様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 １５ 日現在

機関番号：３２６８９
研究種目：基盤研究(B)（特設分野研究）
研究期間：2017～2021
課題番号：１７ＫＴ００８５

研究課題名（和文）100億Webページ収集に基づくWebコンテンツの信頼性解析

研究課題名（英文）Credibility Analysis of Web contents based on 10 billion Web pages

研究代表者
山名　早人（YAMANA, HAYATO）

早稲田大学・理工学術院・教授

研究者番号：４０２３０５０２

交付決定額（研究期間全体）：（直接経費）　　14,300,000円

研究成果の概要（和文）：Webコンテンツの信頼性解析を目指して，Webページ収集クローラ（収集プログラム），Webページコンテンツ解析手法，Webコンテンツにアクセスすることなくコンテンツの安全性を推定する手法，従来のベンチマークの問題点の明確化と人間の判断思考に沿った信憑性解析，本分野研究活性化のためのWebページ信頼性解析研究のサーベイ公開に取り組んだ．特に，クローラでは従来手法に比較して10%の効率化，URLのみを用いて信憑性を判定できる仕組み（精度99.4%を達成）では，コンテンツへのアクセスをせずにURLのみでの判定を可能としたことから今後の実用化に向けて大きな成果を得ることができた．

研究成果の学術的意義や社会的意義
日々の暮らしに必要不可欠な存在となったWebコンテンツについて，その信頼性を判定する指標（判定手法）を考案することで，今後さらに巧妙となってくる信憑性・信頼性が低いWebコンテンツを自動判定する仕組みを構築することができた．構築された基盤技術を用いて今後ツールを構築していくことで，インターネット利用者が安心してWebコンテンツを利活用できる基盤を築くことができた．さらに，本分野の研究において欠くことのできないベンチマークの問題点を明らかにし，今後の本分野の研究のあり方を提言することができた．

研究成果の概要（英文）：In this research project, efficient web page crawlers (gathering programs), web page content analysis methods, methods for estimating web content reliability without accessing web contents (i.e., using only URLs), revealing the problems of previous benchmarks where the ground truth is usually based on human first-impression decisions, and distributing the related research survey of web content reliability have been completed. Especially, the crawler achieved a 10% improvement in efficiency compared to previous methods, and the method that can judge credibility using only URLs (achieving an accuracy of 99.4%) achieved significant results for future practical use, as it can judge credibility using only URLs without accessing content.

研究分野：ビッグデータ解析

キーワード：Webコンテンツ　信憑性　信頼性　フィッシング　Webクローラ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究開始当初（2016.10 時点）における全世界の Web サイト数は 14 億を超え，アクセス可能な Web ページ数は約 900 億ページを超えていた．しかし，Web コンテンツは玉石混合であり，検索エンジンのみならず SNS やフィッシングメール等からの誘導により，利用者が危険な Web サイトに誘導される事例が増加していた．Web コンテンツの信頼性に関する従来の研究は，「コンテンツ自体」と「コンテンツにたどり着くための検索エンジン」の信頼性の両面から研究が実施されていたが，コンテンツ自体にアクセスするとマルウェア等への感染の危険性があり，安心して Web を利用できる新しい仕組みが必要とされた．

２．研究の目的

本研究では，日々の暮らしに必要不可欠な存在となった Web コンテンツについて，その信頼性を判定する指標を考案し，コンテンツに対する信頼性を示すと共に，安心して Web を利用できる環境を提供することを目指した．本研究では，　100 億程度の Web ページを収集可能な Web クローラを構築し，実際に Web ページを収集すると共に，Web ページに紐付く特徴量（URL 構造，Web ページ構造，Web ページ執筆著者数）などの手法により Web コンテンツの信頼性を判定すると共に信頼性判定のためのベンチマークを公開することを目的とした．

３．研究の方法

本研究開発は，以下の５項目で実施する．

（１）Web ページ収集クローラ（収集プログラム）の研究開発

大規模に Web ページを収集するためには，効率的な Web クローラ（Web ページ収集プログラム）が必要となる．本研究実施者の研究室でこれまでに開発した Web クローラ構築のノウハウを活かし，収集効率の高い Web クローラの研究開発を行う．特に，Web サーバの負荷やネットワーク負荷によって，収集中の収集効率が常に変化することに着目し，「Web サーバに大きな負荷をかけず，収集時の収集速度や収集対象コンテンツに応じて収集を効率化する仕組み」の研究開発を行う．

（２）Web ページコンテンツ解析手法の研究開発

Web ページコンテンツ解析手法においては，①時系列データ解析，②著者特徴量の抽出，③Web ページ構造解析の３分野での研究開発を進める．まず，SNS 等の時系列で頻繁に変化するデータに対しては，時系列での解析が必要となることから，時系列データ解析を進める．次に，Web コンテンツの信頼性は，同コンテンツを執筆した著者の信頼性に依存するという仮説のもと，著者特徴量を抽出する手法に取り組む．最後に，信頼性や信憑性の低い Web ページは，一般に収益を目的とした Web ページである場合が多いことに着目し，Web ページの構造に着目し，広告と本来のコンテンツの配置から，Web ページの信頼性判定を行う手法に取り組む．

（３）Web コンテンツにアクセスすることなくコンテンツの安全性を推定する手法の研究開発

信頼性の低い Web コンテンツの中にはマルウェア等を拡散するものも含まれる．利用者の安全を守るためには，上記（２）の Web ページコンテンツを解析するだけでは不十分（当該 Web ページにアクセスすること自体が脅威となるため）であり，Web コンテンツにアクセスすることなくコンテンツの安全性を推定する手法が重要となる．そこで，Web コンテンツに頼らない信頼性判定の仕組みについて，Web ページへたどり着くためのアドレスである URL のみを用いてフィッシングサイト等の危険なサイトを検知する仕組みの研究開発を行う．

（４）従来のベンチマークの問題点の明確化と人間の判断思考に沿った信憑性解析

本研究項目は，当初計画時点では予定していなかったものである．（１）～（３）の研究を実施中に，本研究分野で一般的に利用されているベンチマークの正しさに疑義を感じたことをきっかけとして，新たに実施するものである．具体的には，従来の Web ページ信頼性判定に用いられるベンチマークの問題点を明らかにすると共に，その原因を明らかにすることを目指す．

（５）Web ページの信頼性解析研究のサーベイの公開

本分野の研究は，本研究開発だけで完了するものではなく，多くの研究者が取り組むことによって，継続的に現れる新たな脅威に対処していくことが必要となる．そこで，本研究の一つの目標である本分野での研究をさらに活性化させることを目指して，本研究分野での最新サーベイを行い広く一般に公開する．

４．研究成果

上記「研究の方法」で示した５項目に対する研究成果を示す．なお，主要論文発表先は本成果内でも示す．

（1）Web ページ収集クローラ（収集プログラム）の研究開発

Web コンテンツを効率的に収集するための手法として，「History-enhanced Focused Website Segment Crawler」を提案・実装した（2018 International Conference on Information Networking（ICOIN)で発表．Journal of New Generation Computing で発表)．具体的には，Web ページ収集中の挙動（ページ収集効率の変化，目的とする収集カテゴリの割合）をリアルタイムに解析し，収集対象の Web コンテンツに対して動的に収集優先度を更新することで，従来の Best-First クローラに比較し約 10%の収集効率向上を実現した．

（2）Web ページコンテンツ解析手法の研究開発

①時系列データ解析

時系列でのコンテンツ変化を捉えるために，「A Variable-Length Motifs Discovery Method in Time Series using Hybrid Approach」を提案・実装した（19th International Conference on Information Integration and Web-based Applications & Services で発表．International Journal of Web Information Systems で発表.)．本手法は，時系列データを文字情報にマッピングすることで解析のための計算量を抑えるだけでなく，類似する時系列を持つデータを抽出するための特徴的パターン（Motifs）を抽出できる．本手法を用いることで，類似するコンテンツの効率的な検出が可能となった．

②著者特徴量の抽出

Web コンテンツを執筆した著者の特徴から Web ページコンテンツ解析を行う方法として，コンテンツ内さらに，Web コンテンツを分類するための手法として，「単語重要度 CrRv」を提案し，著者専門性推定を可能とした．これにより信憑性が問題となる SNS のように短い文章にも対応できる．さらに，著者人数推定に基づく Web ページの信頼性判定に取り組んだ．これは，一般的に多くの人が共同して書いた文章の信頼性は高くなるという事実に基づいている．具体的には，文章をスライディングウィンドウにより分割し，単一ウィンドウは１名の著者で書かれたという前提のもと，前後のスライディングウィンドウとの類似度（n-gram，係り受け等の特徴量を利用）の変化により著者を推定した．当該分割点からの距離（文字数）によって特徴量の重みを指数関数的に変化させ精度向上を行った． １人によって記述された文章に対する執筆者数推定の正解率 56.3%，２人の場合 81.8%，３人の場合 74.8%，４人の場合 65%を達成した．

③Web ページ構造解析

Web ページの構造に着目することで，Web ページの信憑性度合の判定を行う手法を開発・実装した（IEEE International Conference on BigData 2018 で発表)．本手法では「信憑性や信頼性が低い Web ページは主に収益を目的としている点」に着目し，オンライン広告とコンテンツそのものの Web ページ内での配置を用いることで，従来のコンテンツ信憑性判定を精度を向上（71.1%から 74.5%への改善）できることを確認した．

（3）Web コンテンツにアクセスすることなくコンテンツの安全性を推定する手法の研究開発

URL の文字列に含まれる情報のみを用いて当該 Web ページの信頼性を判定する手法に取り組んだ．研究にあたっては，本分野に関連する全論文を調査し，従来利用されている判定のための特徴量を採用するだけでなく，以下の表１に示す特徴量を最終的に採用し，図１に示す深層学習の仕組みを提案し「Hybrid Phishing URL Detection Using Segmented Word Embedding」手法を完成させた（The 21st International Conference on Information Integration and Web-based Applications & Services で発表．The 17th International Conference on Availability, Reliability and Security 投稿中)．特徴的な点は，ワード単位（URL 全体，URL 内ドメイン部），

表１　採用した特徴量

Feature Name	Definition	Location
Feat_tld	Top-level domain name	Domain
Feat_ip	IP address	Domain
Feat_sus_pattern	Occurance of suspicious words	Domain
Feat_long_word	Ratio of the longest word length to the domain length	Domain
Feat_max_letter	Max. length of consec. letter	Domain
Feat_max_digit	Max. length of consec. digit	Domain
Feat_max_symbol	Max. length of consec. symbol	Domain
Feat_min_letter	Min. length of consec. letter	Domain
Feat_min_digit	Min. length of consec. digit	Domain
Feat_min_symbol	Min. length of consec. symbol	Domain
Feat_max_letter	Max. length of consec. letter	Path
Feat_max_digit	Max. length of consec. digit	Path
Feat_max_symbol	Max. length of consec. symbol	Path
Feat_min_letter	Min. length of consec. letter	Path
Feat_min_digit	Min. length of consec. digit	Path
Feat_min_symbol	Min. length of consec. symbol	Path
Feat_dnan	Non-alphanumeric character	Domain
Feat_pnan	Non-alphanumeric character	Path
Feat_at	Inc. "@" redirection or not	Path
Feat_dslash	Inc. "//" redirection or not	Path
Feat_link	Inc. embedded link or not	Query
Feat_subdomain	# of subdomains	Subdomain
Feat_sp	# of special characters	Subdomain
Feat_len_subdomain	Length of subdomains	Subdomain

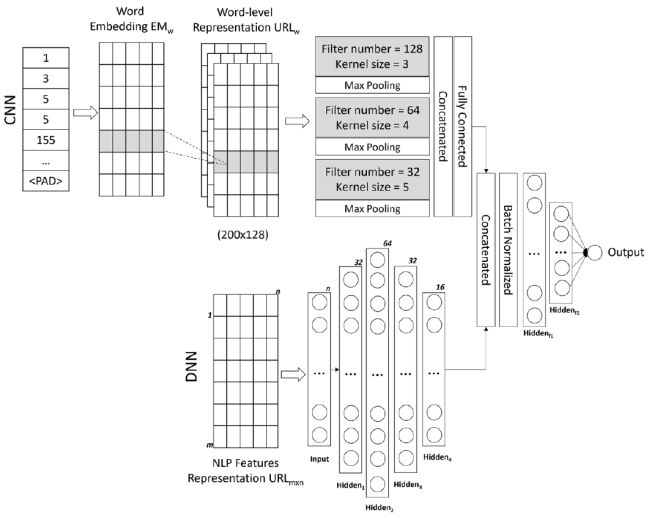

図１　信頼性判定アーキテクチャ

文字単位(URL 内パス部)での解析を統合，さらに自然言語処理で用いられる各種統計特徴量(表１)を統合した点にある．最終的に 99.2%(本研究で収集したインバランスデータセット)，99.4%（従来他論文で利用されているバランスデータセット）の正解率を達成し，従来手法を超える正解率を確認し，実用性の向上を果たした．

(４)従来のベンチマークの問題点の明確化と人間の判断思考に沿った信憑性解析

信憑性判定のためのベンチマーク（例：Fake News Corpus）には，Web ページの URL とその判定結果（信頼性有無）が保存されており，多くの研究が同ベンチマークを用いてその検出精度を競っている．しかし，同ベンチマークで正解とされる判定結果（信頼性有無）は，人手による判定（多くは 3 名以上の判定者による多数決により決定）である．このため，同判定が真に正しいかどうかは疑問である．そこで，（１）により独自に収集した Web ページを対象に人手による判定として「第一印象での判断」と「熟考後の判断」の二種類の判定を行った．その結果，「第一印象」と「熟考後」の判断での一致度は 37.5%に留まった．次に，第一印象に近い特徴量のみを用いて従来のベンチマークに対して信頼性判定（Google が Web ページ評価のために公開している Google Lighthouse 特徴量を利用）したところ，従来手法の精度が 73.7%～81.1%に対して，89.8%を達成することができた．以上から，人間が行っている信頼性判定をコンピュータで実現する場合は，人間が「第一印象で感じる特徴量」を中心とした信頼性判定が求められることが分かった．一方，本来の意味で信頼できるかどうかを判定するためには，これまでのベンチマークの精査が必要であり，ベンチマーク作成において人手による判定を行う場合は，熟考後の判定が必須であることが分かった(IEEE International Conference on BigData 2019 で発表．The 14th International Symposium on Mining and Web で発表)．

(５)Web ページの信頼性解析研究のサーベイの公開

Web ページの信頼性解析研究の最新サーベイを電子情報通信学会論文誌（IEICE Transactions on Information and Systems）に発表した．

以上，2017 年度～2021 年度の 5 年間，Web コンテンツの信頼性解析に取り組み，それぞれの分野で成果を上げた．特に，従来の本分野で用いられているベンチマークは必ずしもその判定結果が正しいとは判断できず，我々が構築・公開したベンチマークセットを含め，より信頼性のおけるベンチマークが公開されることが望まれる．また，本研究期間の中でも長期間にわたり研究を実施した URL のみからの信頼性判定は，その利便性が大きく(判定のための計算機資源コストが低く，かつ，危険なサイトにアクセスする前に判定可能であるため)，今後ブラウザへのプラグインとしての提供等を含め，社会への還元に結び付けていきたい．

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Kenta Yamada, Hayato Yamana	４．巻 1
２．論文標題 Effectiveness of Usability & Performance Features for Web Credibility Evaluation	５．発行年 2019年
３．雑誌名 Proc. of IEEE BigData 2019	６．最初と最後の頁 6257-6259
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BigData47090.2019.9006419	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tanaphol Suebchua, Bundit Manaskasemsak, Arnon Rungsawang, Hayato Yamana	４．巻 36-2
２．論文標題 Efficient Topical Focused Crawling Through Neighborhood Feature	５．発行年 2018年
３．雑誌名 New Generation Computing	６．最初と最後の頁 95-118
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00354-017-0029-8	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Chaw Zan, Hayato YAMANA	４．巻 -
２．論文標題 A Variable-Length Motifs Discovery Method in Time Series using Hybrid Approach	５．発行年 2017年
３．雑誌名 Proc. of the 19th International Conference on Information Integration and Web-based Applications & Services	６．最初と最後の頁 49-57
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3151759.3151781	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tanaphol Suebchua, Bundit Manaskasemsak, Arnon Rungsawang, Hayato YAMANA	４．巻 -
２．論文標題 History-enhanced Focused Website Segment Crawler	５．発行年 2018年
３．雑誌名 Proc. of IEEE the 32nd International Conference on Information Networking	６．最初と最後の頁 80-85
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICOIN.2018.8343090	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Ken MISHIMA, Hayato YAMANA	4．巻 E105-D, No.7
2．論文標題 A Survey on Explainable Fake News Detection	5．発行年 2022年
3．雑誌名 IEICE Transactions on Information and Systems	6．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2021EDR0003	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Ent Sandi Aung, Hayato YAMANA	4．巻 1
2．論文標題 Segmentation-based Phishing URL Detection	5．発行年 2021年
3．雑誌名 Proceedings of WI-IAT '21: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology	6．最初と最後の頁 550-556
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3486622.3493983	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計11件（うち招待講演　0件／うち国際学会　0件）

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
タイ	カセサート大学			