Vol.2009-NL-194 No.9
2009/11/17

特定トピックの日英ブログ収集・分析・類型化: 事例研究

中崎寛之[†1]　阿部佑亮[†2]　宇津呂武仁[†1]
河田容英[†3]　福原知宏[†4]　神門典子[†5]
吉岡真治[†6]　中川裕志[†7]　清田陽司[†7]

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て用いた．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．

Collecting/Analyzing/Categorizing Japanese and English Blogs with Specific Topics: A Case Study

Hiroyuki NAKASAKI ,[†1] Yusuke ABE ,[†2]
Takehito UTSURO ,[†1] Yasuhide KAWADA ,[†3]
Tomohiro Fukuhara ,[†4] Noriko KANDO ,[†5]
Masaharu Yoshioka ,[†6] Hiroshi NAKAGAWA [†7]
and Yoji KIYOTA [†7]

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.

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．具体的なサービスの例として，*Technorati*[⋆1]，*BlogPulse*[⋆2]，*kizasi.jp*[⋆3]，*blogWatcher*[⋆4]などが挙げられる．これらの検索サービスは，巨大なブログ空間の索引付けという観点から見ると，キーワードや評判，時系列変化や人手によって作成されたカテゴリ情報などを索引として用いて，利用者の求めるブログ記事やブログサイトを

†1 筑波大学大学院 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
†2 筑波大学 第三学群工学システム学類
College of Engineering Systems, Third Cluster of Colleges, University of Tsukuba
†3 (株) ナビックス
Navix Co., Ltd.
†4 東京大学 人工物工学研究センター
Research into Artifacts, Center for Engineering, University of Tokyo
†5 国立情報学研究所
National Institute of Informatics
†6 北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
†7 東京大学 情報基盤センター
Information Technology Center, University of Tokyo
⋆1 http://technorati.com/
⋆2 http://www.blogpulse.com/
⋆3 http://kizasi.jp/(日本語のみ)
⋆4 http://blogwatcher.pi.titech.ac.jp/(日本語のみ)

© 2009 Information Processing Society of Japan

検索する．また，多言語ブログサービスとしては，*Globe of Blogs* [⋆1]が言語横断ブログ記事検索機能を提供している．ほかにも，アジア言語ブログの検索機能を提供している *Best Blogs in Asia Directory* [⋆2]や，多言語ブログ記事の分析を行っている *Blogwise*[⋆3]がある．

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る研究を行った[4)]．この研究では，ブログ空間におけるトピックの分布を調べるために，トピックである各 Wikipedia エントリに対して，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対応づける．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．また，この索引付けを日英 Wikipedia で行い，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，日英各言語について検索し，その記述内容を二言語間で対照分析する研究を行った[5),6)]．これにより，ブログ特有の個人レベルの情報や意見における国間差異を多数観測することができた．しかし，トピックである日英 Wikipedia エントリに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．例として，トピック「オークション詐欺」を挙げると，あるブロガーは自分の父親がインターネットオークション詐欺にあった時の被害経験について記述している．しかし，別のブロガーはオークションの売り手の立場からオークション詐欺にあわないための対策法を詳しく記述しているなど，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．したがって，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においても，利用対象となる情報や知識の種類・内容の多様性に応じて，収集されたブログサイトやブログ記事を適切に分類して提示することが重要である．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，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，犯罪分野のトピックを対象としたブログの類型化を題材とした．さらに，犯罪分野のカテゴリから，「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のみを類型化の対象にした．これは，これらの分野の犯罪が悪質でありながら多くの人の生活に身近なものであるため，被害者によるブログが多く存在すると考えたためである．立場の例として，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詳しく掲載しているブログなどが挙げられる．犯罪行為の被害者によるブログは，被害者自身の犯罪の被害経験などの貴重かつ独自の記述が書かれている．犯罪行為の報道記事を引用しているブログは，その報道を自身のブログ記事で取り上げることで周囲に犯罪行為に対する注意を喚起している．犯罪行為に対する対策の仕方について詳しく掲載しているブログは，周囲の人が犯罪行為に巻き込まれないよう犯罪行為に対する対策を詳しく記載している．これらのブログ記事の情報は，特定の犯罪行為から身を守るための方法を探している人や，既に被害に遭ってしまったが，その状況を打開する方法を探している人にとって有益である．特にその国の文化や犯罪に関して不慣れな外国人にとっては，非常に価値のある情報であると考えられる．このような観点から，本研究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する．

2. 本研究の全体的枠組み

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 [⋆4] 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．本研究では今回 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．これにつ

⋆1 http://www.globeofblogs.com/
⋆2 http://www.misohoni.com/bba/
⋆3 http://www.blogwise.com/

⋆4 http://{en,ja}.wikipedia.org/.

© 2009 Information Processing Society of Japan

Vol.2009-NL-194 No.9
2009/11/17

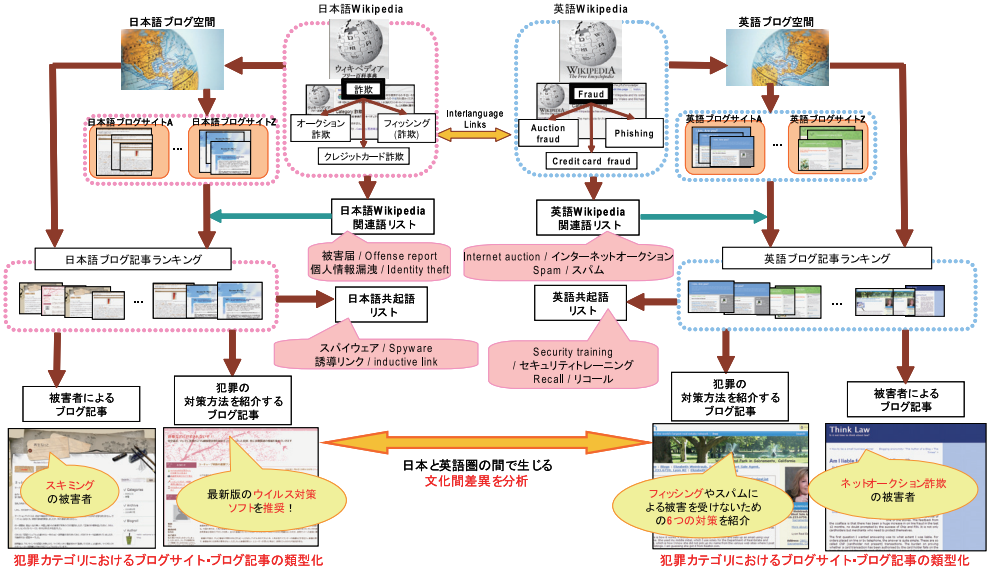


図 1 二言語対照ブログ分析およびブログサイト・ブログ記事の類型化の全体的枠組み

いては 4 節で詳しく説明する．

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[5] 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．

最後に，収集したブログサイトおよびブログ記事集合を以下のタイプに分類した．

(1) 被害者もしくはその知人・目撃者のブログ (未遂を含む)
(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：ブロガーの意見のみ記述されているブログ）

これらの詳細は，5 節で説明する．

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 ⋆1 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「振り込め詐欺」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 ⋆2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．

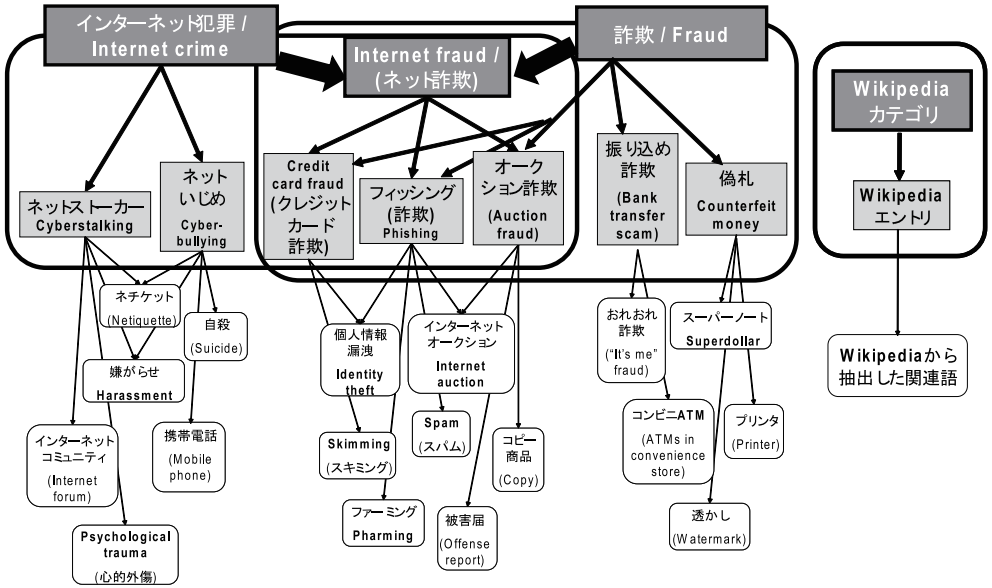
「詐欺」カテゴリおよび「インターネット犯罪」カテゴリにおけるトピックの例を図 2 に示す⋆3．ここで,「詐欺」カテゴリおよび「インターネット犯罪」カテゴリの下位に位置す

⋆1 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 11 社 (FC2.com,yahoo.co.jp,rakuten.ne.jp,ameblo.jp,goo.ne.jp,livedoor.jp,Seesaa.net,jugem.jp,yaplog.jp,webry.info.jp,hatena.ne.jp)，英語ブログでは大手 12 社 (blogspot.com,msnblogs.net,spaces.live.com,livejournal.com,vox.com,multiply.com,typepad.com,aol.com,blogsome.com,wordpress.com,blog-king.net,blogster.com) のブログ会社のドメインに限って検索を行った．

⋆2 http://www.eijiro.jp/

⋆3 参考までに，これらの各トピック名に該当する犯罪について，日本及び米国における検挙件数 (出典：警察庁 (http://www.npa.go.jp/)，警察庁サイバー対策 (http://www.npa.go.jp/cyber/)，米国 Internet Crime Complaint Center(IC3) (http://www.ic3.gov/)) およびブログ検索ヒット数を表 1 に示す．

© 2009 Information Processing Society of Japan



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

図 2 「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

るカテゴリとして「ネット詐欺」カテゴリがある．この「ネット詐欺」カテゴリ下に属するトピックの例として，三つのトピックを図中に示す．さらに，図 2 には，各トピックの Wikipedia エントリから抽出した関連語の例も示す．これらの関連語は 4.1 節で説明する「関連語を用いたブログサイト順位付け」で用いる．また，同じカテゴリ下に属する複数のトピックがいくつかの関連語を共有する場合も多い．

4. ブログサイト・ブログ記事検索手法

2 節で述べたように，本研究ではブログの検索に二種類の検索手法を用いた．本節では，それぞれの手法について説明する．

4.1 手法 1：関連語を用いたブログサイト順位付け

一つ目の検索手法は,「関連語を用いたブログサイト順位付け」(以下,「手法 1」と呼ぶ) で

表 1 「詐欺」カテゴリおよび「インターネット犯罪」カテゴリにおける各犯罪の検挙件数およびブログ空間中のヒット数

トピック名	検挙件数 (2008 年)		ブログ空間中のヒット数 (2009 年 9 月)	
	日本	アメリカ合衆国	日本語ブログ	英語ブログ
インターネット詐欺	N/A	72,940	61,600	21,300
(オークション詐欺)	1,140	18,600	44,700	1,760
(クレジットカード詐欺)	N/A	6,600	8,590	43,900
(フィッシング詐欺)	N/A		136,000	479,000
振り込め詐欺	4,400	N/A	349,000	30
偽札	395	N/A	40,500	16,800
ネットストーカー	N/A		32,100	20,300
ネットいじめ	N/A		45,700	38,900

ある．現在のブログ検索サービスでは，被リンク数の多い人気ブログサイトの記事から優先的に検索されるため，被リンク数は多くないが，特定トピックについて詳細な記述のあるブログサイトが検索されにくい．そこで，以下の手順に従って，ブログサイト・ブログ記事の検索を行った．

まず，Wikipedia エントリをトピックとしてブログサイトの検索を行う．ブログサイトを検索するために，本研究では日本語ブログの検索には Yahoo! Japan 検索 API を，英語ブログの検索には米 Yahoo!検索 API を利用し，2 節の注釈で述べたブログ会社のドメイン (日本語ブログでは大手 11 社，英語ブログでは大手 12 社) に限って検索を行った．検索の際には，Wikipedia エントリのエントリ名を検索クエリとして，複数のブログホストを一度に指定して検索し，1000 件の記事を取得する．しかし API の検索ではブログ記事単位の検索になるので，同一著者のブログ記事は一つのブログサイトにまとめるという作業を行った．その後，各ブログサイトにおいて，Wikipedia エントリのエントリ名のヒット数を求め，ヒット数が下限未満 (本論文では，10) のブログサイトを削除した．

次に，収集した日英ブログサイト集合の中から，トピックについて詳しく書かれたブログ記事を選定する．手法としては，トピック名がタイトルである各言語の Wikipedia エントリのリダイレクト，さらに Wikipedia エントリの本文から太字，他エントリリンクをブログ記事検索のための関連語として抽出する．そして，収集したブログサイト集合中から，抽出した関連語のいずれかが出現する各言語のブログ記事を選定する．各トピックの 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 2 に示す．

© 2009 Information Processing Society of Japan

表 2 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
振り込め詐欺 / 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

最後に，よりトピックについて詳しく記述しているブログサイト・ブログ記事を得るために，各言語において収集したブログサイトおよびブログ記事をそれぞれ順位付けする．順位付けには，Wikipedia エントリから抽出した関連語を用いる．ブログ記事は，以下のスコアの降順に順位付けする．

$$PostScore(p) = \sum_{t}(weight(type(t)) \times freq(t))$$

$weight(type(t))$ は，Wikipedia から抽出した関連語 t の種類 $type(t)$ に付与する重みで，$freq(t)$ は，ブログ記事 p 内における Wikipedia から抽出した関連語 t の出現頻度である．また，Wikipedia から抽出した関連語 t の種類 $type(t)$ がリダイレクトの場合は重みを 3，太字の場合は重みを 2，他エントリリンクの場合は重みを 0.5 とする．ブログサイトについては，各ブログサイトに含まれるブログ記事のスコアの総和の降順に順位付けする．

本研究では，これまでの調査結果から,「手法 1」は，トピック名である犯罪行為に関するニュースを収集しそれらに独自の意見をしているブロガーや，犯罪行為による被害の予防となる対策を多く紹介しているブロガーの検索に適している手法だということがわかった．これらの類型化については，5 節で詳しく説明する．

4.2 手法 2：検索エンジンによるブログ記事順位付け

二つ目の検索手法 は,「検索エンジンによるブログ記事順位付け」(以下,「手法 2」と呼ぶ) である.「手法 2」では，既存の Web 検索エンジンでブログ記事を検索し，その検索結果のランキングをそのまま用いる.「手法 1」と同様に日本語ブログ記事の検索には Yahoo!Japan 検索 API を，英語ブログ記事検索には米 Yahoo!検索 API を用いた．4.1 節で述べたように，これらの検索エンジンでは，被リンク数の多い人気ブログサイトの記事から優先的に検索される．

「手法 2」では，2 節で述べたブログサイト・ブログ記事のタイプ (1)，タイプ (2)，タイプ (3) でそれぞれ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，あるトピック t について，タイプ (1) の被害者もしくはその知人・目撃者によるブログ記事を検索する際には，日本語では「t AND 被害」，英語では「t AND victim」あるいは「t AND 'I was a victim'」を検索クエリとして検索を行った．その結果，多くのトピックにおいて，被害者もしくはその知人・目撃者によるブログ記事が検索結果の上位に多く出現した．例えば，トピック「オークション詐欺」の場合，各言語のブログ記事ランキング上位 20 件のうち，被害者もしくはその知人・目撃者によるブログ記事が日本語で 8 件，英語で 3 件出現した．これらのブログ記事はいずれも「手法 1」のブログサイトランキングでは収集することができなかったブログサイトの記事であった．これは，犯罪行為の被害者によるブログサイトは，犯罪行為に関するニュースを引用したり犯罪行為の被害を防ぐ方法を多く紹介したりするブログサイトとは違い，定期的に犯罪行為に関するブログ記事を書かないためだと考えられる．このような結果から,「手法 2」は，犯罪行為の被害者によるブログ記事を「手法 1」よりも多く検索できることがわかった．また，同様に他のタイプの検索クエリとして，タイプ (2) の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記事を検索する場合，日本語では「t AND 引用」，英語では「t AND reference」を用い，タイプ (3) の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記事を検索する場合は，日本語では「t AND 対策」，英語では「t AND tips」をそれぞれ用いた．しかし，今回「手法 2」で用いた検索クエリの設計は十分ではなく，改善の余地がある．今後の課題として，新たな検索クエリの設計が挙げられる．

5. 「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.「詐欺」カテゴリに属する「オークション詐欺」,「フィッシング」,「振り込め詐欺」[⋆1] のそれぞれに関連するブログサイト・ブログ記事の割合,「手法 1」と「手法 2」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 3 に

⋆1 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 3 と表 4 には記載していない．

© 2009 Information Processing Society of Japan

表 3 「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 / 手法 1 で収集した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 2 で収集した関連ブログ記事数	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				
(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

示す.

また，2 節で述べたように，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)
(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：ブロガーの意見のみ記述されているブログ）

表 3 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 4 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.

表 4 「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) 犯罪被害の予防となる対策を紹介	(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		
(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) 犯罪被害の予防となる対策を紹介	(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		
(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

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

© 2009 Information Processing Society of Japan

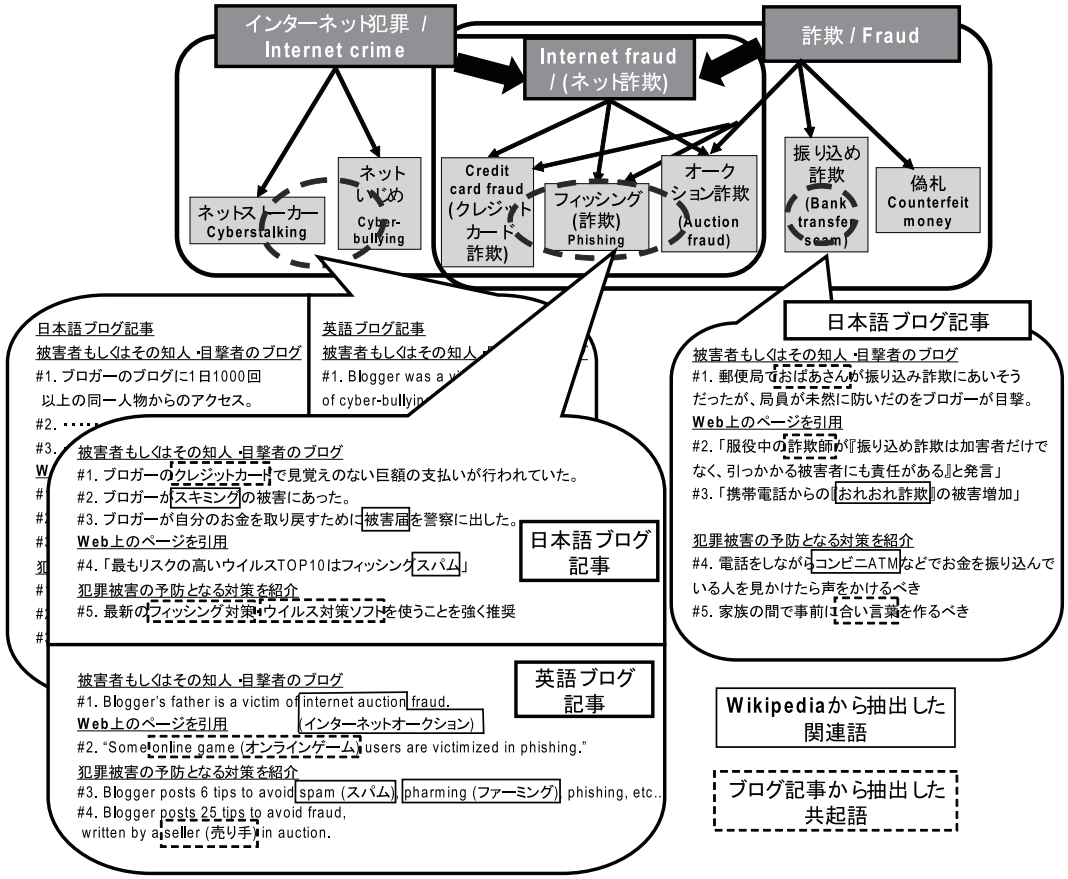


図 3 ブログ記事の類型化および要約の事例

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 3 に示す．図中では，図 2 で示した 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

© 2009 Information Processing Society of Japan

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.

6. 関連研究

関連研究として，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[1),7),10),11)]が挙げられる．文献 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．文献 7) 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．文献 11) 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．文献 1) 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．また，その他の関連研究として，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[2),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の特定に適用可能であると考えられる.

7. 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[3),9)] 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.

参 考 文 献

1) Bautin, M., Vijayarenu, L. and Skiena, S.: International Sentiment Analysis for News and Blogs, *Proc. ICWSM*, pp.19–26 (2008).
2) DeSaeger, S., Torisawa, K. and Kazama, J.: Looking for Trouble, *Proc. 22nd COLING*, pp.185–192 (2008).
3) Evans, D.K., Ku, L.-W., Seki, Y., Chen, H.-H. and Kando, N.: Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task, *Proc. 3rd Inter. Cross-Language Information Processing Workshop (CLIP2007)*, pp.456–463 (2007).
4) 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).
5) Nakasaki, H., Kawaba, M., Utsuro, T. and Fukuhara, T.: Mining Cross-Lingual/Cross-Cultural Differences in Concerns and Opinions in Blogs, *Computer Processing of Oriental Languages: Language Technology for the Knowledge-Based Economy: 22nd International Conference, ICCPOL 2009* (Li, W. and Mollá-Aliod, D., eds.), Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence, Vol.5459, Springer, pp.213–224 (2009).
6) 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).
7) Pouliquen, B., Steinberger, R. and Belyaeva, J.: Multilingual Multi-document Continuously-updated Social Networks, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp.25–32 (2007).
8) Torisawa, K., De Saeger, S., Kakizawa, Y., Kazama, J., Murata, M., Noguchi, D. and Sumida, A.: TORISHIKI-KAI, an Autogenerated Web Search Directory, *Proc. 2nd ISUC*, pp.179–186 (2008).
9) Wiebe, J., Wilson, T. and Cardie, C.: Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210 (2005).
10) Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D. and Steinberger, R.: Combining Information about Epidemic Threats from Multiple Sources, *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp.41–48 (2007).
11) Yoshioka, M.: IR Interface for Contrasting Multiple News Sites, *Prof. 4th AIRS*, pp.516–521 (2008).

© 2009 Information Processing Society of Japan