報利用者が Web コンテンツを構成するテキスト情報の信憑性を評価する基準を与えることを目的として、以下の条件を満たす、テキスト情報の分析技術を開発する。

- ・ブログなどを含む Web コンテンツを構成しているテキストについて、誤字・脱字、語彙の多様性、文章の読みやすさ等、テキストの表層的特徴を多様な観点から分析し、テキストの品質を分析・提示する表層的特徴分析エンジンを開発する。
- ・発信者の情報発信履歴等を解析し、情報発信者が取り扱うことの多い話題や話題構造等の特徴を分析・提示する発信者分析エンジンを開発する。

課題イ：意味内容の時系列分析技術

Web コンテンツにおけるテキスト情報は、情報利用者が Web コンテンツの内容そのものを理解するために重要な役割を持つ。従ってテキスト情報については、課題アにおけるデータ解析的な手法だけでなく、その内容に基づく意見について分析することも重要となる。そこで、情報利用者が意思決定を行う際に、情報の信憑性に関する判断支援を行うことを目的として、Web 上のテキストの集合中に潜在している意見を抽出するとともに、意見間の論理的関係等を抽出する意味理解評価技術や、この意味理解評価技術をもとに各意見の時間的変遷や相互の関係、情報伝搬過程等を分析する時系列分析技術を、言語解析技術を用いて研究開発する。

課題イー１：意味理解評価技術

大量・多岐にわたる Web 上のテキスト情報から、情報利用者が着目する領域・観点に関係する情報を、意味解析を通じて分析・整理し、情報利用者が個々の判断基準で Web 情報を容易に活用可能となるために、語レベルの知識ベース構築技術、意見間の論理的関係抽出技術、意見の分析・要約技術を研究開発する。

- Web上のテキスト情報の意味解析を可能とするため、語の間の論理的関係（因果関係、含意関係、上位下位関係、全体部分関係等）と、語が示す概念の実世界での利活用に関連する知識（例えば、語を「風邪薬」とすると、「風邪薬」の利用方法、「風邪薬」の効果情報、「風邪薬」に關係する場所情報等）について記述可能な知識ベース（上記のような体系化した知識をデータベース化したもの）を設計する。さらに大規模Webコーパス¹に適用することで、意見の論理的関係を認識するための大規模な語彙知識ベースを構築する技術を開発する。
- Web上のテキスト情報から、そのテキスト中に記述された内容や出来事の特徴的に表現する意見を抽出し、さらに類似の意見をまとめ上げる意見抽出エンジンを開発する。また、照応・省略関係を含む深い言語解析を行うことで、任意の2つの意見間の論理的関係を識別する論理関係認識エンジンを開発する。
- 大量・多岐にわたるWeb上のテキスト情報から、情報利用者が着目する領域や観点に關係する情報を分析・整理し、各意見間の論理的関係や重要度と共に提示する意見分析エンジンを開発する。意見に付与する重要度を基に、情報利用者が着目する領域や観点における重要な情報発信者を発信者重要度と共に出力する。さらに、情報利用者が着目する領域・観点の意見の全体像を概観できる要約文書作成エンジンを開発する。

課題イー2：時系列分析技術

情報利用者が着目する意見の有効性を判断する手がかりの提示を目的とし、Web上から抽出された意見の時間的な変遷に関する表現手法や、情報利用者が着目する意見と關係する意見（背景、対立、派生など）の変遷や相互の關係を、Web上のテキスト情報間の默示的な参照關係や論理的な關係に関する解析をした上で分析する技術、各意見の時間的な安定性や情報伝搬過程における段階を分析する技術を開発する。

- Web上のテキスト情報に記述されている情報の変遷（背景、対立、派生など）を追跡可能とするため、課題イー1の意見抽出エンジンで抽出された意見の変遷を定量的に取り扱う手法を定式化する。また変遷に伴う変化部分の表現手法を確立する。
- 情報利用者の情報要求（キーワードリスト等）を入力とし、この入力に応じて複数のWebテキスト上から、情報検索要求に関連する意見の変遷と關係の種類を出力する時系列クラスタリング手法を確立する。時系列クラスタリングの結果、クラスタリングされた各意見に対して、それが既に確立した意見であるのか、それとも急速に多くの人々に拡散しているフェーズなのか、既に廃れたフェーズなのかなど、情報伝搬過程におけるフェーズを判定する伝搬フェーズ判定エンジンを開発する。

¹ 情報通信研究機構知識処理グループで構築する数億ページ規模のWebコーパス。大規模Webコーパスの概要、利用条件等に関しては、別添資料を参照すること。

5. 研究テーマ選定の背景、研究開発の必要性及び他で実施されている類似研究との切り分け

1) 当該研究テーマを取り巻く現状

インターネット上に流通する膨大な情報は、その内容も事実や知識に関するもの、意見を述べるものなど多岐にわたる。しかし、情報の内容の真偽や正確性を検証することは、膨大な背景知識を蓄積した知識ベースを構築して論理的演算を行ったとしても、実現することは困難である。そもそも、情報が意見を述べるものである場合には、利用者により考え方や受け止め方は異なり、その真偽・正確性を決定することはできない。情報の内容の真偽や正確性はわからないが、それを推定するために行われるのが信憑性に関する判断であり、最終的には個々の情報利用者が行うものであると考えられる。

信憑性の判断支援については、技術開発のアプローチの視点として以下の4つが重要となることが指摘されている。(ただし、これらに関する研究は近年取り組みが始められてきた段階であり、実用レベルとして社会に提供されるまでには到っていない。)

- ①情報の外観
- ②情報発信者
- ③情報の意味内容
- ④情報に対する社会的評価

①情報の外観

文章の形式的特徴(文の長さ、句中の文節数、漢字・カナの割合、語の出現頻度、接続表現等)をもとに、文章の読みやすさ、語彙の多様性、論理構成等を分析し、小論文の自動採点を行うことを目指しているシステムの研究事例がある。また、小説家等の著者推定を行う計量文献学による研究は行われているが、Webページの信憑性との関連を評価する研究事例は少ない。文体やページデザイン等の情報の外観は、情報利用者が信憑性を評価する上での手がかりの一つとなると考えられるが、研究事例は少なく、未だ技術は確立されていない。

また、Webページのデザイン等に基づく研究としては、スタンフォード大学のFoggらが情報の信憑性判断に影響を与える要素に対して行った被験者実験があり、この成果に基づいて、信憑性に関して重要な要素を自動処理する方法の提案等を行う研究事例がある。

②情報発信者

情報発信者の属性(公的機関、報道機関、学術研究者、個人等)や、これまでの情報発信内容やコミュニティの中で役割等は、情報利用者が信憑性を評価する上での手がかりの一つとなると考えられる。

発信者の特定に関する技術やリンク関係・リンク間の信頼性表現を基に情報発信者

に関する分析を行う技術等がある。発信者の特定に関連する研究開発を行っている事例は少ないが、情報通信研究機構知識処理グループによる情報発信者クラス分類に関する研究では、Web ページの URL や、テキスト文字列等を基に、個人（有識者、一般等）、団体（営利団体、非営利団体、報道機関等）の発信者クラスを自動的に分類する研究が行われている。

また発信者に関する分析を行っている研究事例として、ブログを対象としたリンク解析等を通じて、ブログコミュニティにおいて話題の質的・量的な変化を促すなど重要な役割を果たしている投稿者の検出を行う研究等がある。

③意味内容

利用者が着目する領域や観点に関連するインターネット上の複数の文書について、情報の意味内容の解析をもとに、それぞれの言論・意見の全体での位置づけや相互の関係性等を提示することによって、利用者による情報の信憑性に関する判断の支援に寄与するものと考えられる。

特に、あるタイミングでの評価だけではなく、情報が時間的にどのように変化しているか、相互にどのような影響を与えているのか等の動的な分析と組み合わせることにより、情報利用者が情報の意味内容について理解及び評価する基準を与えることができる。関連する研究事例としては、テキストの時系列データをもとにトピックの出現検出や特徴分析、時系列的なクラスタリング等を行うといったものがある。また、インターネット・アーカイブを活用して、Web ページの生成時期を考慮した情報の評価に関するもの等もある。

しかし、これらの研究を通じて、Web 上の特定の話題やキーワードに対する時間的な出現数の変化を計数する技術の確立がなされてきているが、情報の関係性や変遷を表現・分析する技術は未だ確立されていない。

④情報に対する社会的評価

情報が社会からどのように評価されているのかを解析することによって、情報の信頼性を分析するものであり、他者が Web ページやブログ等で行っている情報に対する評価を言語解析等によって抽出、分析することが考えられる。近年、Web 上のテキストから商品等に関する評価情報を抽出・整理する研究が進められてきているが、情報に対する評価の抽出に関しては十分な研究がなされているとは言えない状況である。

表層的なアプローチでは評価対象となるキーワードに近接している評価語をもとに、肯定評価・否定評価を判別するといった研究がなされているが、トピックにより評価語の意味が異なることから評価表現辞書等を構築する必要がある。但し、情報に対する社会的評価は、必ずしも肯定・否定の 2 値に留まらないことから、自然言語処理に基づく意味内容の解析等を踏まえ、その関係性を分析していく必要がある。

これらの技術開発を進めていくアプローチとしては、構文解析や意味解析等を通じて文書間の関係（因果関係、含意関係、上位下位関係、全体部分関係）等を明確にする深い自然言語処理に基づくものと、文書内のキーワードの出現頻度等を基にした特徴ベクトルで文書のトピックを表し、特徴ベクトルの類似等をもとに文書のクラスタリングを行ったり、キーワード等の共起頻度等を統計的に観測することで特徴を見いだしたりする表層的なアプローチとがある。

自然言語処理に基づく言語の深い解析のアプローチでは、言語理解に必要となる語彙知識等を整理した知識ベースを基に情報の意味内容や照応・省略関係等を解析し、意味内容を踏まえた上で、文書間の類似、反対、理由、サポート等の関係を調べたり、論理的整合性を検証したりすることを目指す。解析結果は意味内容を踏まえた正確なものとなるが、大規模な解析では演算量が膨大となり長い処理時間を要する。また、予め知識ベースを構築する必要がある。情報の意味内容を解析することが困難であることなどもあり、従来は基礎的な研究に留まっていたが、近年の自然言語処理技術の進展とともに、情報の意味内容を取り扱う研究が行われつつある。

一方、表層的なアプローチでは、文の意味内容には踏み込まずに、キーワードの出現頻度等を基に数理的な解析を通じて、文書のクラスタリングを行うといったものである。意味内容を踏まえていないことから、文書間の関係性を解析することが難しいこと、ノイズデータが混入する可能性があるなどの課題があるものの、必要とされる計算量が少ないことから利用者の入力に対し、現実的な応答時間で出力が期待できることや、知識ベースを構築するといった大規模な前処理が必要ないなどの特徴がある。

以上のような現状を踏まえ、本研究開発の課題アでは、①情報の外観、②情報発信者に関する分析技術について、情報利用者に対して現実的な処理時間による分析結果の提示を可能とする表層的なアプローチにより研究開発を行う。課題イでは、情報の意味内容の解析等が重要となる③情報内容、④情報に対する社会的評価に関する分析技術について、自然言語処理による深い解析のアプローチにより研究開発を行うこととする。

2) 研究開発の必要性

インターネット上に流通する情報量は近年増大しており、特に、ブログや SNS 等の普及に伴い、一般の個人であっても、誰でもが簡単に情報発信をするようになっていく。ブログ等を通じて、既存のメディアとは異なるような様々な視点による情報が得られるようになっていく一方で、出所が明らかではないものや正しくないものなど、情報の多様性が増し、信頼できる情報を得ることが難しくなっている。

このため、誰でもが信頼してコンテンツを取り扱い、高度に利活用できる環境を実現することが必要であるが、情報の信憑性に関する研究開発は、大学や民間企業等において手がけられはじめたところであり、未だ社会的に実用化レベルに達するまでに

は到っていない。このため、安心・安全な社会インフラの利活用を確保するための研究開発として早急に推進する必要性がある。

3) NICT 及び他で実施されている類似研究との切り分けと NICT 委託研究における本テーマの位置づけ

- ・通信・放送機構（現：情報通信研究機構）では、平成 11 年度から平成 13 年度にかけて「情報通信不適正利用対策技術の研究開発」を実施し、インターネット上で誹謗中傷やわいせつ情報の流通、違法物品の売買広告等を行っている不適正利用に対し、プロバイダー等による発信者への警告・掲載情報の削除等の対応を支援する技術、不適正利用の抑止に資する技術に関する研究開発が行われた。この中で、不適正利用に係る情報の構造分析・特徴抽出技術の研究開発が行われ Web ページの特徴値をもとに「高い確率で適正」「高い確率で不適正」「どちらとも言えない」の 3 つに分類することを実現している。これに対し、本研究テーマでは、不適正利用に限定せずに広く情報の信憑性に関する検証技術の実現を目指すものである。また、上記研究においては、Web ページ単独の特徴量を基に分類分けを行っているが、本研究テーマでは、文書相互の論理的関係性や時間的変遷などを基に信憑性検証技術の実現を目指すものである。
- ・情報通信研究機構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では、信頼性分析プロジェクトを推進しており、クロールによる日本語の Web 文書の収集を行い、約 2 億ページの大規模 Web コーパスを構築している。この大量の Web コンテンツの情報の信頼性を、情報発信者の信頼性、Web ページの外観的な特徴に基づく信頼性、情報内容の信頼性、内容／発信者の評価情報に基づく信頼性の 4 つの視点で分析・提示することによって、利用者が Web コンテンツの信憑性を判断する上での支援を行えるシステムの構築を目指しており、これまでに情報発信者クラスの間接分類技術を開発している。これに対し、本研究テーマでは、インターネット上の文書だけではなく、画像、音声、映像をも対象とした信憑性の検証技術の実現を目指している。また、情報の時間的変化といった視点からの信憑性の検証技術の実現を目指すものである。

4) 標準化の動向

本研究テーマに関連する技術に関する標準化は現時点では行われていない。

6. 研究開発の到達目標

【全体的な留意事項】

- ・各研究開発課題では一般の情報利用者を想定し、分析結果の提示等においては、一般の情報利用者にとって、わかりやすい提示方法等の実現を図ること。
- ・課題アと課題イでは、最低限、日本語による Web コンテンツを分析対象とすること。
日本語による Web コンテンツに加えて、他の言語による Web コンテンツも対象に含む提案を除外するものではない。
- ・目標到達状況に関する評価に評価用トピック、評価用データを用いる研究開発課題では、評価用トピック、評価用データの客観性・中立性を担保するために、情報通信研究機構の指示のもと、その選定や作成を行うこと。
- ・応募にあたってはそれぞれの研究開発課題について、中間報告（平成 20 年 10 月頃に実施予定）までに達成する研究開発目標と、最終目標について、それぞれ数値目標を含めた具体的な目標を提案すること。
- ・以下に記載の目標は全て最低目標であり、提案を行う際には、同等かそれ以上の目標設定とすること。

【課題別目標】

課題ア全体における留意事項

- ・課題アでは、データ解析的なアプローチにより、Web コンテンツを構成するテキスト、画像、音声、映像情報等の信頼性を、ISP や検索エンジンなどの実環境において、現実的な処理時間で分析・提示することにより、情報利用者が個々の判断基準で Web コンテンツを容易に活用可能となる技術を開発すること。
- ・各課題及び各項目の成果を、ISP などの実環境における実証実験を通じて確認すること。
- ・応募にあたっては、各課題について、利用者からの分析要求後、分析結果を出力するまでに要する時間に関して、数値目標を含めた具体的な目標を提案すること。

課題アの個別課題目標

課題アー 1 画像・音声・映像情報の分析技術

- ・画像・音声・映像情報とそれに対応するテキスト情報に対して行った分析が不整合を検出するために有効であるか、評価用に選定した 10 以上のトピックについて人間による主観評価を行い、有効性 80%以上を達成すること。

課題アー 2 テキスト情報の分析技術

- ・情報利用者が入力した情報要求に関連するテキストに対して、表層的特徴を分析・提示することを実現するとともに、ISP、検索エンジンなどの実環境において、評価用に選定した 10 以上のトピックについて、人間による評価と比較して、精度

80%以上を達成すること。

- ・また、評価用に選定した 10 以上のトピックに関して任意に選択したブログ等の情報発信者の特徴を分析・提示することを実現し、人間による評価と比較して、精度 70%以上を達成すること。

課題イ全体における留意事項

- ・課題イでは、情報の意味内容の解析を含む自然言語処理による深い解析を通じて、情報相互の論理的関係性や時間的変遷等进行分析・提示することにより、情報利用者の意思決定を支援する技術を開発すること。
- ・情報通信研究機構知識処理グループが構築した数億ページ規模の大規模 Web コーパスを活用して研究開発を行うこと。
- ・課題イー 1、課題イー 2 の研究開発は、相互に十分に連携して行うこととし、最終年度の実証実験においても、課題イー 1、課題イー 2 のそれぞれの実証装置を単純に接続して行うのではなく、両者を融合させた上で ISP などの実環境において実証実験を実施すること。

課題イの個別課題目標

課題イー 1 意味理解評価技術

- ・開発した知識獲得手法を大規模 Web コーパスに適用することで、200 万エントリ規模の上位下位関係を獲得し、その全体に付与された知識（語の間の論理的関係、利活用に関連する知識）に対し、精度 80%以上を達成すること。
- ・与えられたトピックにおいて主要な対立論点を認識し、それに関する意見をテキストから 80%以上の精度で抽出する意見抽出器を開発すること。
- ・また、10 以上のトピックについて人手で作成した評価用データに対して、2つの意見間の類似、対立、矛盾、論拠等の論理的関係の識別において、精度 80%以上を達成すること。
- ・さらに、10 以上のトピックについて人手で作成した評価用データに対して、情報利用者が着目する領域や観点における重要な意見や発信者の抽出において、再現率、適合率の双方について 80%以上を達成する意見分析エンジンを開発すること。分析結果は、利用者からの分析要求後、2～3 分程度の時間内で出力する。
- ・また、要約文書生成エンジンについては、10 以上のトピックについて人手で作成した評価用データに対して、人間が作成した参照要約と比較し精度 80%以上を達成すること。分析結果は、利用者からの分析要求後、2～3 分程度の時間内で出力する。

課題イー2 時系列分析技術

- ・評価用に選定した 10 以上の特定トピックにおいて、着目する意見とその関連意見の時系列変化を出力。さらに、意見数、意見背景、関連意見との関係などから総合的に変化点とその時点での代表文書を出力。試験トピックの知識を持つ人間による分析結果と比較し、出力結果の適切度 80%以上を達成すること。
- ・また、評価用に選定した 10 以上の特定トピックにおいて、出力された時系列クラスタリング結果の意見間の関係が妥当であるかどうか、人間による主観評価を行い、評価者の 80%以上の妥当性を達成すること。また上記試験トピックにおける着目する意見の分析結果に対して伝搬フェーズ判定エンジンが出力したフェーズが適切であるか人手で判断し、精度 70%以上を達成。
- ・上記の分析結果は、利用者からの分析要求後、2～3 分程度の時間内で出力する。

7. 期待される波及効果

1) 類似研究開発に期待する波及効果

本研究テーマを通じて開発される、知識処理技術等は、情報の信憑性検証への適用に留まらず、計算機による自然言語の理解や人との自然なコミュニケーション、知識の体系化の実現など、より高度な技術の基盤ともなる技術であり、これらの発展を促進する効果が期待される。

2) 実用化に期待する波及効果

本研究テーマを通じて、情報信憑性検証技術が開発されることにより、利用者の経験や知識といったリテラシーによらない安心、安全なインターネット利活用環境の実現に寄与することから、インターネット上での社会経済活動を活発化させるなどの効果が期待される。また、従来の検索エンジンなどを使った利活用では気がつかなかった情報へのアクセス機会をも生み出すことにも繋がり、国民の社会経済生活の高度化にも資することが期待される。また、企業のマーケティング活動などへ応用することにより、企業活動に消費者ニーズ等を効率的に反映することが可能になるものと考えられる。

3) 標準化活動面に期待する波及効果

本研究により、情報信憑性検証技術について研究開発が進められ、その実用性が広く認識されれば、本研究の成果がインターネット上のテキスト、画像、音声、映像情報等の信頼性分析に関する先導的研究例となる。本研究を通じて、情報信頼性の分析に適したコンテンツの記述法等の普及を図ることなどにより、信頼性の高い情報流通を促す情報発信プロセスの実現にも大きく寄与するものと考えられる。

8. 研究開発スケジュール

- ・本研究テーマの研究開発期間は、平成 19 年度から平成 22 年度の 4 年間であり、スケジュールは概ね以下のとおりである。

	19 年度	20 年度	21 年度	22 年度
課題ア Web コンテンツの分析技術	設計・開発	技術開発	技術開発	実験・評価
課題イ 意味内容の時系列分析技術	設計・開発	技術開発	技術開発	実験・評価
		△ 中間評価	△ 中間評価	

なお、本研究テーマは、総合科学技術会議 科学技術連携施策群「情報の巨大集積化と利活用基盤技術開発」の中で、経済産業省「情報大航海プロジェクト」、文部科学省「革新的実行原理に基づく超高性能データベース基盤ソフトウェアの開発」と連携を図りながら進めることから、情報通信研究機構 プログラムディレクターのもと進捗管理が行われる。そのため、年に複数回の関係者による会合を開催すること。また、必要に応じて研究開発の方向性に関する助言等がプログラムディレクターにより行われ、進捗状況に応じて、予算額の変更も平成 20 年度以降あり得る（参照「3. 研究開発期間及び予算」）。

なお、最終年度には課題ア、イを連携した実証実験を実施すること。

情報通信研究機構知識処理グループ 「大規模 Web コーパス」について

1. 概要

インターネット上から自動収集し、日本語テキストのページのみを自動選択した数億ページ規模の Web ページの集合である。

以下のフォーマットで、URL と本文をページ区切り文字列で分割して、複数のページを 1 ファイルに保存している。

- ページ区切り文字列:

フォーマット: <<<10 桁の数字 5 桁の数字>>>

10 桁の数字は次のページまでのバイト数

5 桁の数字は URL のバイト数

- URL

- HTTP ヘッダと本文: HTTP ヘッダを含む html テキストで、文字コードは取得時の文字コードのまま

(データ例)

```
...
<<<0000025778 00023>>>
http://www.nict.go.jp/
HTTP/1.1 200 OK
Date: Mon, 27 Aug 2007 07:26:13 GMT
Server: Apache
Last-Modified: Mon, 27 Aug 2007 04:55:33 GMT
ETag: "2a1-63b0-2aa51b40"
Accept-Ranges: bytes
Content-Length: 25520
Connection: close
Content-Type: text/html

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<HTML>
<HEAD>
<META http-equiv="Content-Type" content="text/html; charset=iso-2022-jp">
<TITLE>NICT 独立行政法人 情報通信研究機構</TITLE>
<LINK href="css/style.css" rel="stylesheet" type="text/css">
</HEAD>
<BODY leftmargin="0" topmargin="1" marginwidth="0" marginheight="1">
<DIV align="center">
<TABLE width="800" border="0" cellpadding="0">
```

```

<TR bgcolor="#FFFFFF">
  <TD width="497"><a href="index-J.html"><IMG src="images/header_logo.gif" alt="NICT 独立行政法人 情報通信研究機構" width="267" height="36" hspace="10" vspace="5" border="0"></a></TD>

</HTML>
<<<0000007107 00039>>>
http://www.nict.go.jp/about/index.html
HTTP/1.1 200 OK
Date: Mon, 27 Aug 2007 07:31:17 GMT
Server: Apache
Last-Modified: Fri, 06 Jul 2007 02:46:48 GMT
ETag: "e1e-1ab2-4e737600"
Accept-Ranges: bytes
Content-Length: 6834
Connection: close
Content-Type: text/html

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<HTML>
...

```

2. 提供方法

以下の方法により無償で提供する。

(1)ハードディスクでの提供

採択者がネットワーク接続可能なハードディスク（例えば、イーサネットに接続可能な PC+ハードディスク）を NICT けいはんなセンターに持ち込み、NICT の担当者の指示の元でコピーすること。

(2)リモートでの利用

JGN2に接続された NICT けいはんなセンター内のマシンに外部からリモート接続して、取得・利用する。インターネット上からアクセス可能だが、高速アクセスが必要な場合の JGN2 への接続などのネットワーク環境は採択者側で整えること。

3. 利用条件

本委託研究の研究開発目的に限って利用することができる。