2012 Seventh International Conference on Availability, Reliability and Security

A Private Information Detector for Controlling Circulation of Private Information through Social Networks

Midori Hirose
Department of Informatics,
University of Electro-Communications
Tokyo, Japan
e-mail: hirose@edu.hc.uec.ac.jp

Akira Utsumi
Department of Informatics,
University of Electro-Communications
Tokyo, Japan
e-mail: utsumi@se.uec.ac.jp

Isao Echizen
Digital Content and Media Sciences Research Division,
National Institute of Informatics
Tokyo, Japan
e-mail: iechizen@nii.ac.jp

Hiroshi Yoshiura
Department of Informatics,
University of Electro-Communications
Tokyo, Japan
e-mail: yoshiura@hc.uec.ac.jp

***Abstract*—A "private information detector (PID)" is described that helps control the circulation of a user's private information through a social network. It checks texts to be posted by the user on a social network and detects potential revelations of private information so that it warns the user or modifies the texts automatically. It can cope with a wide variety of expressions that might be revealing by using public information accessible through the Internet rather than a large knowledge base. Evaluation using about 7000 blog sentences showed that it has reasonably good performance in terms of true and false detection rates.**

Keywords-component; social network; privacy; Security

I. INTRODUCTION

A. Background

Online social networks are important infrastructures used by more and more users, as shown by the growing numbers of Facebook users (currently more than 800 million) and of Twitter tweets (more than 200 million per day). While they provide bases for communicating more effectively and enjoyably, they can also result in the revelation and circulation of private information, as is reported in the news almost every week.

Revelation can occur from various parts of a social network such as the user profile, posted texts (blog posts, comments, tweets, etc.), photos, and videos. Revelation from the user profile can be prevented to some extent by setting the disclosure scopes properly (e.g., "friends only") because the types of private information in the user profile are limited and relatively static. In contrast, revelations from posted texts are much more difficult to prevent because these texts are posted almost every day, and it is difficult to predict what information they will contain. A user should therefore not use a static disclosure setting for these texts but should consider the setting every time a new text is posted. However, this is a tiresome and error-prone task. Moreover, existing disclosure control is typically all-or-nothing; i.e., the whole text is disclosed or hidden, leading to either possibly revealing private information or impairing the enjoyment of SNS communication.

Much research has been done on controlling the circulation of private information in social networks. It includes surveying information disclosure in social networks [1][2][3][4], clarifying privacy risks by identifying attack methods [5][6][7][8][9][10], defining metrics for measuring privacy risks [11][12][13], and developing techniques for controlling information circulation [14][15][16]. There has been little research, however, on techniques for controlling the circulation of information in posted texts.

B. Our Contribution

We have developed a set of algorithms, comprising a "private information detector (PID)," that check texts to be posted by a user on a social network. If the PID finds phrases that could reveal private information about the user, the user is warned or the text is automatically modified so that the text does not reveal the private information. We expect that the PID will play a unique and important role in the privacy enhancement of social networks because the circulation of private information in texts is difficult to control by other methods, as mentioned above, and because revealing various types of information, such as the actual names of the user and his/her family members, affiliations, address, current location, political opinion, and schedule, can be harmful depending on the user and his/her situation.

The biggest challenge in designing the PID was enabling it to cope with a wide variety of expressions that might be revealing phrases. The definition of "private" depends on the user, and there is a wide variety of words and combinations of words for each item of private information. In addition, there are words that do not directly express private information but could enable its deduction. Suppose, for

978-0-7695-4775-6/12 $26.00 © 2012 IEEE
DOI 10.1109/ARES.2012.89

example, a user wants to keep the company for which he or she works secret because of various potential risks, e.g., the company could fire him/her for posting negative opinions about the company, a powerful party might be angered by the posted opinions and pressures the company to fire him/her, the posted opinions might draw the attention of an extremist, and people could establish unwanted contact with him/her. The company the user wants to hide depends on each user, and there are various words and combinations of words that represent a company including its formal name, common names, and abbreviations. Moreover, there are numerous expressions that, although they do not directly mention the company, could be used to deduce it (e.g., names of executives and famous employees, name of building where the company is located, and combinations of city name and business category).

One approach to constructing a detector that can handle this type of problem is to use a large set of vocabularies (with synonyms and ontologies) and inference rules. Such a detector would not be practical, however, because constructing a large knowledge base is costly, and customizing it for each u ser is virtually impossible. Another approach is to use techniques from text classification and learning (e.g., [17]). However, because the definition of privacy depends on the user, using these techniques for detection would be problematic because the classification and rules learned from a set of sample texts posted by specific users could not be applied to ones posted by another user.

The approach we have thus taken for our PID is to use public information accessible through the Internet. Evaluation of its performance showed that it has reasonably good performance in terms of true and false detection rates.

II. PREVIOUS RESEARCH

Previous research on controlling the circulation of private information in social networks can be divided into four types: surveying information disclosure in social networks, clarifying privacy risks by identifying attack methods, defining metrics for measuring privacy risks, and developing techniques for controlling information circulation. Regarding the first type, in 2005, Gross et al. analyzed 4,000 user profiles on Facebook and found that 99% of them had unlimited disclosure scope (i.e. to everyone) [1], 89% disclosed actual names, 88% disclosed birth dates, and 51% disclosed current residence. Viegas surveyed people's consciousness of privacy in writing blogs and found that 55% of the writers disclosed their actual name, 21% wrote the names of their friends, 66% wrote about their friends without permission, and 76% used the unlimited disclosure setting [2]. In 2008, Lewes et al. reported that about one-third of 232 users used limited disclosure settings while the remaining two-thirds used the unlimited settings [3]. In 2010, Meeder et al. analyzed 2.7 billion Twitter messages and 80 million Twitter profiles. They found that retweeting tweets containing private information would violate the original tweeter's privacy because the original tweets were intended only for that person's followers [4].

Regarding the second type, clarifying privacy risks, Novak showed that online messages written by the same person can be identified on the basis of text level similarity in 2004 [5]. In 2007, Berthold proposed a formal concept analysis for linking a message and its author on the basis of the message content [6]. Backstrom identified active and passive attack methods for de-anonymising social network accounts [7]. A subgraph in the anonymised network is identified actively or passively and its nodes (the target user and his/her friends) are de-anonymised. In 2008, Lam found that real names of 80% users can be identified by analyzing one-line comments posted by friends in a social network called Wretch [8]. In 2010, Narayanan et al. also proposed a linking method based on subgraph matching and used it to link the Twitter and Flicker accounts of the same users [9]. Hasel showed that private information can be obtained through APIs of social networks [10].

Regarding the third type, identifying potential metrics for measuring privacy risks, Maximilien et al. described a metric for measuring privacy risk caused by information disclosure on Facebook in 2009 [11]. The metric is based the sensitivity of the disclosed information and the scope of the disclosure. In 2010, Yasui described a probabilistic metric for quantifying the degree of personal information disclosure from a blog [12]. A metric developed by Ngoc et al. uses joint entropy to measure risk caused by multiple information disclosures on different attributes of a person [13]. Regarding the fourth type, developing techniques to control information circulation, most social network services provide mechanisms for controlling the disclosure of information to, for example, friends, friends of friends, and everyone. Gurses et al. criticized these mechanisms, saying that the disclosure scope is not clear to novice users and that it is cumbersome for a user to have multiple identities based on this scoping [14]. In addition, as mentioned in Section 1, setting the scopes for posted texts is tiresome and error-prone, and existing disclosure control either reveals private information or impairs the enjoyment of SNS communication because of its all-or-nothing feature.

There has been little research, however, on countermeasures against revelations of information in posted texts [15][16]. The research that has been done includes the work by Hart et al. on a plan to develop a language for defining access policies for blogs and on an access control mechanism based on the policies defined and on the blog content, but the details were not reported [15]. They also developed an algorithm that learns the relationship between blog topics and disclosure scopes from the user's previous scope settings [16].

III. MODELS AND REQUIREMENTS

We modeled the PID application environment as shown in Figure 1. When the user inputs a text to a social network client system, the PID reads the text, recognizes the component words, and identifies any word or any combination of words that could reveal private information. The detection results are sent to a warning process that enables the user to control the circulation of his/her private information by modifying the text manually. They are

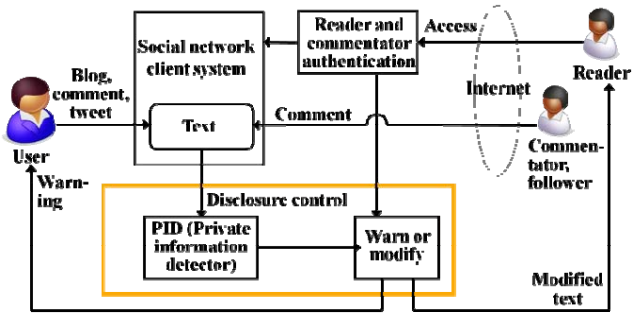

Figure 1. Environmental model.

alternatively sent to a modifying process that automatically modifies the text so that it is less revealing.

We modeled an attacker as a person looking for private information disclosed in posted texts. If a text contains a direct expression of private information, the attacker immediately recognizes the information. If the text contains expressions that give a clue about private information, the attacker guesses the information. Though there are various ways of guessing, most attackers use the Internet; i.e. the attacker selects words from the text that he/she thinks are related to the user's private information and uses a search engine with these keywords to find the private information.

On the basis of these models, our PID should take into account various expressions for the same item of private information and should take into account not only each word but also combinations of words. It also should take into account not only direct expressions of private information but also indirect expressions from which an attacker might deduce private information.

IV. ALGORITHMS FOR PID

The PID is implemented as illustrated in Figure 2. As a preprocess (not shown), a natural language analyzer recognizes words in the text. Then, using a given list of ***sensitive phrases*** that directly express private information about the user, the PID applies two types of detection (direct and indirect) to the recognized words and phrases, sentence by sentence, sensitive word/phrase by sensitive word/phrase. Note that a sensitive phrase can be a single word as a special case.

The direct detection algorithm identifies words that directly express private information by searching the text for words that are sensitive phrases. It cannot cope with abbreviations and synonyms of a sensitive phrase because it simply matches words in the text to the sensitive phrases in the list. It also cannot cope with indirect expressions of private information.

The indirect detection algorithm can cope with those varieties of expression by judging ***reachability*** from the words in the text to the sensitive phrases in the list. It detects abbreviations and synonyms for a sensitive phrase if reachability from these abbreviations and synonyms to the sensitive phrase holds and detects words that indirectly express private information if reachability from these words to the sensitive phrase holds. The reachability judgment simulates an attacker using the Internet to find private information. It extracts keywords from the text, using them to retrieve online content related to the text, and checking the retrieved content for its relevance to sensitive phrases.

The algorithm currently works sentence by sentence. It will be extended so that it can handle multiple sentences at a time. First, consider the situation in which a judgment is made using only the first sensitive phrase (Figure 3). The algorithm takes all the words in the sentence as input and eliminates words, such as articles, prepositions, and conjunctions, that are not suitable search keywords, so that ***n*** words remain. It then generates all the combinations of at most ***m*** words from the ***n*** words, resulting in ***S*** combinations, where

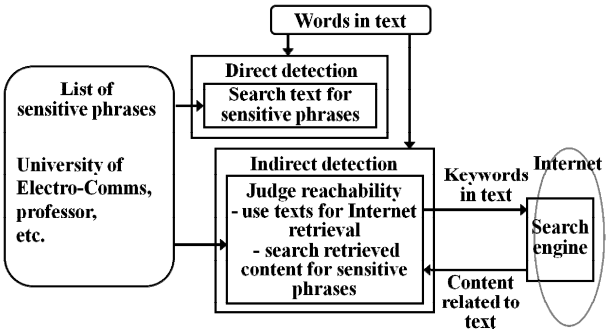

Figure 2. Private information detector

$$S = \sum_{i=1}^{m} \binom{n}{i} \tag{1}$$

The algorithm uses a search engine ***S*** times, once for each combination as a keyword list, where C_j is the j-th combination. For each search, the first ***k*** retrievals are checked for the number of times the sensitive phrase appears, resulting in count v_j These counts are summed up, and the sum is normalized by being divided by ***S***:

$$TRS = \frac{1}{S} \sum_{j=1}^{S} v_j \tag{2}$$

where TRS, or ***total risk score***, is the normalized total count for that sentence and that sensitive phrase. If TRS is larger than or equal to a threshold *T*, the algorithm judges that reachability from the sentence to the sensitive phrase holds, meaning that the sentence reveals private information. This process would be repeated for each sensitive phrase.

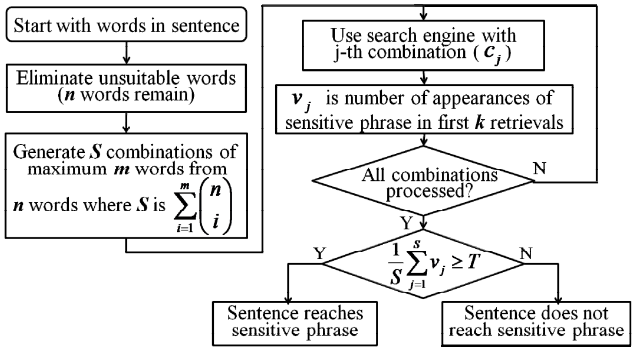

Figure 3. Algorithm for reachability judgment.

A. *Defining sensitive phrases*

Because the definition of privacy depends on the user, sensitive phrases need to be defined for each user. A simple but reasonable way of defining sensitive phrases is to use the security settings on the user's profile. Most social network services provide security settings with which the user can control the disclosure of each attribute (e.g., name, affiliation, address, marital status) and to which extent (e.g., to friends, friends of friends, everyone). Unless the extent is set "to everyone," the attribute is taken as a sensitive phrase. For example, if the user's affiliation is "company A" and the extent is set "to friends," "company A" is taken as a sensitive phrase. That is, if the disclosure of an information item in the user's profile is limited by the user, direct and indirect revelations of this item in texts to be posted are detected by the PID.

V. EVALUATION

A. *Implementation*

Since the current implementation is for Japanese sentences, we used the Mecab [18] Japanese language analyzer to recognize words in the sentence. For English sentences, an English language analyzer, such as TreeTagger [19], could be used because all the processes except for the language analysis are language independent. We used the Yahoo API for the search engine. The number of appearances of each sensitive phrase (v_j) is counted in the titles of the first ***k*** retrieved items. We used "Mixi," the largest social network in Japan (23 million users), as the social network client system. The PID was automatically activated when the user posted any text.

B. *Evaluation*

We evaluated only the indirect detection algorithm because the performance of the direct detection algorithm is obvious; i.e., it can detect word sequences in the sentence if it is included in the list of sensitive phrases. We used 7462 sentences in 158 actual blog posts posted on Mixi by a male professor at our university, the University of Electro-Communications. When writing these posts, he wanted to hide his university affiliation so that he could express his opinions without worrying about potential risks (of being monitored, his opinions being copied on public electronic boards and criticized, etc.). He also wanted to hide his occupation so that his being a professor would not affect readers' opinions about his posts. He therefore registered one sensitive phrase, "University of Electro-Communications", for his university and another one for his occupation, "professor", on the system's list.

We first checked the 7462 sentences manually and first eliminated 415 sentences that directly mentioned the sensitive phrases to evaluate performance of the indirect detection only. Among the remaining 7047 sentences, we found that 53 and 52 sentences respectively indirectly expressed "University of Electro-Communications" and "professor." We used this manual judgment as the reference for the evaluation.

We then applied the indirect detection algorithm to the 7047 sentences. The maximum number of words in a keyword list, ***m***, and the number of top retrievals to check, ***k***, were respectively set to 3 and 24 because studies have shown that people use search engines with an average 2.21 keywords [20] and look at the top 23.5 retrievals on average [21]. The results for "University of Electro-Communications" are shown in Figure 4. The horizontal axis represents the total risk score for revealing the university, and the vertical axis represents the number of sentences having the corresponding TRS. The black bars show the distribution of the 53 sentences that reveal the user's university, and the white ones show the distribution of the 6994 sentences that do not. If we set the threshold to 0.30, the algorithm detects 74 sentences: 48 sentences that indirectly express the university's name and 26 that do not. Thus, we have a true detection rate of 0.91 (48/53) and a false detection rate of 0.35 (26/74).

Table 1 lists example sentences judged to reveal the user's university with the number of keyword lists (***S***), the number of times the phrase "University of Electro-Communications" appeared in the retrievals, and the TRS. These example sentences include a specific department name ("ICE") and an education topic ("information security") (sentence 1), a specific building name ("West-6") and the phrase "job fair" (2), the name of an area near the university ("Chofu") and the word "students" and phrase "graduate research" (3), or the abbreviation for the university name ("UEC") (4).

In sentences (1), (2), and (3), the university name is not directly expressed but can be deduced from the combination of words. In sentence (4), an abbreviated name of the university is used. Our indirect detection algorithm coped with these various expressions and combinations by using neither corresponding vocabularies nor inference rules but by simply using one vocabulary item, "University of Electro-Communications."

The corresponding results for the user's occupation are shown in Figure 5 and Table 2. A threshold of 0.30, the same value used for the university, results in the detection of 82 sentences, including 36 that indirectly express "professor" and 46 that do not. Thus, we have a true detection rate of 0.69 and a false detection rate of 0.56. Sentences judged to reveal the occupation include the phrase "supervising students' research" (sentence 1), the words "paper,"

TABLE I. SENTENCES JUDGED TO REVEAL USER'S UNIVERSITY

No	Sentence (translated into English)	S	$\sum_{j=1}^{S} v_j$ (TRS)
1	In the ICE Department, we teach information security.	3	5 (1.67)
2	Job fair in the West-6 Building.	7	3 (0.34)
3	In Chofu, I must always be quiet because there are students doing graduate research even late at night.	14	5 (0.36)
4	Big advertising display of UEC just in front of the station exit.	14	50 (3.57)

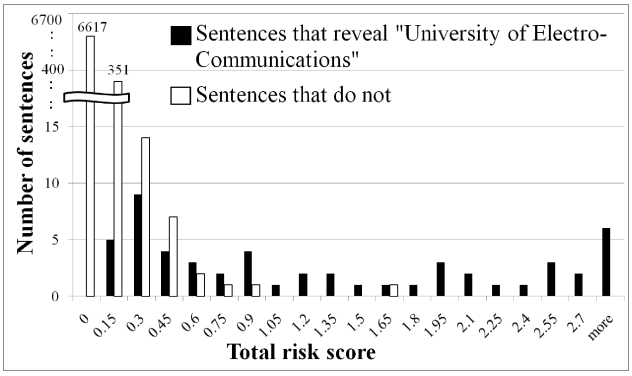

Figure 4. Evaluation result for "University of Electro-Communications"

TABLE II. SENTENCES JUDGED TO REVEAL USER'S OCCUPATION

No	Sentence (translated into English)	S	$\sum_{j=1}^{S} v_j$ (TRS)
1	Supervising students' research for 8 years, but it is still fun for me.	7	10 (1.43)
2	A paper written by a student of mine received an award.	7	3 (0.43)
3	Too busy attending conference on day I returned from overseas business trip.	3	1 (0.33)

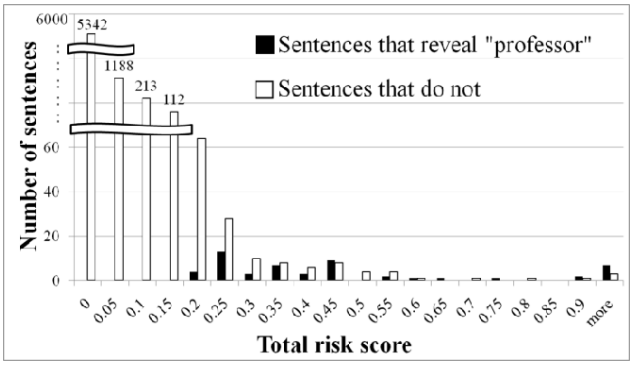

Figure 5. Evaluation result for "professor".

"student," and "award" (2), or include the word "conference" and the phrase "overseas business trip" (3). As for "university," our indirect detection algorithm coped with these various expressions and combinations by simply using only one vocabulary item, "professor."

Although the detection performance was not excellent, particularly the high false detection rate for "professor" (i.e., 56% of the alarms produced by the PID would be false), the rate of false alarms was low in terms of total blog sentences, less than 1% (46 alarms out of 7047 sentences), for the "professor" case. In addition, the false alarms could be made less annoying by using a passive user interface[1] such as one typically used for grammar checkers. Coping with a wide variety of words and phrases that could reveal private information is not an easy undertaking. Considering the challenging nature of this problem, it is fair to say that the performance of our algorithm is reasonably good.

VI. CONCLUSION

Detecting phrases that could reveal private information plays an important and unique role in controlling circulation of private information through social networks. We developed a private information detector (PID) that can cope with a wide variety of expressions that might be revealing phrases. Evaluations using 7047 blog sentences showed that the PID has reasonably good detection performance. Our follow-up projects include:

- Develop a method for identifying the words that reveal private information. This can be done by observing the number of occurrences of a sensitive phrase in the retrievals made with the j-th combination of words (as shown in Figure 3).
- Implement a system that warns the user so that the user can control the circulation of his/her private information by modifying the text.
- Implement a system that automatically modifies texts so that they are less revealing. Research would be needed to establish an intelligent modification system that replaces revealing phrases with less revealing ones (e.g. those at a higher level in an ontology) while keeping the communication informative.
- Apply the PID to other social networks, such as Facebook and Twitter, and to various other types of media.

To the best of our knowledge, the PID is the first one that makes use of public information accessible through the Internet to detect privacy risks. Our research should thus inspire new research in this direction.

REFERENCES

[1] R. Gross, and A. Acquisti, "Information revelation and privacy in online social networks," In Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society (WPES), pp. 71-80, New York, 2005.

[1] This means an interface that simply shows (e.g., underlines) sentences judged to be ungrammatical. It shows a detailed explanation only if the user requests one.

[2] F. Viegas, “Bloggers' expectations of privacy and accountability: an initial survey,” Journal of Computer-Mediated Communication, 10(3), article 12, 2005.

[3] K. Lewis, J. Kaufman, and N. Christakis, “The taste for privacy: an analysis of college student privacy settings in an online social network,” Journal of Computer-Mediated Communication, 14(1), pp. 79-100, 2008.

[4] B. Meeder, J. Tam, P. Kelly, and L.F. Cranor, “RT@ IWantPrivacy: widespread violation of privacy settings in the Twitter social network,” In Proceedings of the Web 2.0 Privacy and Security Workshop, Oakland, 2010.

[5] J. Novak, P. Raghavan, A. Tomkins, “Anti-aliasing on the Web,” In: Proc. 13th International World Wide Web Conference (WWW2004), pp.30-39, New York, 2004.

[6] S. Berthold, and S. Clauss, “Linkability estimation between subjects and message contents using formal concepts,” In Proceedings of the Workshop on Digital Identity Management, pp. 36-45, Virginia, 2007.

[7] R.Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou R3579X? anonymized social networks, hidden patterns, and structural steganography,” In Proceedings of the 16th International World Wide Web Conference (WWW2007), pp. 181-190, Banff, 2007.

[8] I. Lam, K. Chen, and L. Chen, “Involuntary information leakage in social network services,” In: Proc. the 3rd International Workshop on Security (IWSEC), LNCS 5312, pp.167--183, Takamatsu, 2008.

[9] A. Narayanan, and V. Shmatikov, “De-anonymizing social networks, In Proceedings of the 30th IEEE Security & Privacy,” pp. 173-187, Oakland, 2009.

[10] M. Hasel, and L. Iacono, “Security in OpenSocial-Instrumented Social Networking Services,” In Proceedings of the Communications and Multimedia Security, pp. 40-52, 2010.

[11] E. Maximilien, T. Grandison, T. Sun, D. Richardson, S. Guo, K. Liu, “Privacy-as-a-Service: Models, Algorithms, and Results on the Facebook Platform,” In Proceedings of the Web 2.0 Security & Privacy Workshop, 2009.

[12] R. Yasui, A. Kanai, T. Hatashima, and K. Hirota, “The Metric Model for Personal Information Disclosure,” In Proceedings of Fourth International Conference on Disital Society (ICDS2010), pp.112-117, 2010.

[13] T. Ngoc, I. Echizen, K. Kamiyama, H. Yoshiura: New Approach to Quantification of Privacy on Social Network Sites, in Proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 556-564, 2010.

[14] S. Gurses, R. Rizk, and O. Gunther, “Privacy design in online social networks: learning from privacy breaches and community feedback,” In Proceedings of the 29th International Conference on Information Systems, pp. 1-10, Paris, 2008.

[15] M. Hart, R. Johnson, and A. Stent, “More Content - Less Control: Access Control in the Web 2.0,” Proceedings of the Web 2.0 Security & Privacy Workshop, 2007.

[16] M. Hart, C. Castille, R. Johnson, and A. Stent, “Usable Privacy Controls for Blogs,” In Proceedings of the International Conference on Computational Science and Engineering, pp. 401-408, 2009.

[17] E. Stamatatos, “Author identification: Using text sampling to handle the class imbalance problem,” Information Processing & Management, 44(2), pp.790-799, 2008.

[18] T. Kudo, “MeCab. Yet another part-of-speech and morphological analyzer,” http://mecab.sourceforge.net/ (in Japanese).

[19] TreeTagger - a language independent part-of-speech tagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[20] iProspect: search engine user behavior study, 2006. http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf.

[21] B. Jansen, A. Spink, and T. Saracevic, “Real life, real users, and real needs, a study and analysis of user queries on the Web,” Information Processing and Management: an International Journal, 36(2), pp. 207-227, 2000.