3B3-1

特定トピックのブログの収集および類型化 — 「犯罪」分野を事例として —

A Case Study of Blog Distillation and Blog Categorization with "Crime" Domain

阿部 佑亮[*1] Yusuke Abe　中崎 寛之[*2] Hiroyuki Nakasaki　横本 大輔[*1] Daisuke Yokomoto　宇津呂 武仁[*1] Takehito Utsuro　河田 容英[*3] Yasuhide Kawada
福原 知宏[*4] Tomohiro Fukuhara

[*1]筑波大学大学院システム情報工学研究科 Grad. Sch. Systems and Information Engineering, University of Tsukuba
[*2]株式会社 NTT データ NTT DATA CORPORATION
[*3](株) ナビックス Navix Co., Ltd.
[*4]独立行政法人 産業技術総合研究所 サービス工学研究センター Center for Service Research, National Institute of Advanced Industrial Science and Technology

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 people who are not familiar with crimes. This paper proposes a framework of categorizing people's concerns, reports, and experiences on crimes in their own blogs. First, we refer to *Wikipedia* as a terminological knowledge base, and search for Wikipedia entries describing criminal acts. We categorize blog feeds/posts into four types including experts in the crime domain and victims of criminal acts. We then propose how to detect blog posts by 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．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 [川場 09]．そして，その対応づけによって収集可能となった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」で類型化することが必要だと考えた．

我々はまず，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試みた [中崎 09]．その結果，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．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や，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が存在した．また，特に「被害者によるブログ記事」には，被害者自身の犯罪の被害経験などといった，貴重かつ独自の記述が書かれていることも分かった．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．そして，[中崎 09] の成果を踏まえて，本稿では，特に「被害者によるブログ」に注目し，それらを自動収集する手法を提案する [阿部 10].

連絡先: 阿部 佑亮，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 1-1-1，029-853-5427

2. 「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，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 以上のトピックのみを選定の対象とした．その結果,「詐欺」カテゴリでは 20 トピック,「インターネット犯罪」では 8 トピックとなった． さらに，それらの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 10 トピック,「インターネット犯罪」カテゴリからは 5 トピックが選定された．

「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 1 に示す.「ネット詐欺」カテゴリは,

[*1] ブログの検索には Yahoo!Japan 検索 API(http://www.yahoo.co.jp)，を用い，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) のブログ会社のドメインに限って検索を行った．

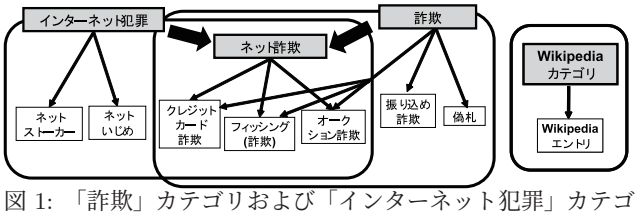


図 1: 「詐欺」カテゴリおよび「インターネット犯罪」カテゴリにおけるトピックの例

「詐欺」カテゴリおよび「インターネット犯罪」カテゴリの下位カテゴリに属し，そのカテゴリ下に属する 3 つのトピックを図中に例として挙げた.

3. 「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，まず,「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.

その結果，犯罪分野におけるブログを以下のタイプに分類することができた.

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）

さらに，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」は三つに,「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」については二つに，それぞれ分類した.「被害者もしくはその知人・目撃者によるブログ」は,「被害者によるブログ」と,「被害未遂の人のブログ」,「被害者の知人または目撃者によるブログ」の三種類に分類し,「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.

4. 「被害者によるブログ記事」の自動収集手法

次に，前節で述べた類型のうち，特に「被害者によるブログ記事」について，それらを自動収集する手法について述べる. 本研究で提案する手法は，検索エンジン API を用いて「被害者によるブログ記事」の候補を収集し,「被害者によるブログ記事」同定規則を用いて順位付けをする，というものである.

4.1 検索エンジン API によるブログ記事収集

本稿では，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，3. で紹介した記事のタイプ (1)，(2)，(3) 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，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「t AND 被害」を用い，同様にタイプ (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の際は「t AND 引用」を，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の際は,「t AND 対策」を，それぞれ用いた.

表 1: 「被害者によるブログ記事」同定規則において用いる手がかり表現およびそのスコア

手がかり表現のタイプ		手がかり表現の例	スコア	種類数
係り受け関係	基本形	被害 - 遭う, 詐欺 - 引っ掛かる	10	19
	派生形	被害 - あいました, 詐欺 - 引っ掛かった		84
文節単位の表記パターン	高スコア	やられた, 騙された, 詐欺られた	2	13
	中スコア	音信不通，凹む, 被害届，不審	1	113
	低スコア	警察，連絡，あって	0.5	17

以上の全 4 種類のクエリを用いて，それぞれトピックに関する記事を 1000 記事収集し，アーカイブを除去する．そして，クエリ間での重複記事を除いたものを，上述の同定規則による順位付けの対象記事集合とした.「オークション詐欺」「フィッシング詐欺」「クレジットカード詐欺」の各トピックで，順位付けの対象となった記事数はそれぞれ，1063 記事，1388 記事，968 記事であった.

4.2 手がかり表現に注目した「被害者によるブログ記事」の順位付け

本稿では，検索エンジンによって収集されたブログ記事集合の中から「被害者によるブログ記事」を収集するために，記事文中の手がかり表現に注目した．具体的には,「被害者によるブログ記事」の同定の手がかりとなる表現を用いて同定規則を作成しておく．そして，検索エンジンによって収集されたブログ記事に対して，同定規則を用いたスコアリングを行い，スコアの合計によって順位付けを行う.

手がかり表現は，大別して,「係り受け関係」と「文節単位の表記パターン」の 2 種類がある．この内，特に「係り受け関係」の方が「被害者によるブログ記事」同定で重要となる．例えば,「オークションで見事に騙された.」という文があった場合,「オークション - 騙された」という係り受け関係から，このブログ記事を書いたブロガーはオークション詐欺の被害者である可能性が高いと考えられる．そのため,「係り受け関係」には「文節単位の表記パターン」と比べて高スコアを付与した．一方,「やられた」「騙された」「不審」などの「文節単位の表記パターン」に対しては，表現によって 3 種類のスコアのいずれかが付与されている．手がかり表現の例とそのスコアを表 1 に，係り受け関係の例とその例文を表 2 に，それぞれ示す．なお，規則作成の際,「オークション詐欺」の被害者自身が記述した 20 件の記述および,「フィッシング詐欺」の被害者自身が記述した 3 件の記述を，それぞれ参照した.

4.3 評価

前節で述べた「同定規則を用いた順位付け」(提案手法) とベースラインとを比較し,「被害者によるブログ」の収集性能の評価を行った.

ベースラインとしては，既存の Web 検索エンジン (Yahoo!Japan 検索 API を使用) で「t AND 被害」をクエリとした時の順位付けを用いた．これらの検索エンジンでは，被リ

表 2: 「被害者によるブログ記事」同定規則において用いる係り受け関係の例および例文

基本形	派生形	例文
被害 - 遭う	被害 - 遭う	まさか自分がそんな被害に遭うなんて !!
	被害 - 遭いました	いわゆるネットオークションで詐欺被害に遭いました.
	被害 - あいました	私 詐欺の被害にあいました.
詐欺 - 引っ掛かる	詐欺 - 引っ掛かった	運悪くオークション詐欺に引っ掛かったみたいなんです.
	詐欺 - 引っ掛かっていた	ブログの更新が滞っていたと思ったら，実は詐欺に引っ掛かっていたのでした.
詐欺 - 遭遇	詐欺 - 遭遇	ヤフーオークションにて詐欺に遭遇してしまいました.

表 3: 「被害者によるブログ記事」収集手法の詳細分析

(a) 提案手法

トピック名	評価数	係り受けが適用された記事				係り受け関係が適用されなかった記事			
				不正解数		正解数		不正解数	
		総数	正解数	「被害者によるブログ記事」以外の類型	トピックと無関係	トピックに特化した係り受け関係により同定可能	主観情報を利用することで同定可能	「被害者によるブログ記事」以外の類型	トピックと無関係
オークション詐欺	50	50	22	17	11	0	0	0	0
フィッシング詐欺	50	43	3	34	6	0	1	4	2
クレジットカード詐欺	50	14	6	2	6	0	2	13	21
結婚詐欺	50	39	0	20	19	0	0	4	7
偽札	50	4	0	1	3	1	2	15	28
誹謗中傷	50	1	0	1	0	2	5	18	24
おとり商法	48	0	0	0	0	0	1	26	21

(b) ベースライン

トピック名	評価数	正解数			不正解数	
		「提案手法」と共通	係り受け関係の拡張により同定可能	その他	「被害者によるブログ記事」以外の類型	トピックと無関係
オークション詐欺	50	2	1	1	15	31
フィッシング詐欺	50	0	0	0	39	11
クレジットカード詐欺	50	6	3	0	5	36
結婚詐欺	50	0	0	0	22	28
偽札	50	0	2	0	30	18
誹謗中傷	50	0	5	3	27	15
おとり商法	8	0	0	0	6	2

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．また,「t」のみでの検索よりも「t AND 被害」での検索のほうが「被害者によるブログ記事」をより上位に収集できることが分かっている [中崎 09].

トピック「オークション詐欺」「フィッシング詐欺」「クレジットカード詐欺」についての評価結果を図 2 に示す.「オークション詐欺」と「フィッシング詐欺」では，提案手法の方が検索エンジンよりも，上位により多く「被害者によるブログ記事」を収集できたが,「クレジットカード詐欺」では，逆に検索エンジンの方が提案手法よりも上位に多く「被害者によるブログ記事」を収集できた.

各トピックについて収集された上位記事を分析したところ，提案手法によって収集された「被害者によるブログ記事」の多くは，同定規則中の係り受け関係によって上位に収集されていた．そして，各トピックについて上位に収集された記事のうち，同定規則中の係り受け関係を含む記事は,「オークション詐欺」では上位 75 記事,「フィッシング詐欺」では上位 43 記事,「クレジットカード詐欺」では上位 14 記事であった.

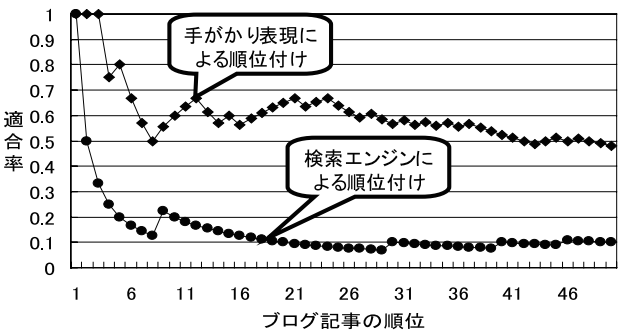
本稿で評価対象とした 7 トピックについて，提案手法およびベースラインの性能を詳細に分析した結果を表 3 に示す.

まず，提案手法で収集したブログ記事に「トピックと無関係なブログ記事」が，トピック「クレジットカード詐欺」「偽札」を中心に多く見られた．これについては，Wikipedia 関連語を用いた「トピックと関係のあるブログ記事」の同定手法 [川場 09] の導入によって対応できると考えられる．また，係り受け関係が適用された記事の中に，3. で紹介した「被害者によるブログ記事」以外の類型も見られた．それらに関しては，各類型について類型を同定する手法を考案し,「被害者によるブログ記事」以外の類型であると判定することで対応できると考えられる.

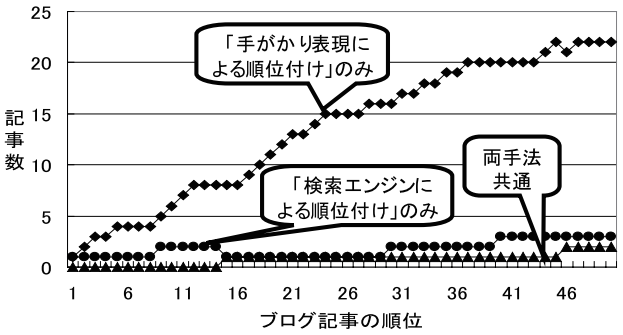
係り受け関係が適用されなかったが,「被害者によるブログ記事」であったものに関しては，2 種類の対応策が考えられる．1 つは,「トピックに特化した係り受け関係による同定」である．例えば，トピック「偽札」であれば,「偽札 - つかまされました」といった係り受け関係を導入することが考えられる．もう 1 つは,「主観情報の利用」である．例えば，トピック「クレジットカード詐欺」においては,「が〜〜〜ん」のような表現や絵文字などを用いて，被害に遭ったことに対する主観を表現しているブログ記事があり，これらの「被害者によるブログ記事」は，提案手法では同定できなかった．このように，ブログにおいて被害者が被害に遭ったことを表現する際には，係り受け関係以外の主観表現が用いられることも多いと考えられる.

5. 関連研究

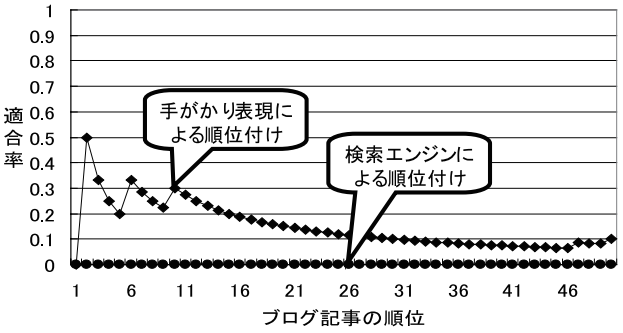
関連研究として，Web 上のページからトラブルを表す文の抽出を行っている研究 [De Saeger 08, Torisawa 08] が挙げられる．この研究でのトラブル表現抽出技術は，我々の研究における「被害者によるブログ記事」同定においても有用な可能性がある．また，Web 上の膨大なブログから人々の経験情報を収集し，意味的に索引付けて DB 化する手法 [乾 08] について



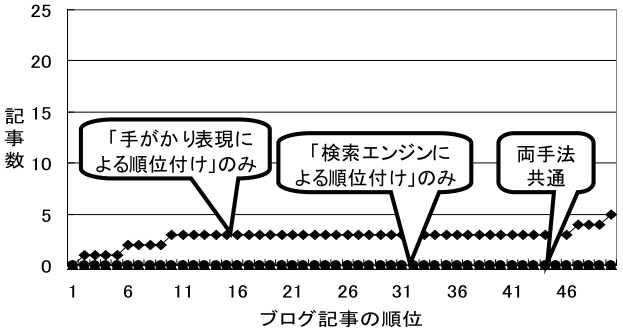
(1-a) 「オークション詐欺」(適合率)



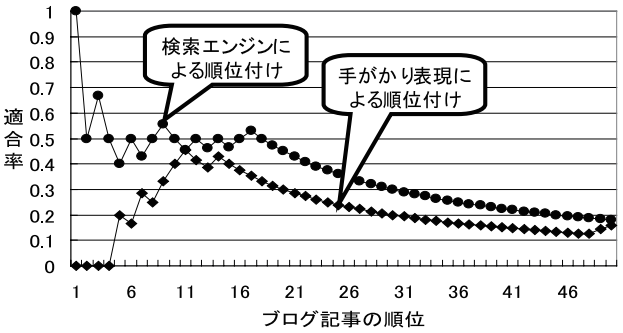
(1-b) 「オークション詐欺」(収集記事数)



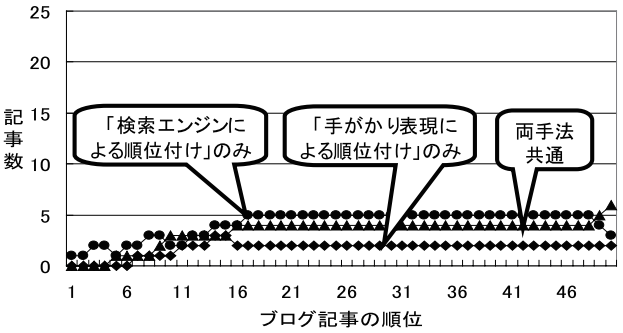
(2-a) 「フィッシング詐欺」(適合率)



(2-b) 「フィッシング詐欺」(収集記事数)



(3-a) 「クレジットカード詐欺」(適合率)



(3-b) 「クレジットカード詐欺」(収集記事数)

図 2: 「被害者によるブログ記事」収集性能の評価

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．

6. まとめと今後の課題

本稿では，まず，事例研究として，犯罪分野に関する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．特に,「被害者によるブログ」について，それらを自動収集する手法を提案した．提案手法を用いて「被害者によるブログ」を自動収集した結果，検索エンジンでの順位付けよりも多くの「被害者によるブログ記事」を上位に収集することができた．

今後の課題としては，[川場 09] の手法を用いて，トピックと関係のあるブログ記事のみを対象として順位付けを行うことや,「被害者によるブログ記事」以外の類型の同定手法の考案，トピックに特化した係り受け関係の拡張，主観情報抽出手法の導入が挙げられる．

参考文献

[阿部 10] 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)

[De Saeger 08] De Saeger, S., Torisawa, K., and Kazama, J.: Looking for Trouble, in *Proc. 22nd COLING*, pp. 185–192 (2008)

[乾 08] 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)

[川場 09] 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)

[中崎 09] 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)

[Torisawa 08] Torisawa, K., De Saeger, S., Kakizawa, Y., Kazama, J., Murata, M., Noguchi, D., and Sumida, A.: TORISHIKI-KAI, an Autogenerated Web Search Directory, in *Proc. 2nd ISUC*, pp. 179–186 (2008)