平成２１年度 新規委託研究
「インターネット上の違法・有害情報の検出技術の研究開発」
研究計画書

独立行政法人 情報通信研究機構

１．研究開発課題

『インターネット上の違法・有害情報の検出技術の研究開発』

２．研究開発の目的

インターネットは、国民の社会活動、文化活動、経済活動等のあらゆる活動の基盤として利用されるようになり、国民生活に必要不可欠な存在となっている。一方で、インターネット上を流通する膨大な情報の中には、違法・有害情報も含まれ、犯罪を助長するサイトや自殺を誘引するサイト等、社会的に大きな影響を与える事案も発生している。

こうした中で、第 169 回国会において「青少年が安全に安心してインターネットを利用できる環境の整備等に関する法律」が成立した。同法では、表現の自由等に配慮し、国による過度な規制は採用せずに、これまで行われてきた民間における違法・有害情報に対する自主的かつ主体的な取組を、一層推進していくことが強調されている。

民間における自主的取組みでは、ISP やコンテンツ提供事業者、コンテンツ監視事業者等が、契約約款や利用規約等に基づき、規約違反の書込のパトロールや削除等の対応を行っている。そこでは、NG ワード[1]検出等の簡易な検出技術等が活用されているが、通常、単語が表す意味は使用される文脈によって異なることから、違法・有害情報とは無関係な情報が多く検出されてしまうという問題がある。また、いわゆる隠語については、単純な文字列処理では解析することができない。そのため現状では、違法・有害情報にあたるかどうかを、最終的に、人手で確認することが必要となっている。

ブログや SNS 等の CGM（Consumer Generated Media）の普及もあり、インターネット上の情報量は、今後、爆発的に増加していく。こうした中で、ISP 等が自主的な取組を適切に行っていくためには、違法・有害情報の候補となる情報の検出精度を高め、人手にかかる負荷を軽減することによって、監視業務の効率化を図る必要がある。但し、ISP 等によって違法・有害性の判断基準や、対処する情報の種類は異なっていることから、個々の ISP 等の業務にあわせた支援ツールが必要となる。

そのため、本研究テーマでは、現在、ISP 等が行っている監視業務の効率化を支援するための基盤技術として、NG ワードが使われている前後の文脈等を解析することで、単純な文字列処理では得られない非表層的な意味（例えば隠語等）の分析を可能とする技術等を研究開発する。これら基盤技術を、個々の ISP 等が自己の業務形態にあわせて活用することによって、違法・有害情報の候補となる情報の検出精度を高め、人手にかかる負荷を軽減し、監視業務の効率化に繋げていくことを目的とする。

[1] 違法・有害情報に関係する文字表現

３．研究開発期間及び予算

研究開発期間：平成２１年度から平成２３年度までの３年間。
予算：平成２１年度は１８０百万円程度を上限とする。

なお､平成 22 年度以降は対前年度比で 6%削減した金額を上限として提案を行うこと。

４．研究開発課題の内容

ISP やコンテンツ監視事業者等が行っている違法・有害情報の監視業務では、NG ワード検出等の簡易な検出技術が活用されているが、違法・有害情報とは無関係な情報が多く検出されてしまい、人手にかかる負荷が大きくなっているという問題がある。また、違法・有害情報では、いわゆる隠語や意図的な誤字・脱字、当て字、併せ字[2]、口語的表現、視覚的な言語表現（絵文字や顔文字・アスキーアート等）といった表現形態が活用されることがあることなどから、既存の言語処理技術では対処が難しい。

以上のような状況から、本研究開発課題では、ISP 等の監視業務の効率化を支援するための基盤技術として、NG ワードが使われている前後の文脈等を解析すること等によって、ISP 等が監視対象とする掲示板や SNS、Web 等に掲載されたテキスト情報から、違法・有害情報の候補を高精度に抽出することを可能にする技術の研究開発を求めており、そのコンセプト・設計案・プロセス等の具体的な提案を公募する。

別紙に上記を実現するための研究開発方法の一例を示すが、この例に捉われることなく、異なる方法に基づく違法・有害情報検出技術の研究開発の提案も歓迎する。

但し、社会的に違法・有害情報への迅速かつ的確な対応が求められていることから、本研究開発では、新たな検出原理の発見といった基礎的・学術的な貢献を追求するものではなく、研究開発期間終了後に成果が ISP 等の業務に活用され得る、実用的・実践的な技術を開発することを目指していることに留意すること。

尚、本研究開発課題には、上記に示したような違法・有害情報で使用されることの多い表現形態等、違法・有害情報の検出を困難にしている要因について分類し、対処方法等を整理した上で取り組むものとする。

５．研究開発の到達目標

全体目標

・ISP やコンテンツ監視事業者等が行う Web や掲示板、ブログ等の実運用環境における監視業務について、同一の業務を現在の半分程度の人員で行うことを可能とする、監視業務の効率化を支援するための基盤技術を開発する。

[2] （例えば「終」を「糸」と「冬」の二文字で表すなど）

・通常とは異なる意味を持たせたいわゆる隠語や、意図的な誤字・脱字、当て字・併せ字、口語的表現、視覚的表現等、単純な文字列処理による違法・有害情報の検出を困難としている要因に対して対処すること。
・ISP やコンテンツ監視事業者等の実運用環境における実証実験を通じて、研究開発の到達目標について実証するとともに、研究成果の有用性について検証すること。

個別技術に関する目標

・個々の ISP やコンテンツ監視事業者により、監視ツールの仕様や監視に関する業務形態に違いがあることを考慮した上で、個々の ISP 等が活用する監視ツール等から、解析対象となる情報を収集し、違法・有害情報検出技術との間の入出力を可能とする、標準インタフェースを設計・開発すること。
・違法・有害情報に関する知識ベースの設計を行い、実証実験等を通じて実用規模の違法・有害情報の知識ベースを構築すること。（実用規模と考える規模について、数値等を用いて提案書に明記すること（例えば、違法・有害情報の文例等を 1 万件以上　等））
・ISP 等が監視対象とする掲示板や SNS、Web 等に掲載されたテキスト情報から、実運用環境に近い違法・有害性判断基準のもとで、違法・有害情報を再現率[3]90%以上、適合率[4]50%以上の精度で検出する技術を開発する。

実用に向けた目標

・委託研究開発終了後、成果が継続的に活用される予定であり、ISP やコンテンツ監視事業者等において活用されうるものであること。活用方策については、提案書に具体的に記載すること。
・日々新しい違法・有害情報が出現していることから、研究開発終了後においても、違法・有害情報に関する知識ベースのアップデート等の継続的な取組みが必要となるため、新たなコンピュータ・ウイルスが発見された際にウイルス・パターン情報等を関係者間で共有する事例などと同様の仕組み等についても考慮すること。

留意事項

・応募にあたっては、本研究開発において、どのような分野[5]の違法・有害情報を対象に、どのような情報検出を目的として研究開発及び実証を行うのか、具体的に提案

[3] 再現率とは、検出対象となるべき対象の内、実際に検出されたものの比率（違法・有害情報全数の内、違法・有害情報の候補として検出できたものが占める比率）
[4] 適合率とは、実際に検出されたものの中で、検出対象となるべきものが占める比率（違法・有害情報の候補として検出されたものの中で、実際に違法・有害情報であるものが占める比率）
[5] ここでの分野とは、麻薬、自殺など、研究開発対象とする違法・有害情報の種類を指します。

すること。
・解析要求後から解析結果を出力するまでに要する時間等、提案する研究開発方法の性能面に関する数値目標等、その他必要な目標について、具体的に提案すること。
・上記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。また、これらの到達目標は、研究開発の社会や産業面での着実な活用を推進する観点から、本委託研究期間中に受託者と協議の上弾力的に見直しを行う可能性がある。

６．研究開発スケジュール

本委託研究の研究開発期間は、平成２１年度から平成 ２３年度までの３年間であり、スケジュールは概ね以下のとおりとする。

	２１年度	２２年度	２３年度
インターネット上の違法・有害情報の検出技術	設計	試作・評価	改良・評価
			△ 最終評価

７．研究開発の運営管理及び評価に関する事項

情報通信研究機構は、総務省と密接な関係のもと、プログラムコーディネータの助言を受けつつ、本委託研究の適切な運営管理を実施する。外部有識者（評価委員）、プログラムコーディネータの意見を運営管理に適宜反映させるほか、必要に応じて、半年から１年に１回程度本委託研究の進捗状況について受託者から報告を求めることとする。

また、最終評価は平成２３年度に行うが、評価の時期については、本委託研究に係る技術動向、政策動向や本委託研究の進捗状況等に応じて、追加の実施、前倒しする等、適宜見直すものとする。

８．特記事項

受託者は、本委託研究から得られる研究開発成果の産業面での着実な活用を図るため、本委託研究の終了後に実施すべき取組みのあり方や研究開発成果の社会や産業面での活用方策について立案するとともに、立案した取組みのあり方や産業面での活用方策について、研究開発の進捗状況等を考慮して、本委託研究期間中に必要な見直しを行うこととする。

別紙

本件委託研究の実施方法の一例

個々の ISP 等が用いる監視ツールから解析対象とする情報を、課題アで開発する情報収集・解析結果提供技術を通じて入力し、通常のテキスト情報のみならず、口語的な表現や絵文字、顔文字等の視覚的な言語表現を含むテキスト情報についても、課題ウで開発するネット言語解析技術による単語分割[6]を行った上で、課題ウの文書単位での違法・有害情報検出技術による解析を行う。さらにその解析結果について、課題ウの NG 表現検出技術による解析を行い、違法・有害情報の候補となる情報の検出精度の向上等を図る。解析結果は、課題アの情報収集・解析結果提供技術を通じて、個々の ISP 等が用いる監視ツールに提供するものとする。課題イでは、課題ウの文書単位での違法・有害情報検出技術、NG 表現検出技術を実現するための知識ベースとして、違法・有害情報コーパス及び NG 表現辞書に関する技術を開発する。課題エでは、課題ア～ウの研究開発の成果を実証するため、ISP やコンテンツ監視事業者等の実運用環境において試作システムによる実証実験を行う。

課題ア　情報収集・解析結果提供技術

ISP やコンテンツ監視事業者等が、それぞれの分析対象(掲示板やブログ、SNS など)や業務形態にあわせて活用する監視ツール等から、解析対象となる情報を収集し違法・有害情報検出技術に効率的に受け渡すこと、及び、違法・有害情報検出技術による解析結果を監視ツール等に提供することを可能とするために、通信プロトコルや標準フォーマット仕様などの標準インタフェースを開発する。

課題イ　違法・有害情報知識ベース構築技術

違法・有害情報コーパス

課題ウにおける違法・有害情報の候補となる情報の検出を可能とするために、掲示板、SNS、Web 等から抽出した文書・例文と、それらの人間による違法・有害度等の判断結果、違法・有害性ありと判断した理由について記述可能な知識ベース（違法・有害情報コーパス）を設計する。ISP やコンテンツ監視事業者等の協力のもとに、予め人手で違法・有害性の判断を行った文書・例文等を蓄積することにより、精度の高い初期知識ベースを構築する。さらに、違法・有害情報の検出技術に基づいて違法・有害情報の候補として分類された情報について、人間が再評価・修正等を行った結果について、知識ベースに効果的に反映させ、知識ベースを更新する技術を開発する。

---

[6] アスキーアートに関しては、その中に埋め込まれているテキスト情報に対して単語分割を行うことを想定しています。

NG 表現辞書

課題ウにおける違法・有害情報の候補となる情報の検出を可能とするために、NG ワードに関係する単語の上位下位関係や類義語等についても記述可能な知識ベース（NG 表現辞書）を設計する。また、NG 表現辞書では、違法・有害情報に関する表現を、NG ワードを含む単語の組み合わせ（例えば「麻薬の防止」等）や構文パターン（例えば「麻薬を販売する」等）として記述することも可能とする。

さらに違法・有害情報コーパスの統計的解析や例文の抽象化等により、NG ワードや NG ワードを含む単語の組み合わせ、構文パターンを自動獲得する技術を開発する。また、更新された違法・有害情報コーパス情報により、NG 表現辞書を更新する技術を研究開発する。

課題ウ　違法・有害情報検出技術

ネット言語解析技術

通常の表現によるテキストとともに、口語的かつ視覚的な言語表現（絵文字や顔文字、アスキーアート等）を含むテキストに対しても、文書単位での違法・有害情報検出技術、NG 表現検出技術を適用することを可能とするために、口語的かつ視覚的な言語表現に対して、単語分割を実現する、頑健な形態素解析技術またはそれと同等な効果を実現する技術（これらをネット言語解析技術と称することとする）を開発する。なお、アスキーアート等がある場合には、そこに含まれるテキスト情報を同定した上でネット言語解析技術を適用するものとする。

文書単位での違法・有害情報検出技術

課題アの情報収集・解析結果提供技術を通じて入力された文書について、ネット言語解析技術を活用するとともに、課題イで構築する違法・有害情報コーパスに蓄積された、人手により違法・有害性の判断がなされた文書を教師データとし、例えば文書の外形的特徴に対する統計的解析等を通じて、違法・有害性が未知の文書の違法有害性を分類し、文書単位での違法・有害情報の候補を自動的に抽出する違法・有害情報検出エンジンを開発する。

さらに、違法・有害情報のジャンルによっては、違法・有害情報コーパスに蓄積された教師データが少数となることも想定されることから、例えば、教師データだけではなく、同等の分布をもつ教師データ以外のデータを活用することなどによって、頑健かつ高精度な自動分類を可能とする技術を開発する。

NG 表現検出技術

文書単位での違法・有害情報検出技術において出力された違法・有害情報の候補及び違法・有害性の判別がつかなかった情報について、NG ワード等の前後の文脈に関する解析

を行い、当該文書における違法・有害性のある表現の候補を抽出することにより、人手による違法・有害性に関する確認作業の効率化や、違法・有害情報の候補となる情報の検出精度の向上を実現するために NG 表現検出技術を開発する。

ネット言語解析技術、課題イで構築する違法・有害情報知識ベースを活用し、NG ワードをその近似表現（誤字、脱字、同音異義語、併せ字等）を含めて自動抽出する。さらに、抽出した NG ワード等について、その前後の文脈について深い言語解析を行うことによって、当該 NG ワード等の違法・有害性を判別し、違法・有害情報の候補を自動的に抽出する NG 表現の違法・有害性検出エンジンを開発する。

ここで深い言語解析とは、NG ワード等の前後の文脈において、出現する語句、係り受け構造、述語項構造、照応省略関係、トピックなどを分析し、単純な文字列処理では得られない語句の非表層的な意味（例えば隠語等）を解明する能力をもつ言語処理技術を指す。

NG 表現辞書に、NG ワードを含む単語の組み合わせや構文パターンとして登録した場合には、自動抽出した NG ワード等の近傍に、当該単語の組み合わせや構文パターンを構成する全ての単語が存在することを確認し、当該単語の組み合わせや構文パターンの違法・有害性を判別する。尚、単語の組み合わせや構文パターンについては、構文的な変形(受身使役、関係節、名詞化句など)や照応省略、同義語や上位語下位語における置換等についても対処する。

課題エ　ISP やコンテンツ監視事業者等の実運用環境における実証実験

本研究開発テーマの目標は、ISP 等の監視業務の効率化を支援するための基盤技術を開発することにある。このため、上記の課題ア、課題イ、課題ウの研究成果をもとに、ISP 等の監視ツールから入力した情報を解析し、解析結果の監視ツールへ提供するとともに、知識ベースの更新等を行う一連の機能について統合的な評価を可能とする試作システムを構築し、ISP 等の実運用環境において、研究開発の到達目標について実証するとともに、研究成果の有用性について検証する。

研究開発課題選定の背景、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、標準化動向、期待される波及効果

１）当該研究テーマを取り巻く現状

現在、違法・有害情報の検出に用いられているのは、単語レベルで一致したものを検出する技術が一般的である。しかしながら、これらの単語レベルでの検出技術では、違法・有害情報には該当しない多くの情報をも検出するなど、検出精度に問題がある。また、単語等に通常とは異なる意味を持たせた、いわゆる隠語が利用されることも少なくない。そのため、最終的には人手による確認が必要とされている。今後、インターネット上の情報量が急増していく中で、ISP 等が自主的な取組を適切に行っていくためには、検出精度の向上等を通じて、ISP 等の負担を軽減していくことが必要である。そのため、深い言語解析を行い、前後の文脈の意味を解析することで、単純な文字列処理では解析することができない非表層的な意味（例えば隠語等）の解析についても実現することにより ISP 等を支援する技術を開発することが必要となる。

深い言語解析において、形態素解析技術や、構文解析技術は必須の基盤技術となる。形態素解析技術、構文解析技術については、これまで新聞記事等のような整った文章を対象として開発が行われており、新聞記事を対象とした場合には、形態素解析 99%、構文解析 90%と高い精度が実現されている。しかしながら、ブログや掲示板といった CGM における文章を対象とした場合には、解析精度が低下する。特に、本研究テーマが対象とする違法・有害情報では、口語的な表現や視覚的な言語表現等が用いられることも多く、従来の形態素解析技術、構文解析技術だけでは十分に解析することができない。

自然言語処理技術を活用し、違法・有害情報の自動検出を目指した研究として、国内では、有害情報として判断される情報とそうでない情報とを学習データとして、単語の有害度合いを示す重みを自動学習し、Web 上のコンテンツに含まれる単語の重みの総和によって、有害性を判別するフィルタリングソフトに関する開発事例がある。また、インターネット上の投稿サイトを対象に、人手で不適切であると判定し削除された投稿と削除されていない投稿からなる学習コーパスを作成し、SVM による学習を通じて不適切な投稿の検出を実現することを目指した研究事例がある。

海外では、EU 委員会の Safer Internet Action Plan の中で、POESIA（Public, Open-source Environment for Safer Internet Access）プロジェクトが実施され、英語、イタリア語、スペイン語の 3 言語を対象に、Web 上のテキスト情報がポルノグラフィに該当するかどうかを、自然言語処理技術を用いて分類する取組みがなされた。分類は統計モデルを利用した Light Filter とキーワードの前後 3 単語による文脈パターンに基づく

Heavy Filter との組み合わせにより行われている。また、Safer Internet Action Plan では、NETPROTECTⅡプロジェクトにおいて、有害情報に関するテキスト分類器の開発が行われた。有害情報として、ポルノグラフィ、爆弾製造、薬物、暴力を取り上げ、それぞれに関連するページと、関連しないページを訓練データとして利用し、SVM を利用した機械学習によるテキスト分類が行われている。

また、韓国では、ETRI（韓国電子通信研究院）が有害情報か非有害情報かラベル化を行った評価データセット（EHDS-2000）を作成している。全北大学では EHDS-2000 を使い、NG ワードと特定の共起語によるフィルタリングと SVM を利用した機械学習によるフィルタリングとを組み合わせることにより、有害情報か非有害情報かを分類する研究を行っている。

しかしながら、何れの研究事例においても、前後の文脈に関する深い言語解析を行うことによって、違法・有害情報の候補を検出するといった技術は開発されていない。また、日本の違法・有害情報に特徴的な、口語的かつ視覚的な言語表現についても解析対象として含めた技術開発は行われていない。

２）研究開発の必要性

深い言語解析等を活用することによって、検出効率の高い技術を開発し、現在の監視業務の効率化を支援する必要がある。しかしながら、深い言語解析等を活用した研究開発には、大量の文書の処理技術や大規模計算機環境等を要するとともに、比較的長期にわたる研究開発を行っていく必要がある。また、これらの研究開発は、一般に多額の開発費用を要する等リスクが非常に高いため、大学や民間企業が単独で実施することは困難である。違法・有害情報対策について社会全体として迅速かつ的確な対応が求められている中で産学官の力を結集し早急に研究開発を行うことが必要である。

本研究開発を通じて、違法・有害情報の検出を効率化する技術を開発することによって、ISP 等の検出に係る人的な負担が削減される。また、研究開発成果を活用することにより、これまでは資金面や人材面の制約などから十分な対応を図ることが困難であった事業者等における自主的な対応を促進すること等が期待され、インターネット上の情報流通の適正化を図ることができる。

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

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、平成 18 年度から「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（情報信頼性プロジェクト）」を推進している。クローリングによる日本語の 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。情報信頼性プロジェクトでは、情報内容の信頼性、情報発信者の信頼性、情報外観の信頼性の 3 つの観点から人間の情報信頼

性評価の支援を行うために、情報の解析・組織化を行うシステム「WISDOM」の構築を行っている。これまでに情報発信者の解析、社会的評価の一部について自動分析を実現している。

また、情報通信研究機構では、平成 19 年度から平成 22 年度にかけて「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」を実施している。情報信頼性プロジェクトの強化・補完を図る研究開発として、Web コンテンツの分析技術（クロスメディア分析）、意味内容の時系列分析技術に関する研究開発が行われている。

CGM で利用される口語的な日本語表現に関する分析技術の研究開発がいくつか取り組まれている。情報大航海プロジェクトでは、CGM で利用される口語的な日本語表現に関する技術開発として「CGM 理解のための日本語解析基盤」を行っている。CGM から収集した自然言語データベースを用いて形態素辞書、構文解析辞書のチューニングを行い、形態素解析や構文解析の解析精度の向上がなされている。また、機能語に着目し、口語的な日本語表現を解析する技術を開発し、CGM での商品等の評価を分析するサービスを提供している民間企業もある。

但し、何れの研究においてもインターネット上にある役に立つ信頼性の高い情報を獲得することを目的とした研究開発であり、本研究テーマで対象とする違法・有害情報のように、隠語や併せ字などを用いたくだけた表現の解析等を行うことは目的としていない。

その他、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 ページを構成する文字及び画像の特徴値をもとに「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。但し、基本的にはキーワードベースの検出を目的として、キーワードの拡張、コンパラブルコーパスを利用した語義の曖昧性解消等の技術開発を行ったものであり、文脈の意味を解析して文章単位で違法・有害情報の候補を検出するといった深い言語解析を行うものとはなっていない。

違法・有害情報検出に関連する民間企業による研究開発では、インターネット上から収集したデータに対して、大量の NG パターンについて高速にパターンマッチングすることを可能にするアルゴリズムの開発やベイジアンフィルターを応用する研究等が行われているが、基本的にはキーワードベースの検出を目的としたものであり、本研究テーマで開発を行う深い言語解析を行うものではない。

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。

５）期待される波及効果

本研究テーマを通じて、違法・有害情報で利用されるくだけた表現にも対応した頑健な自然言語解析技術が開発されることにより、ブログや掲示板などの CGM における表現等の解析に応用することが可能となる。今後、重要性が増す CGM の意味解析等の研究開発にも資するものと考えられる。

違法・有害情報の候補の検出精度を高める技術を確立することによって、ISP 等の検出に係る人的な負担が軽減される。これにより、今後爆発的に情報が増加していく中でも、違法・有害情報への適切な対処が可能となり、安心安全な情報の流通を促進することができる。安心安全なインターネットの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。