

「犯罪」分野に関連する日英ブログの収集および類型化

中崎 寛之[†] 阿部 佑亮^{††} 横本 大輔^{††} 宇津呂武仁[†] 河田 容英^{†††}

福原 知宏^{††††}

[†] 筑波大学大学院システム情報工学研究科 〒305-8573 茨城県つくば市天王台 1-1-1

^{††} 筑波大学第三学群工学システム学類 〒305-8573 茨城県つくば市天王台 1-1-1

^{†††} (株) ナビックス 〒141-0031 東京都品川区西五反田 8-3-6

^{††††} 東京大学人工物工学研究センター 〒277-8568 千葉県柏市柏の葉 5-1-5

あらまし 我々は、これまで、ブログ空間中に存在する有用な情報や知識に的確にアクセスし利用するという目的のもとで、体系化された知識体系である Wikipedia とブログサイトを対応づける研究を行ってきた。しかし、同一のトピックについて記述しているブログであっても、ブロガーの立場や環境は大いに異なることがわかった。そこで、本研究では、ブログ空間の情報や知識を類型化するための方式の一つとして、「ブロガーの立場」に着目してブログサイト・ブログ記事を類型化するというアプローチをとる。具体的には、本稿では、特定トピックについてのブログに含まれる情報の有用性を検証するための事例研究として、「詐欺」、「インターネット犯罪」の分野を対象として、日英ブログサイトおよびブログ記事の収集を行い、ブログでの記述内容を被害者・ニュース記事引用・防止対策に類型化した結果を報告する。日英各言語のブログサイト・ブログ記事の収集には、ある同一トピックについて詳しい記述をしているブログサイトを検索する手法と、検索エンジンによるブログ記事検索手法を組み合わせ用いた。そして、収集したブログサイト・ブログ記事にはどのようなタイプの立場があるかを分析し、それらを類型化した。

キーワード 二言語ブログ分析, ブログサイト検索, 犯罪レポート, Wikipedia

Collecting and Categorizing Japanese/English Blogs related to “Crime” Domain

Hiroyuki NAKASAKI[†], Yusuke ABE^{††}, Daisuke YOKOMOTO^{††}, Takehito UTSURO[†], Yasuhide KAWADA^{†††}, and Tomohiro FUKUHARA^{††††}

[†] Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan

^{††} College of Eng. Sys., Third Cluster of Colleges, University of Tsukuba, Tsukuba, 305-8573, Japan

^{†††} Navix Co., Ltd. 8-3-6 Nishi-Gotanda, Shinagawa-Ku Tokyo 141-0031, Japan

^{††††} Research into Artifacts, Center for Engineering, University of Tokyo, Kashiwa, Chiba 277-8568, Japan

Abstract Among other domains and topics on which some issues are frequently argued in the blogosphere, the domain of crime is one of the most seriously discussed by various kinds of bloggers. Such information on crimes in blogs is especially valuable for outsiders from abroad who are not familiar with cultures and crimes in foreign countries. This paper proposes a multilingual framework of categorizing people's concerns, reports, and experiences on crimes in their own blogs. First, we refer to *Wikipedia* as a multilingual terminological knowledge base, and search for Wikipedia entries describing criminal acts. In the retrieval of blog feeds/posts, we take two approaches, focusing on various types of bloggers such as experts in the crime domain and victims of criminal acts. We further categorize the retrieved blog feeds/posts into four types including experts in the crime domain and victims of criminal acts.

Key words cross-lingual blog analysis, blog feed retrieval, crime report, Wikipedia

1. はじめに

近年、世界中でブログサービスやブログツールが普及し、各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった。それに伴い、さまざまな情報がブログに記載され、商用ブログ検索サービスを利用することでそれらの情報を取得することができるようになった。ここで、これらの既存のブログ検索サービスは、ブログ空間に対する索引付けの粒度と体系化の二点において不十分であると言える。まず、カテゴリ式のブログ検索サービスにおいては、人手により設定されたカテゴリの体系が十分な網羅性を持つとは言えず、また、実際の検索要求に比べて、カテゴリの粒度が粗すぎる傾向がある。一方、キーワードや評判、時系列変化などによるブログ検索サービスの場合は、個々の索引の粒度が細かく、また、それらの索引全体を体系化してとらえることが困難である。したがって、利用者が、検索要求に対して適切な索引を想起することができなければ、巨大なブログ空間に対して容易にはアクセスできない。

そこで、我々は、ブログ空間への効率的なアクセスを実現するにあたって、より適切な粒度で、十分に体系化された索引付けの一つの方式として、あらゆる事柄が詳細に体系化された知識体系である Wikipedia とブログサイトを対応づけた [1]。これにより、ブログ空間に対して索引付けが行われ、ブログ空間におけるトピック分布を推定することができるようになった。

しかし、Wikipedia とブログサイトを対応づけただけでは、様々な立場のブロガーによって書かれたブログサイトが同一トピック内で混在していたことも明らかになった。例えば、トピック「オークション詐欺」では、父親の被害経験について記述しているブロガーもいれば、オークション出品者の立場から詐欺にあわないための対策法を紹介しているブロガーもいる。このように、同じ「オークション詐欺」について記述しているブログサイトでもブロガーの立場や環境が大いに異なる。そこで、我々は詳しい記述をしているブログサイトをより細かく分類するために、ブログサイト群を「ブロガーの立場」で類型化する。具体的には、ブログ空間中で頻繁に議論され、かつ「ブロガーの立場」がはっきりと分かれているという理由から、まずは犯罪分野のトピックを対象としたブログの類型化を行う [2]。犯罪分野の立場として、犯罪行為の被害者、犯罪行為の報道記事を引用しているブログ、犯罪行為に対する対策の仕方について紹介しているブログなどに分類された。

本研究の全体的枠組みを図 1 に示す。まず、本研究では、日英 Wikipedia ^(注1) から犯罪行為の事例として「詐欺」カテゴリと「インターネット犯罪」カテゴリを選定した。この二つのカテゴリ下に属する犯罪行為の日英 Wikipedia エントリのタイトルをトピック名とした。次に、トピック名を検索語としてブログサイト・ブログ記事を検索し順位付けする。そして、上記の手法で検索された日英両言語のブログサイト・ブログ記事集合を、トピックに対する関心や意見の文化間差異の発見支援

をするシステム [3] に適用する。このシステムの本来の目的は、日英ブログから抽出した共起語単位で日英ブログの文化間差異の発見を支援することであるが、本研究にこのシステムを適用することで、各地域における犯罪行為の文化間差異の発見支援につながる。例えば、近年日本では「振り込め詐欺」や「おれおれ詐欺」が多数発生しており、大きな話題として取り上げられているが、これは日本特有の犯罪であるため、欧米ではほとんど語られていない。

また、本稿の 2. では、本研究で用いる検索手法の有効性を示すために、あるトピックについて記述しているブログサイトを対象として、Wikipedia 関連語によるスコア付けを用いる手法 [4], [5]、および、機械学習を用いる手法 [1] で順位付けした結果の人手評価について記述する。比較対象として、既存の検索エンジン API によるブログ記事の順位付けを変更せずに、ブログサイト単位にまとめた順位付けをベースラインとして用いた。

さらに、本稿の 4. で、「犯罪」ドメインである「詐欺」カテゴリおよび「インターネット犯罪」カテゴリに属するトピックに関するブログの類型化 [2] について説明する。本研究では類型化の対象となるブログを収集するために、2 種類の検索手法を用いた。一つ目は Wikipedia から抽出した関連語を用いたブログサイト順位付け、二つ目は検索エンジンによるブログ記事順位付けである。また、検索エンジンによるブログ記事順位付けにおいては、分類するブログのタイプにあわせてブログ記事を収集するクエリを複数設計した。

2. Wikipedia エントリに対応するブログサイト順位付け

2.1 評価手順

本研究では、検索の有効性を評価するために「アルコール依存症」、「リストラ」、「著作権侵害」の 3 トピックの日本語ブログサイトを対象として評価実験を行った。対象トピックのブログサイトに対して、Wikipedia 関連語によるスコア付けを用いる手法 [4], [5]、および、機械学習を用いる手法 [1] で順位付けした結果の人手評価を行った。具体的には、各トピックを対象として順位付けされた日本語ブログサイトのうち、それぞれ、上位 20 ブログサイト、および、100 位までの 20 ブログサイトを等間隔にサンプリングした合計 40 ブログサイトを手動で評価した。評価結果としては、「当該トピックについて詳細な記述を含むか否か」の判定を付与した。また、比較対象として、検索エンジン API によるブログ記事の順位付けを変更せずに、ブログサイト単位にまとめた順位付けに対しても、同様に合計 40 ブログサイトを手動で評価した。

2.2 評価結果

評価結果を比較するために、横軸に各ブログサイトの順位をとり、その順位までのブログサイトをすべて「当該トピックについて詳細な記述を含む」と自動判定した場合の精度を縦軸にとって、その推移をプロットしたものを図 2 の「Wikipedia 関連語によるスコア付けを用いた順位付け」、「機械学習を用いた順位付け」、および「ベースライン」に示す。ただし、このプロットには、3 トピック分を平均した結果を示す。この結果が

(注1) : <http://en.ja.wikipedia.org/>.

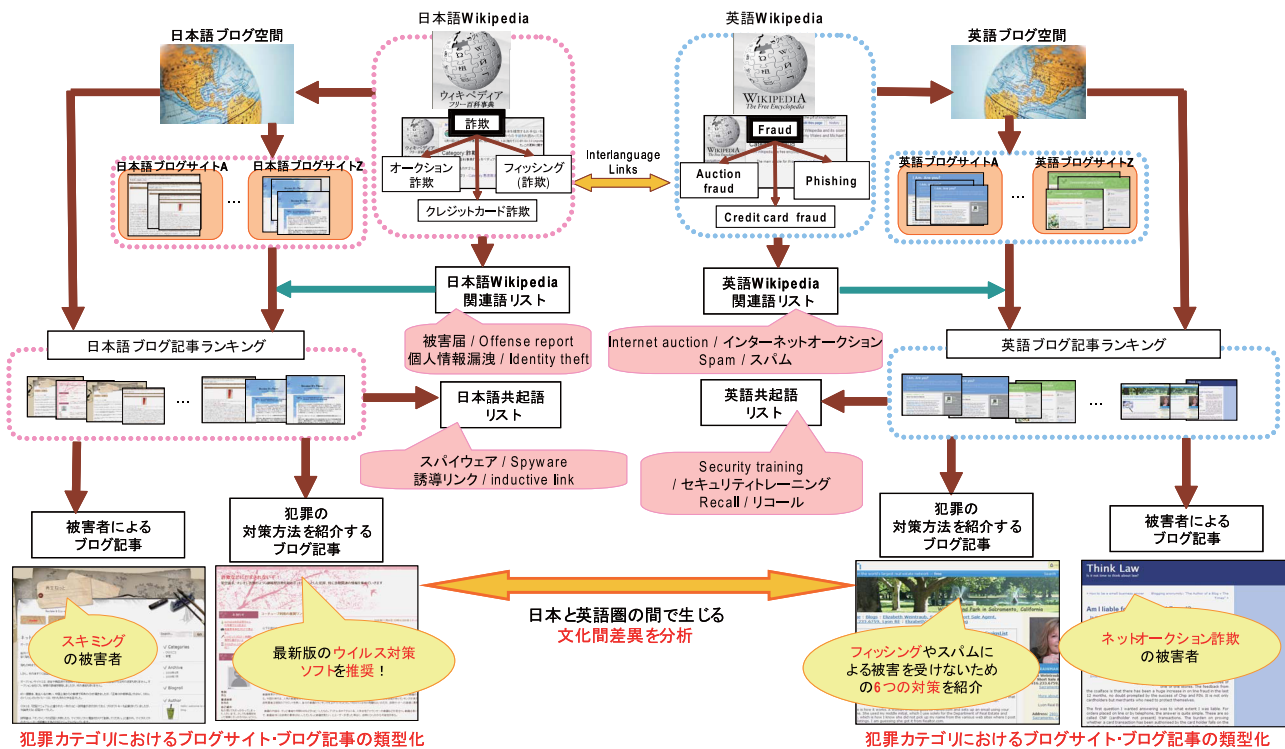


図1 二言語対照ブログ分析およびブログサイト・ブログ記事の類型化の全体的枠組み

ら分かるように、「Wikipedia 関連語によるスコア付け」方式、「機械学習」方式のどちらにおいても、ベースラインの精度を大幅に改善することが分かる。

また、図3には、横軸に各ブログサイトの順位をとり、「トピックについて詳細な記述を含むブログサイトの数」(3トピック分)の推移を、「Wikipedia 関連語によるスコア付け」方式とベースラインの間、「機械学習」方式とベースラインの間、および「Wikipedia 関連語によるスコア付け」方式と「機械学習」方式の間でそれぞれ比較した。具体的には、各順位までで、両手法によって共通に出力されたブログサイト数、片方の手法によってのみ出力されたブログサイト数(方式ごとに1プロットずつ)の比較を行った。

この結果から明らかなように、「Wikipedia 関連語によるスコア付け」方式とベースラインの比較では、「Wikipedia 関連語によるスコア付け」方式のみによって出力されたブログサイト数は、ベースラインのみによって出力されたブログサイト数よりもはるかに多い。同様に、「機械学習」方式とベースラインの比較でも、「機械学習」方式のみによって出力されたブログサイト数は、ベースラインのみによって出力されたブログサイト数よりもはるかに多い。これにより、「Wikipedia 関連語によるスコア付け」方式、「機械学習」方式は、「トピックについて詳細な記述を含むブログサイト」のうち、既存の検索エンジンAPIにおいて下位に順位付けされたサイトを上位に押し上げることが分かる。また、「Wikipedia 関連語によるスコア付け」方式と「機械学習」方式の比較では、順位付けの上位においては「機械学習」方式の精度が高く、順位付けの下位においては「Wikipedia 関連語によるスコア付け」方式の精度が高い。しかし両方式の差は小さく、共通して出力されたブログサイトの

数も多い。

今後の課題としては、「Wikipedia 関連語によるスコア付け」方式で考慮している情報を機械学習の素性として用いることで、「機械学習」方式の性能を改善することが挙げられる。本論文で用いた手法においては、「Wikipedia 関連語によるスコア付け」方式では Wikipedia 関連語の出現回数を直接考慮しているのに対して、「機械学習」方式の素性としては、出現回数を5段階の範囲に分けて、各範囲に該当するか否かを表す二値の素性を用いている。また、「Wikipedia 関連語によるスコア付け」方式では Wikipedia 関連語の種類に応じて重みを付与しているが、「機械学習」方式では関連語の種類は考慮されていない。一方で、「機械学習」方式でのみ考慮されている素性としては、Wikipedia 関連語のブログ空間におけるヒット数や、Wikipedia 関連語の出現種類数などがある^(注2)。このような差があるにも関わらず、「Wikipedia 関連語によるスコア付け」方式と「機械学習」方式の性能に大きな差がないことから、特定トピックに詳細なブログサイトの選択的検索においては、Wikipedia 関連語の出現回数や種類は、機械学習において考慮している素性と同等に重要な情報であると考えられる。今後は、「Wikipedia 関連語によるスコア付け」方式で考慮されている情報を機械学習でも素性として用いることで、「機械学習」方式の性能をより改善したいと考えている。

(注2)：Wikipedia から収集される関連語には、各トピックとは関連が薄い一般語が含まれている。このような語はトピックに関わらず出現するため、ブログ空間でのヒット数が大きくなる。このとき、Wikipedia 関連語のブログサイト数を考慮することで、このような Wikipedia 関連語の影響を抑えることができると考えられる。

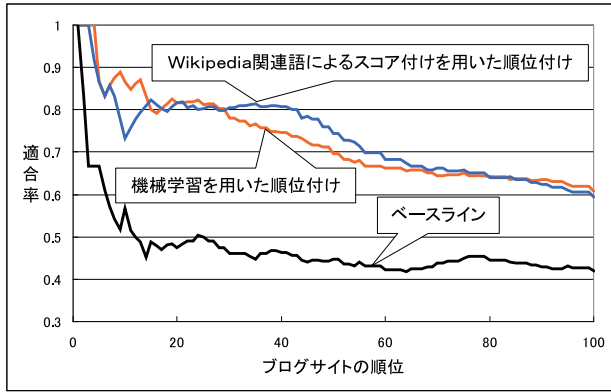


図2 特定トピックのブログサイト検索における適合率の評価

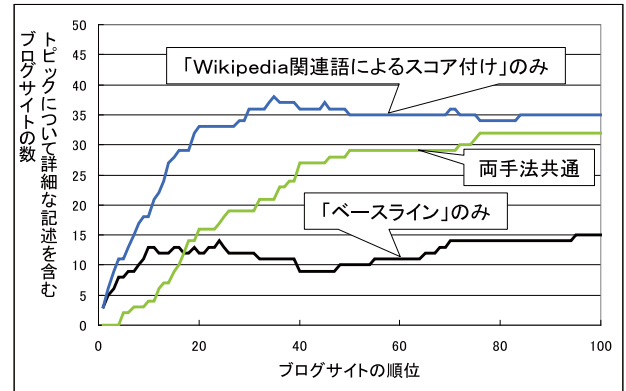
3. 「犯罪」ドメインにおける評価用カテゴリおよびトピック

本研究では、犯罪分野の事例として、「詐欺」カテゴリと「インターネット犯罪」カテゴリを選定した。まず、日英 Wikipedia の「詐欺」カテゴリおよび「インターネット犯罪」カテゴリ下に属するエントリ名を検索語として、ブログ検索ヒット数^(注3)が10,000以上のトピックを検出した。その結果、「詐欺」カテゴリでは日本語で20トピック、英語で68トピックが検出され、「インターネット犯罪」では日本語で8トピック、英語で15トピックが検出された。さらに、検出されたトピックの中から人手でトピックを選定した。その結果、「詐欺」カテゴリからは日本語で10トピック、英語で14トピックが選定され、「インターネット犯罪」カテゴリからは日本語で5トピック、英語で6トピックが選定された。ただし、いくつかのトピックについては、Wikipediaにおいて日本語もしくは英語エントリのいずれかが存在しない場合があった。例えば、トピック「クレジットカード詐欺」は英語 Wikipedia ではエントリが存在するが、日本語 Wikipedia においては「クレジットカード」エントリの一項目となっていた。他にも「オークション詐欺」と「振り込み詐欺」の英語 Wikipedia エントリが存在せず、このようなトピックについては、英辞郎^(注4)でトピック名の対訳を取得し、その対訳をトピック名として検索に用いた。また、一部のトピックでは、日英両言語の Wikipedia にエントリが存在するが、言語間リンクでつながっていないものがあった。この場合も、上記のトピックと同様に英辞郎を用いて対訳を取得して日英 Wikipedia エントリを対応付けた。

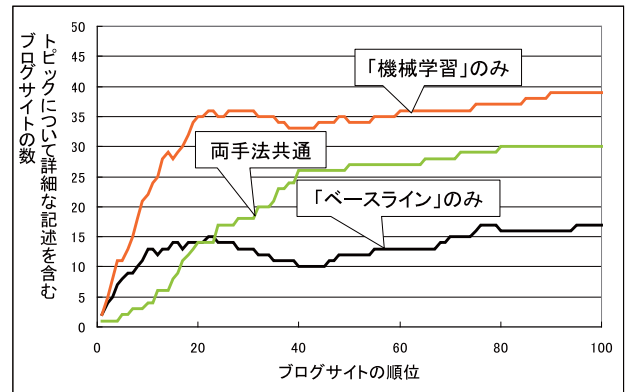
本研究で用いた「詐欺」カテゴリおよび「インターネット犯罪」カテゴリにおけるトピックの一覧を表1に示す。また、各

(注3)：日本語ブログの検索には Yahoo!Japan 検索 API(<http://www.yahoo.co.jp>)、英語ブログの検索には米 Yahoo!検索 API(<http://www.yahoo.com>)を用いた。また、日本語ブログでは大手10社(fc2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, seesaa.net, yaplog.jp, webry.info.jp, hatena.ne.jp)、英語ブログでは大手10社(blogspot.com, spaces.live.com, livejournal.com, vox.com, multiply.com, typepad.com, aol.com, blogger.com, wordpress.com, blogster.com)のブログ会社のドメインに限って検索を行った。

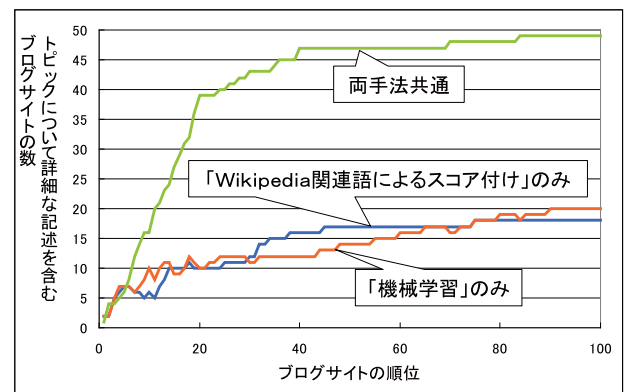
(注4)：<http://www.eijiro.jp/>



(a) 「Wikipedia 関連語によるスコア付け」と「ベースライン」の比較



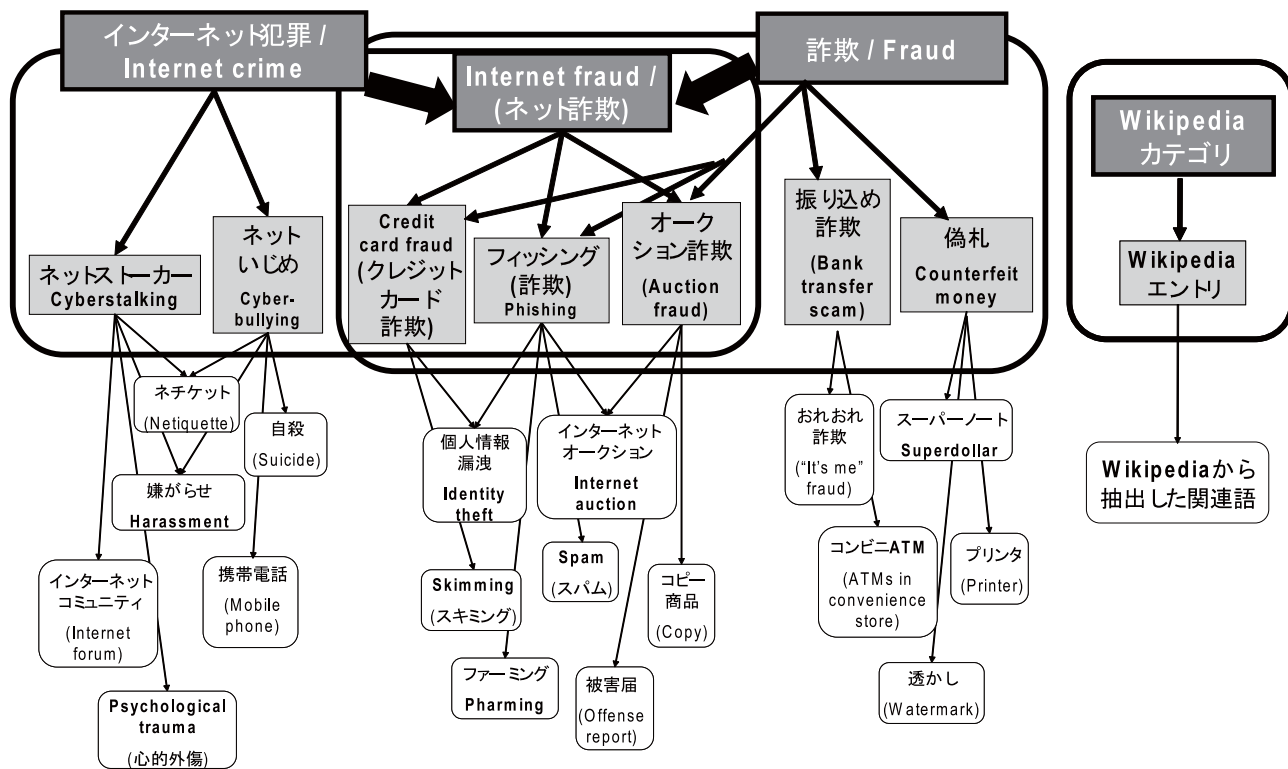
(b) 「機械学習」と「ベースライン」の比較



(c) 「Wikipedia 関連語によるスコア付け」と「機械学習」の比較

図3 特定トピックのブログサイト検索における「トピックについて詳細な記述を含むブログサイトの数」の比較

トピックの Wikipedia から抽出した関連語数、収集したブログサイト数、収集したブログサイト中で Wikipedia 関連語のいずれかが出現したブログ記事数を表中に示す。さらに、各トピックの Wikipedia エントリから抽出した関連語の例を図4に示す。



(注:「クレジットカード詐欺」、「Auction fraud」、「Bank transfer scam」は Wikipedia にエントリーがないため、対訳を用いて英辞郎でトピック名を取得した。)

図4 「詐欺」カテゴリおよび「インターネット犯罪」カテゴリにおけるトピックおよび関連語の例

表1 Wikipedia から抽出した関連語数、収集したブログサイト・ブログ記事数

トピック (日本語/英語)	Wikipedia から 抽出した関連語数 (日本語/英語)	ブログ サイト数 (日本語/英語)	ブログ 記事数 (日本語/英語)
インターネット詐欺 / Internet fraud	76 / 182	60 / 48	353 / 1576
(オークション詐欺 / Auction fraud)	36 / 24	38 / 40	121 / 224
(クレジットカード詐欺 / Credit card fraud)	181 / 28	31 / 50	143 / 1086
(フィッシング詐欺 / Phishing)	63 / 172	118 / 49	1118 / 8982
振り込み詐欺 / Bank transfer scam	96 / 60	132 / 4	2617 / 13
偽札 / Counterfeit money	84 / 175	96 / 41	695 / 186
ネットストーカー / Cyberstalking	29 / 33	39 / 49	242 / 727
ネットいじめ / Cyber-bullying	65 / 52	89 / 49	613 / 4278

4. 「犯罪」ドメインにおけるブログサイト・ブログ記事の類型化

本研究では、「犯罪」ドメインである「詐欺」カテゴリおよび「インターネット犯罪」カテゴリに属するトピックに関するブログの類型化を行った[2].「詐欺」カテゴリに属する「オークション詐欺」、「フィッシング」、「振り込み詐欺」^(注5)のそれぞれに関連するブログサイト・ブログ記事の割合、「関連語を用いたブログサイト順位付け」(以降、「手法1」と呼ぶ)と「検索エ

ンジンによるブログ記事順位付け^(注6)」(以降、「手法2」と呼ぶ)によって収集されたブログサイト・ブログ記事の重複の割合、「手法1」と「手法2」で収集されたブログサイト・ブログ記事の総数を表2に示す。

また、本研究では二つの検索手法によって収集されたブログサイトおよびブログ記事を、以下のタイプに分類した。

(1) 被害者もしくはその知人・目撃者によるブログ(未遂を含む)

(注5): 英語における類型化対象のトピックは「Auction fraud」および「Phishing」の二つである。「Bank transfer scam」については収集したブログ記事数が少なかったため、表2と表3には記載していない。

(注6): あるトピック t について順位付けの対象となるブログ記事を収集する際、「 t 」のみの他に、本節で紹介する記事のタイプのそれぞれについて、個別に選択的にブログ記事を検索するクエリを設計した。具体的には、日本語では「 t AND 被害」、「 t AND 引用」、「 t AND 対策」のクエリを設計し、英語では「 t AND victim」、「 t AND reference」、「 t AND tips」のクエリを設計した。

表 2 「詐欺」カテゴリにおける関連ブログサイト・ブログ記事の割合 (%)

トピック	トピックに関連する ブログサイト・ ブログ記事の割合		両手法で重複する 関連ブログサイト数	両手法で重複する 関連ブログ記事数	手法 1 で 収集した ブログ サイトの 総数	手法 2 で 収集した ブログ 記事の 総数
	手法 1: ブログサイト	手法 2: ブログ記事	手法 1 で収集した 関連ブログサイト数	手法 2 で収集した 関連ブログ記事数		

(a) 日本語

オークション詐欺	92.9	62.3	0	0	14	69
フィッシング詐欺	90.9	92.5	9.1	3.0	11	67
振り込め詐欺	76.9	83.6	7.7	0	13	67

(b) 英語

Auction fraud	90.0	94.0	40.0	20.9	10	67
Phishing	100	92.5	10.0	1.5	10	67

(2) 犯罪行為に関するニュース記事もしくは Web 上の他の公式サイトからの引用を用いて、警告をしているブログ

(3) 犯罪行為の被害を防ぐ方法について紹介しているブログ

(4) 該当トピックに関する記述があるが、上記の 3 タイプには分類されないブログ (例: ブログターの意見のみ記述されているブログ)

表 2 で示したトピックに関連するブログサイト・ブログ記事を対象として、これらの各タイプの割合を算出した結果を表 3 に示す。上記のタイプのうち、「被害者もしくはその知人・目撃者によるブログ」と「犯罪行為に関するニュース記事もしくは Web 上の他の公式サイトからの引用を用いて、警告をしているブログ」については、それぞれさらに二つに細分類した。「被害者もしくはその知人・目撃者によるブログ」は、「被害者自身によるブログ」と、「被害者の知人または目撃者によるブログ」の二種類に分類し、「犯罪行為に関するニュース記事もしくは Web 上の他の公式サイトからの引用を用いて、警告をしているブログ」は、「ニュース記事を引用しているブログ」と、「ニュース以外の公式サイト等を引用しているブログ」の二種類に分類した。

今回は、いずれの検索手法においても、検索クエリのトピックと関連するブログサイト・ブログ記事が多く収集された。特に、日英両言語において、被害者自身が自分の被害経験を記述しているブログが多くみられた。また、ニュースを引用している記事や対策法を紹介している記事が多く収集された。特筆すべき点として、「関連語を用いたブログサイト順位付け」によって収集されたブログサイト・ブログ記事、および、「検索エンジンによるブログ記事順位付け」によって収集されたブログサイト・ブログ記事の間で重複度合いを調査したところ、比較的重複が少ないことが挙げられる。被害者もしくはその知人・目撃者によるブログ記事は、「関連語を用いたブログサイト順位付け」よりも「検索エンジンによるブログ記事順位付け」によってより多く収集された。一方で、「関連語を用いたブログサイト順位付け」では、トピックに関してより詳しい記述をしているブログを「検索エンジンによるブログ記事順位付け」と比較し

てより多く収集できた。したがって、「関連語を用いたブログサイト順位付け」および「検索エンジンによるブログ記事順位付け」はそれぞれ目的に沿ったブログの検索をすることができたといえる。

また、ブログ記事の類型化および要約の事例を図 5 に示す。図中では、Wikipedia エントリから抽出した関連語が出現している箇所を記述した。これらの関連語が出現するブログ記事中の文を特定することで、そのブログ記事の特徴的な内容を発見できる可能性があると考えている。また、そのブログ記事の特徴的な内容を発見することができれば、そこからブログ記事の内容を表現する新たな情報として共起語をブログ本文中から抽出できると考えている。図中では、ブログ記事の内容を表現する共起語例を各記事の要約中に記述した。「振り込め詐欺」については、関連する英語ブログ記事をほとんど収集することができなかったため、日本語特有のトピックであることがわかり、カテゴリにおける文化間差異を観測できた。

今後の課題として、本稿で行ったブログサイト・ブログ記事の類型化の自動化が挙げられる。そこで、筆者らは本稿で示したブログの類型の一つである「被害者によるブログ記事」を記事文中の手がかり表現を用いて収集する手法 [6] を提案している。この手法では、「被害者によるブログ記事」を同定するために記事文中の「係り受け関係」と「文節単位の表記パターン」に着目し、これらの手がかりが「被害者によるブログ記事」の収集において有効であることを示した。

5. 関連研究

関連研究として、Web 上のページからトラブルを表す文の抽出を行っている研究 [7], [8] が挙げられる。この研究でのトラブル表現抽出技術は、我々の研究における「被害者によるブログ記事」同定においても有用な可能性がある。また、Web 上の膨大なブログから人々の経験情報を収集し、意味的に索引付けて DB 化する手法 [9] についても、今後本研究のタスクにおける適用可能性を評価する必要がある。

さらに、多言語情報源を対象とした情報分析の関連研究とし

表3 「詐欺」カテゴリにおいて収集したブログサイト・ブログ記事の類型化およびその割合 (%)

(1) 「手法 1」

トピック	(1) 被害者自身もしくはその知人・目撃者のブログ		(2) Web 上のページを引用		(3)	(4) その他
	被害者自身のブログ	被害者の知人・目撃者のブログ	ニュースを引用	公式サイト等を引用	犯罪被害の予防となる対策を紹介	

(1-a) 日本語

オークション詐欺	28.6	7.1	14.3	7.1	21.4	14.3
フィッシング詐欺	0	0	45.5	27.3	72.7	0
振り込め詐欺	7.7	0	23.1	38.5	38.5	0

(1-b) 英語

Auction fraud	0	0	30.0	40.0	70.0	0
Phishing	0	0	50.0	60.0	90.0	0

(2) 「手法 2」

トピック	(1) 被害者自身もしくはその知人・目撃者のブログ		(2) Web 上のページを引用		(3)	(4) その他
	被害者自身のブログ	被害者の知人・目撃者のブログ	ニュースを引用	公式サイト等を引用	犯罪被害の予防となる対策を紹介	

(2-a) 日本語

オークション詐欺	21.7	2.9	13.0	18.8	18.8	15.9
フィッシング詐欺	9.0	0	31.3	20.9	35.8	9.0
振り込め詐欺	6.0	1.5	46.3	29.9	26.9	14.9

(2-b) 英語

Auction fraud	1.5	4.5	13.4	56.7	61.2	0
Phishing	13.4	4.5	7.5	41.8	52.2	1.5

て、複数情報源からのニュースの多言語間差異分析を行っている研究 [10]～[13] が挙げられる。[10] は、32 言語における 1000 以上の情報源を分析し伝染病に関するレポートをまとめあげる研究を行っている。[11] では、32 言語におけるニュース記事群から特定の人物名を収集し、その人物の人間関係やその人物について言及している各国のニュース記事を継続的に分析する研究を行っている。[12] は、複数の国の代表的なメディアが発信するニュースを情報源として、同一事象に対する各国のニュースの伝え方の差異分析をテーマとしている。[13] では、9 言語間における同一事象に対する主観情報の差異分析の研究を行っている。これらの関連研究は主にニュース記事を対象に分析を行っている点で本論文とは異なる。

6. おわりに

本稿では事例研究として、犯罪分野に関する日英各言語のブログ記事を収集し、それらをブロガーの立場で類型化する枠組みを提案した。今回は、犯罪分野である「詐欺」カテゴリおよび「インターネット犯罪」カテゴリのトピックを対象にブログの類型化を試みた。また、ブログサイト・ブログ記事の検索に

において、我々は犯罪分野の専門家といったさまざまなブロガーの種類や犯罪行為の被害者といった観点に注目して、「関連語を用いたブログサイト順位付け」と「検索エンジンによるブログ記事順位付け」の二種類の手法を比較した。結果として、いずれの手法もトピックと関連したブログサイト・ブログ記事を多く収集することができた。特に重要な点は、両手法の検索結果で重複したブログサイト・ブログ記事が比較的少なかったことである。今後は、多言語で主観情報抽出手法 [14], [15] を組み込み、トピックに関連するブログの類型化を自動化したいと考えている。

文 献

- [1] 川場真理子, 中崎寛之, 横本大輔, 宇津呂武仁, 福原知宏. Wikipedia 概念体系とブログ空間の間のトピック対応の推定. 日本データベース学会論文誌, Vol. 8, No. 1, pp. 17–22, 2009.
- [2] 中崎寛之, 阿部佑亮, 宇津呂武仁, 河田容英, 福原知宏, 神門典子, 吉岡真治, 中川裕志, 清田陽司. 特定トピックの日英ブログ収集・分析・類型化: 事例研究. 情報処理学会研究報告, Vol. 2009, No. (2009-NL-194), 2009.
- [3] 中崎寛之, 川場真理子, 山崎小有里, 宇津呂武仁, 福原知宏. 同一トピックの日英ブログにおける文化間差異の発見支援. データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2009.

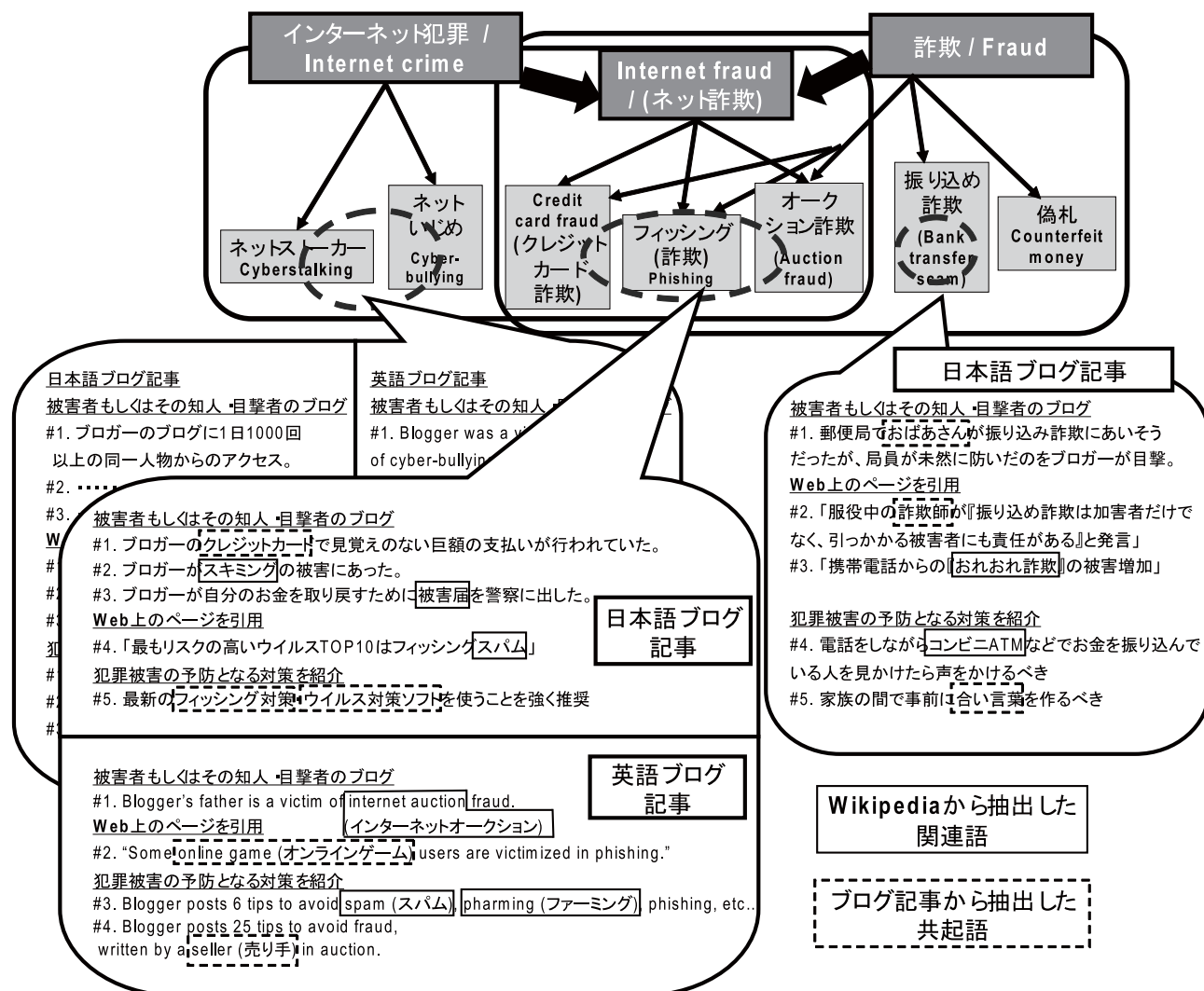


図 5 ブログ記事の類型化および要約の事例

- [4] 中崎寛之. 特定トピックの日英ブログ収集および分析に関する研究. 修士論文, 筑波大学大学院 システム情報工学研究科, 2010.
- [5] 横本大輔, 中崎寛之, 宇津呂武仁, 福原知宏. Wikipedia エントリに関連するブログサイトの収集. 言語処理学会第 16 回年次大会論文集, pp. 126–129, 2010.
- [6] 阿部佑亮, 中崎寛之, 横本大輔, 宇津呂武仁, 河田容英, 福原知宏. 「犯罪」分野に関連するブログの類型化と自動収集. 言語処理学会第 16 回年次大会論文集, pp. 130–133, 2010.
- [7] S. De Saeger, K. Torisawa, and J. Kazama. Looking for Trouble. In *Proc. 22nd COLING*, pp. 185–192, 2008.
- [8] K. Torisawa, S. De Saeger, Y. Kakizawa, J. Kazama, M. Murata, D. Noguchi, and A. Sumida. TORISHIKI-KAI, an Autogenerated Web Search Directory. In *Proc. 2nd ISUC*, pp. 179–186, 2008.
- [9] 乾健太郎, 原一夫. 経験マイニング: web テキストからの個人の経験の抽出と分類. 言語処理学会第 14 回年次大会論文集, pp. 1077–1080. 言語処理学会, 2008.
- [10] R. Yangarber, C. Best, P. von Etter, F. Fuart, D. Horby, and R. Steinberger. Combining Information about Epidemic Threats from Multiple Sources. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 41–48, 2007.
- [11] B. Pouliquen, R. Steinberger, and J. Belyaeva. Multilingual Multi-document Continuously-updated Social Networks. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 25–32, 2007.
- [12] M. Yoshioka. IR Interface for Contrasting Multiple News Sites. In *Prof. 4th AIRS*, pp. 516–521, 2008.
- [13] M. Bautin, L. Vijayarenu, and S. Skiena. International Sentiment Analysis for News and Blogs. In *Proc. ICWSM*, pp. 19–26, 2008.
- [14] D. K. Evans, L.-W. Ku, Y. Seki, H.-H. Chen, and N. Kando. Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task. In *Proc. 3rd Inter. Cross-Language Information Processing Workshop (CLIP2007)*, pp. 456–463, 2007.
- [15] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210, 2005.