Cross-Lingual Analysis of Concerns and Reports on Crimes in Blogs

Hiroyuki Nakasaki[1], Yusuke Abe[1], Takehito Utsuro[1], Yasuhide Kawada[2], Tomohiro Fukuhara[3], Noriko Kando[4], Masaharu Yoshioka[4], Hiroshi Nakagawa[3], and Yoji Kiyota[3]

[1] University of Tsukuba, Tsukuba, 305-8573, Japan
[2] Navix Co., Ltd., Tokyo, 141-0031, Japan
[3] University of Tokyo, Kashiwa 277-8568 / Tokyo, 113-0033, Japan
[4] National Institute of Informatics, Tokyo, 101-8430, Japan
[5] Hokkaido University, Sapporo, 060-0814, Japan
{nakasaki,utsuro}@nlp.iit.tsukuba.ac.jp

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 framework of cross-lingually analyzing people's concerns, reports, and experiences on crimes in their own blogs. 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.

Keywords: cross-lingual blog analysis, blog feed retrieval, crime reports, Wikipedia

1 Introduction

Weblogs or blogs are considered to be one of personal journals, market or product commentaries. 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. One type of such bloggers are those who have expert knowledge in the crime domain, and keep referring to news posts on criminal acts in their own blogs. Another type of bloggers who have expert knowledge also often post tips for how to prevent certain criminal acts. Furthermore, it is surprising that victims of certain criminal acts post blog articles on their own experiences. Blog posts by such various kinds of bloggers are actually very informative for both who are seeking for information on how to prevent certain criminal acts and who have been already victimized and are seeking for information on how to solve their own cases. Such information is especially valuable for outsiders from abroad who are not familiar with cultures and crimes in foreign countries. Based on this observation, this paper proposes

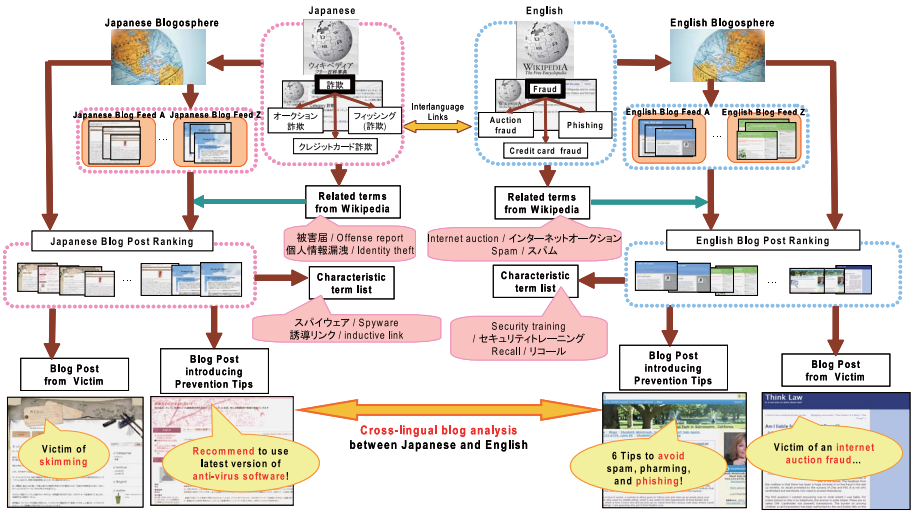


Fig. 1. Overall Framework of Cross-Lingual Blog Analysis

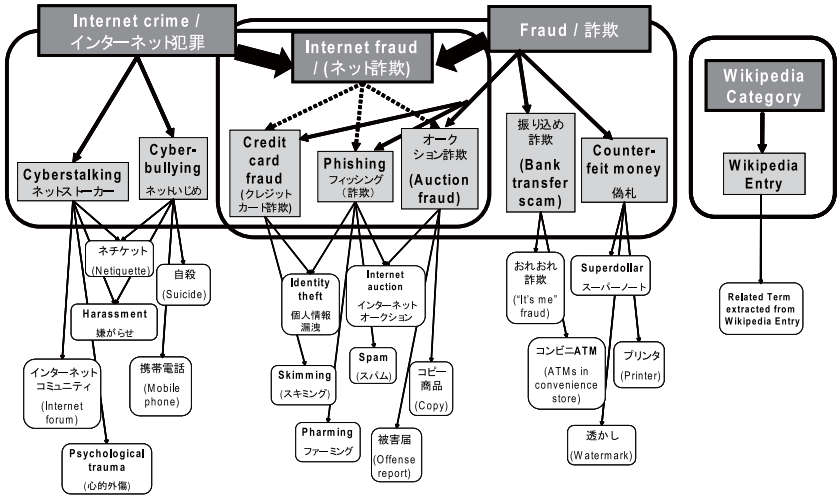
a framework of cross-lingually analyzing people's concerns, reports, and experiences on crimes in their own blogs.

The overview of the proposed framework is shown in Figure 1. First, we start from terms which represent relatively narrow range of concepts of criminal acts. In this paper, we focus on *"fraud"* and *"Internet crime"*. Then, we refer to *Wikipedia* (English and Japanese versions[1]) as a multilingual terminological knowledge base, and search for Wikipedia entries describing criminal acts in the range of *"fraud"* or *"Internet crime"*. Then, from the collected Wikipedia entries, multilingual queries for retrieving blog feeds/posts (in this case *"auction fraud"*, *"credit card fraud"*, and *"phishing"*) are created, where interlanguage links are used for linking English and Japanese translated entries.

Next, in the retrieval of blog feeds/posts, we take two approaches. The first approach [1] focuses on collecting blog feeds rather than on directly collecting blog posts. Its major component is designed as the blog feed retrieval procedure[2] recently studied in the blog distillation task of TREC 2007 blog track [3]. In this first approach, we regard blog feeds as a larger information unit in the blogosphere. We intend to retrieve blog feeds which roughly follow the criterion studied in the blog distillation task, which can be summarized as *Find me a blog with a principle, recurring interest in X*. In the results of empirical analysis of this paper, we found that this first approach is mostly quite appropriate for

[1] http://en.ja.wikipedia.org/. The underlying motivation of employing Wikipedia is in linking a knowledge base of well known facts and relatively neutral opinions with rather raw, user generated media like blogs.

[2] Its detailed evaluation including improvement over the baseline as the original rankings returned by "Yahoo!" API and "Yahoo! Japan" API is published in [2].



(Note: "A term t_x" ("a term t_y") in the nodes above indicates that t_y is not an entry in Wikipedia, nor extracted from any of Wikipedia entries, but translated from t_x by Eijiro.)

Fig. 2. Wikipedia Entries and Related Terms in the "Fraud / Internet Crime" Domain

discovering bloggers (i.e., blog feeds) who are deeply interested in watching news reports on cases of those criminal acts and warning others by referring to those news reports in their own blog posts. This first approach is also quite effective in discovering bloggers who introduce tips for how to prevent such criminal acts.

In the second approach, we simply use existing Web search engine APIs and follow the original ranking of blog posts given a query keyword. Generally speaking, we can not estimate the ranking strategy employed by Web search engine APIs. In principle, however, it is expected that this second approach generally returns blog posts which have more inlinks than other posts. Actually, throughout the empirical analysis of this paper, we found that this second approach returns blog posts by victims of those criminal acts more often than the first approach. Here, we observe that victims of criminal acts such as fraud and Internet crime sometimes post one or two articles to their own blogs just after they were victimized. Though, in most cases, those victims do not keep posting articles related to those crimes, and hence, their blog feeds are not ranked high in the result of blog feeds ranking by the first approach.

Finally, to retrieved English and Japanese blog feeds/posts, we apply our framework of mining cross-lingual/cross-cultural differences in characteristic terms within top ranked blog posts [1] (to be presented in section 3). This framework is originally designed for discovering cross-lingual/cross-cultural differences in concerns and opinions that are closely related to a given topic. With this framework, it becomes much easier for users to discover regional differences in criminal acts which originate from cultural differences. For example, recently, *"bank transfer*

scam", and especially *""it's me" fraud"*[3] is very frequent in Japan, while they are not very frequent in western countries.

2 Topics in the "Fraud / Internet Crime" Domain

Table 1. Statistics of "Fraud / Internet Crime"

ID	Wikipedia Entry	# of Cases (sent to the court in the Year of 2008)		# of Hits in the Blogosphere (checked at Sept. 2009)	
		U.S.A.	Japan	English	Japanese
1	Internet fraud	72,940	N/A	21,300	61,600
2	(Auction fraud)	18,600	1,140	1,760	44,700
3	(Credit card fraud)	6,600	N/A	43,900	8,590
4	(Phishing)	N/A		479,000	136,000
5	Bank transfer scam	N/A	4,400	30	349,000
6	Counterfeit money	N/A	395	16,800	40,500
7	Cyberstalking	N/A		20,300	32,100
8	Cyber-bullying	N/A		38,900	45,700

In this paper, as topics in the domain of criminal acts, we focus on *"fraud"* and *"Internet crime"*. We first refer to Wikipedia and collect entries listed at the categories named as *"fraud"* and *"Internet crime"* as well as those listed at categories subordinate to *"fraud"* and *"Internet crime"*. Then, we require entry titles to have the number of hits in the blogosphere over 10,000 (at least for one language)[4]. Finally, for the category *"fraud"*, we have about 10 to 15 entries both for English and Japanese. For the category *"Internet crime"*, we have about 5 entries both for English and Japanese. Figure 2 shows samples selected from remaining entries[5]. In the figure, the category *"Internet fraud"* is an immediate descendant of both *"fraud"* and *"Internet crime"*, where three entries listed at this category are selected as samples. The figure also shows samples of related terms automatically extracted from those selected Wikipedia entries. For those selected sample entries, Table 1 shows the number of cases actually sent to the

[3] The fraudster makes a phone call to the victim and pretends to be his/her child or grandchild, and then requests the victim to transfer funds to the fraudster's account.

[4] We use the search engine"Yahoo!" API (http://www.yahoo.com/) for English, and the Japanese search engine "Yahoo! Japan" API (http://www.yahoo.co.jp/) for Japanese. Blog hosts are limited to major ones, i.e., 12 for English and 11 for Japanese.

[5] For some of those selected samples, only English or Japanese term is listed as a Wikipedia entry. In such a case, translation is found in an English-Japanese translation lexicon Eijiro (http://www.eijiro.jp/, Ver.79, with 1.6M translation pairs).

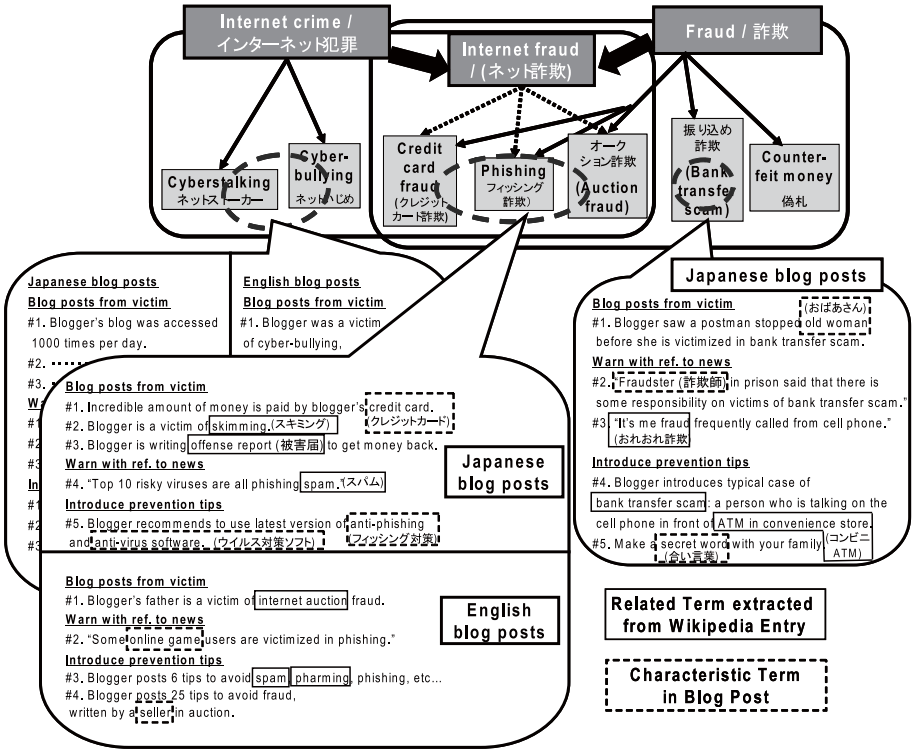


Fig. 3. Topics, Terms in Blogs, and Summaries of Blog Posts: Examples

court both in U.S.A and in Japan[6]. The table also shows the number of hits of those sample entry titles in the blogosphere.

3 Blog Analysis with Terms and Summaries of Blog Posts

For the sample Wikipedia entries shown in Figure 2, both English and Japanese blog feeds / posts are retrieved and ranked according to the first approach of blog feeds/posts ranking described in the previous section. Along with this result of blog feeds/posts ranking by the first approach, blog post ranking by the search engine"Yahoo!" API for English and the Japanese search engine "Yahoo! Japan" API for Japanese are used as the second approach of blog post ranking. Then, characteristics terms within top ranked blog posts are shown through an interface for mining cross-lingual/cross-cultural differences between English and Japanese [1]. The interface also has a facility of showing top ranked blog posts, which enables cross-lingual blog analysis quite efficiently. Figure 3 illustrates

[6] Statistics are taken from the Internet Crime Complaint Center (IC3), U.S.A. (http://www.ic3.gov/), and National Police Agency, Japan (http://www.npa.go.jp/english/index.htm).

results of manual cross-lingual blog analysis, where summaries of blog posts are categorized into the following three types: (1) blog posts from a victim or one who personally knows a victim, (2) blog posts which warn others with reference to news posts on criminal acts, (3) blog posts which introduce how to prevent criminal acts. In the figure, samples of related terms automatically extracted from Wikipedia entries (those shown in Figure 2) are marked. Manually selected characteristic terms included in the blog posts are also marked.

In those results, it is remarkable to note that, both in the English and the Japanese blogosphere, many victims actually mention their cases in their own blogs. It is also important to note that we can collect many blog posts which refer to news posts or which introduce prevention tips. Differences in the English and the Japanese blog posts can be also detected, discovering that *"bank transfer scam"* is unique to Japan, and only was observed in the Japanese blogosphere[7].

4 Conclusion

This paper proposed a framework of cross-lingually analyzing people's concerns, reports, and experiences on crimes in their own blogs. There exist several works on studying analysis of concerns in multilingual news [4–6], but not in blogs. Future works for cross-lingual blog analysis on facts and opinions include incorporating multilingual sentiment analysis techniques and automatic extraction of reports or experiences of victims of crimes.

References

1. Nakasaki, H., Kawaba, M., Yamazaki, S., Utsuro, T., Fukuhara, T.: Visualizing Cross-Lingual/Cross-Cultural Differences in Concerns in Multilingual Blogs. In: Proc. ICWSM. (2009) 270–273
2. Kawaba, M., Nakasaki, H., Utsuro, T., Fukuhara, T.: Cross-Lingual Blog Analysis based on Multilingual Blog Distillation from Multilingual Wikipedia Entries. In: Proc.ICWSM. (2008) 200–201
3. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC-2007 Blog Track. In: Proc. TREC-2007 (Notebook). (2007) 31–43
4. Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D., Steinberger, R.: Combining Information about Epidemic Threats from Multiple Sources. In: Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization. (2007) 41–48
5. Pouliquen, B., Steinberger, R., Belyaeva, J.: Multilingual Multi-document Continuously-updated Social Networks. In: Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization. (2007) 25–32
6. Yoshioka, M.: IR Interface for Contrasting Multiple News Sites. In: Prof. 4th AIRS. (2008) 516–521

[7] We collect only 4 blog feeds and 13 blog posts from the English blogosphere, while we collect 132 blog feeds and 2617 blog posts from the Japanese blogosphere.