Design of Categorization Mechanism for Disaster-Information-Gathering System

Shinsaku Kiyomoto*, Kazuhide Fukushima, and Yutaka Miyake
KDDI R & D Laboratories Inc.
Fujimino, Saitama, Japan
{kiyomoto, ka-fukushima, miyake}@kddilabs.jp

Abstract

Organizations that are responsible for disaster and emergency management are faced with the issue of how to gather reliable and useful information during a major disaster. We consider an information-gathering platform for large-scale disasters and emergencies based on mobile terminals. A simple solution to realize an information-gathering system is to construct a server where information is uploaded and published. However, such a centralized approach is not flexible nor is it robust. For example, it is very hard to find an appropriate system to which the user can upload information during a disaster, and the centralized server may be down because of overload or because it has been physically destroyed. We must consider a distributed and dynamic architecture for the system. Security and privacy issues are another concern that should be addressed for providing information from user's mobile terminals. Here, we focus on the design of a categorization mechanism for the information-gathering system. In this paper, we present a mechanism that has been designed so that it is able to accommodate requirements and we then conduct a feasibility analysis of the mechanism.

Keywords: security, privacy, information gathering, disaster, categorization.

1 Introduction

There are two major issues confronting IT systems during disasters [1].

System continuity management is the first issue for IT systems during disasters. Cloud computing environments have been considered a cost-effective solution for ensuring system continuity. Wood *et al.* performed a pricing analysis to estimate the cost of running a public cloud-based disaster-recovery service and showed significant cost reductions compared to using privately owned resources [2]. Cloud computing environments are also robust in the context of wide-area disasters, and cloud services have been used for system continuity management.

Another issue confronting an organization that is responsible for disaster and emergency management is how to gather reliable and useful information during a major disaster. Internet search engines are not an effective means for searching for information about a disaster, and sometimes information overflow occurs. There are several studies on how to construct a disaster management system using computer networks. The main topic is how to support sharing of information about the current disaster and the status of resource allocation for emergency management. Currently, user-centric systems using mobile terminals are recognized as new approaches to achieving a more efficient information-sharing system. It has been suggested that SNS and micro-blogs are effective systems for communication and sharing information during a major disaster.

In this paper, we consider an information-gathering platform for use during disasters and emergencies based on mobile terminals. A simple solution for setting up an information-gathering system is to

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, volume: 3, number: 4, pp. 21-34
*Corresponding author: 2-1-15 Ohara, Fujimino, Saitama, 356-8502, Japan, Tel: +81-49-278-7885

construct a server to which information is uploaded and published. However, such a centralized approach is not flexible and nor is it robust. For example, it is often difficult to find an appropriate system to which the user can upload information in a disaster, and the centralized server may be down because of overload or because it has been physically destroyed. We must consider a distributed and dynamic architecture for the platform.

Security and privacy concerns should be addressed when providing information from user's mobile terminals using the platform. If the identity of users can be kept anonymous from governmental organizations, users will find it acceptable to send information to such organizations. Thus, we have to consider a privacy-preserving data upload mechanism for a distributed and dynamic architecture when developing an information gathering system for disasters.

We design a categorization mechanism for the disaster information-gathering system. The disaster information-gathering system requires that mobile terminals add a message category tag and location information to their send messages. To improve usability, the tag for each message should be selected from among several categories automatically, and added to the send messages. Furthermore, the concern that users have regarding privacy concern should be considered. Messages sent by the user terminals have to be kept secret from intermediate entities such as the control servers in the system. We design a mechanism that can be executed on the user terminal and generates messages appropriate for privacy requirements. The feasibility of the mechanism is also discussed in this paper.

The rest of the paper is organized as follows; section 2 introduces existing research related to the issue. An overview of the system is provided in section 3 and we present our categorization method in section 4. In section 5, we conduct an evaluation of the method and present the results, and the conclusion of this paper is presented in section 6.

2 Related Work

Developing emergency and disaster management systems is an important issue for our modern computer-oriented society. The primary issue is how to share information about a current disaster and the status of resource allocation to facilitate emergency management. Atteih *et al.* presented a case study [3] on the implementation of an Emergency Management Information System (EMIS) in support of emergency responders. The incident management system (IMS) [4] proposed by Perry was a tool for marshaling pre-identified and pre-assembled resources for responding to an emergency or disaster. Yao *et al.* built a system [5] that allowed virtual teams of experts to create and discuss the emergency scenario. Collabit [6] is a virtual dashboard that facilitates distributed asynchronous sharing of information in an emergency. Wickler *et al.* considered the use of new media technologies, including virtual worlds on the Internet, for collaboration in disasters [7]. Shklovski *et al.* presented evidence of ICT use [8] for reorientation toward the community and for the production of public goods in the form of information dissemination during disasters. Jang and Tsai proposed a MANET-based emergency communication and information system [9] that could support a large number of rescue volunteers during catastrophic natural disasters. Research [10] by Dilmaghani and Rao identified a set of potential network oriented problems in existing inter-organizational communication protocols incorporating the information collected from several drill participations and after interviewing first responders.

Systems using mobile terminals for the management of a disaster have been considered. Fajardo and Oppus proposed a disaster management system [11] that facilitates the logistics for rescue and relief operations. The system provides the optimum route for rescuing people in a disaster. Zeng *et al.* proposed a mobile communication system [12] for evacuations during emergencies.

SNS and micro-blogs are very useful tools for communication and sharing information even under the conditions created by a major disaster. In this paper, we demonstrate the security and privacy require-

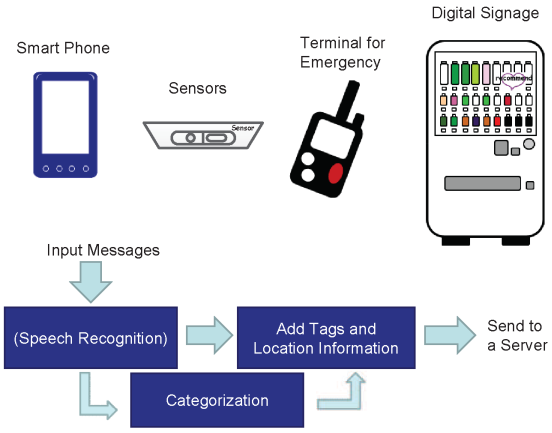

Figure 1: Client Terminals

ments for the information-gathering platform and consider the necessary functions for the platform.

A categorization algorithm is a key function in our mechanism. Spam filtering is a categorization technique for text messages,[13] and some filtering techniques are good examples for a client-based architecture. Spam filtering techniques categorize email messages into two groups: Spam or not Spam. Current practical email filtering systems mainly rely on email-specific features, such as the email header or the domain names of links embedded in the email [14]. The Robinson-Fisher (RF) algorithm [15] is the most commonly used Spam filtering algorithm. The RF algorithm realizes efficient categorization within a feasible transaction time.

Anomaly-based intrusion detection is a special case of data analysis. Two-group categorization of data is required for anomaly-based intrusion detection systems (IDS); the IDS categorizes data into normal traffic or attack traffic. An anomaly detection algorithm based on the Naive Bayes classification algorithm [16] has been proposed by Maxion and Townsend. Julisch and Dacier presented a new conceptual clustering technique for an intrusion detection system [17]. Lee and Stolfo use data mining for the construction and training of classifiers that detect intrusions [18]. SmartSifter [19] is an outlier detection engine addressing fraud detection based on statistical learning theory. For current intrusion detection systems, the trend is to optimize the method of analysis using data mining techniques for particular attacks or applications.

Clustering methods are key techniques for solving the problems addressed in this paper. Several methods have been proposed and compared with other methods in terms of their classification performance [20]. There are three major coefficients for categorization algorithms: the overlap coefficient (Simpson's coefficient) [21], the Jaccard coefficient, and the dice coefficient[22]. In this paper, we selected an appropriate clustering method and optimized it to analyze the importance of the data. The accuracy of the analysis is generally improved by feedback of the analysis results. Thus, we also considered an efficient algorithm for automatic updates of the database without increasing the size of the database or decreasing categorization accuracy.

3 Disaster-Information-Gathering System

In this subsection, we introduce an overview of the disaster-information-gathering (DIG) system.

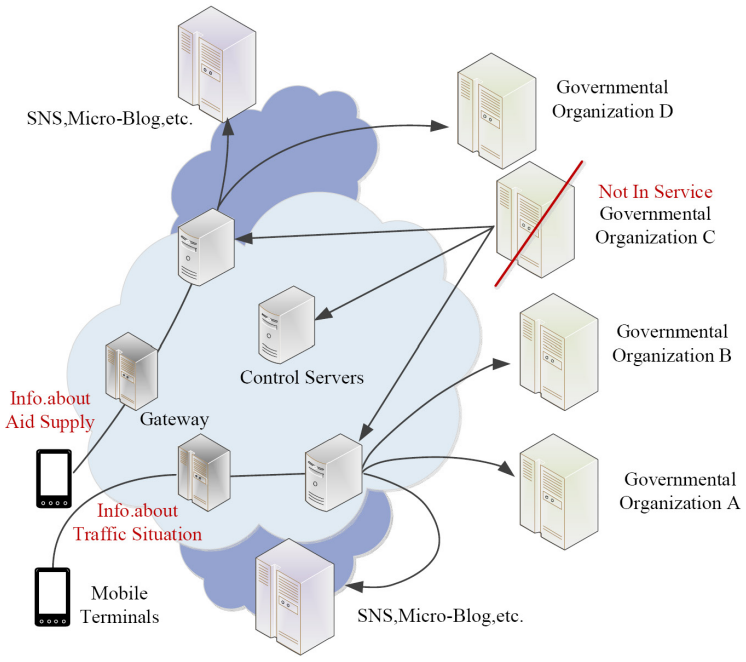

Figure 2: System Overview

3.1 Client Terminal

Client terminals send a message about a disaster to a governmental organization in the disaster-information-gathering system. We assume several client terminals such as smart phones, voice sensors, terminals for emergencies, and digital signage. The message is input into the client terminals as a voice message or text message, and then tag information that indicates the category of the message and location information are attached. For voice messages, a speech recognition function for the client terminals is needed in order to convert the voice message into a text message; a speech recognition mechanism is outside the scope of this paper. Figure 1 shows examples of the client terminals and the procedure used to send a message to a governmental organization.

3.2 System Overview

An overview of the disaster information gathering system is shown in Figure 2. The gateway manages the connections between the mobile terminal and mobile networks. In a disaster, messages to a commercial SNS or micro-blog system are copied and transferred to the systems of corresponding organizations, where a user accepts the responsibility for providing information to the organizations. The messages have a tag that describes the type of information, and the control server selects the appropriate system according to the tag. The organizations can gather information about the disaster and about people who need support. An example of a governmental system for information gathering is shown in Figure 3.

3.3 Insider Threats in the System

The system is a hybrid system that consists of systems and networks operated by governmental organizations and other systems and networks operated by mobile operators. The governmental organizations should be trusted, but the mobile operators are not fully trusted due to malicious insiders. We should

Figure 3: Example of Assumed Governmental Organization System

design security functions to consider malicious insider threats in mobile operator's networks. A control module is executed on the control server by the mobile operators. That is, insider threats from mobile operators need to be taken into consideration for the control servers as well as the network.

We assume that the control servers are to be honest, but users do not want to provide more privacy information than is required. The users may not be willing to accept that the mobile operators are able to obtain messages, even though the users allow the governmental organizations to read the messages. Thus, all the message data except for information required for the operation of control servers should be encrypted on the client terminals. The encryption scheme should not require off-line key pre-sharing between users and governmental organizations; thus, we should use a public key encryption scheme for the message encryption. The user terminal only stores a common public key of governmental organizations for the encryption scheme.

3.4 Requirements for Categorization

The categorization mechanism is an important function for realizing the disaster-information-gathering system. We consider the categorization mechanism in this paper. We design a categorization method that satisfies the following requirements;

- Tag information for each message is needed in order to deliver an emergency message. To avoid leakage of message contents, we execute a categorization mechanism for making the tag on each mobile terminal.
- Messages should be kept secret from intermediate entities between users and governmental organizations. Thus, messages are encrypted on each mobile terminal.

3.5 Control Module

The role of the control module is to copy messages from user terminals and to transfer them to appropriate governmental organizations. The messages that the control server receives include location information

and a tag that indicates the type of information, such as traffic conditions, rescuer information, and aid supply information. The control server has information about governmental organizations, the location and coverage area of the organizations, and the current status of the organizations. The control server chooses the appropriate governmental organizations according to the location information and the tag and sends the copied message to an organization that requires information of that type. The procedure of message forwarding on the control server is as follows:

1. The control server receives a message from a user terminal. Then, the control server makes a copy of the message, removes tags and location information, and sends the original message to the SNS or micro- blog service, if the user wants to upload the message to the service.

2. The control server checks the tag information and the location information and finds an appropriate governmental organization using a database of governmental organizations. Note that the control server simply deletes the copied message when the tag information is ”not important.”

3. The control server sends the copied message to the governmental organization found in the previous step.

The control server is distributed in mobile networks and checks the current status of the systems by frequently accessing the system. If the system of a governmental organization is damaged by a disaster or the organization has insufficient human resources to help people, the control server automatically selects a system from another organization.

3.6 Example Scenarios

Here, we show some example scenarios:

- **Scenario 1.** Users upload traffic information to the SNS or micro- blog services; for example, some trains have stopped running or stations are closed, there are traffic jams, or there are obstructions on the roads that make it hard to walk or drive. In this situation, the information is copied and transferred to the governmental organizations responsible for traffic control in order to provide support for evacuation of a disaster area.

- **Scenario 2.** A user updates information to SNS or micro-blog services about shortages of aid supplies. The information is copied and transferred to the nearest governmental organization responsible for aid supplies. If the governmental organization does not have such supplies, the information is transferred to other governmental organizations near the location of the user.

- **Scenario 3.** If a user discovers an emergency involving the collapse of a house and gas leaks, the user would upload such information to the SNS or micro-blog services. In this case, the information is copied and transferred to governmental organizations (rescuer or police) responsible for the area near the location.

4 Design of Categorization Mechanism

In this section, we explain the categorization mechanism. We design a categorization mechanism according to the following principles;

- The categorization method is executed on the client terminals. The computational cost of the categorization is feasible for mobile terminals and the size of the database for the categorization is appropriately small.

- The accuracy of the categorization is practically high.

The mechanism runs on the mobile terminals and it includes two kinds of functionality: categorization and message encryption.

The categorization algorithm finds the appropriate category for the data using the signatures for categories. The signatures are preliminarily generated using labeled (already categorized) documents. The JNW algorithm evaluates a set of words in the data for comparison with signatures in a signature database. The algorithm is based on two coefficients: the Jaccard coefficient and a new-word-based coefficient. We conducted an experiment to evaluate the three major coefficients for categorization algorithms: the overlap coefficient (Simpson's coefficient) [21], the Jaccard coefficient, and the dice coefficient[22]. These coefficients evaluate the similarity of the document and signatures. The Jaccard coefficient is the best coefficient for small training data sets. Thus, we chose the Jaccard coefficient. Furthermore, we considered how the accuracy of the algorithm could be improved when combined with another coefficient. A coefficient that complements the Jaccard coefficient should be added to improve the categorizing accuracy. We proposed a new-word-based coefficient as a complement to the Jaccard coefficient; the algorithm evaluates the difference between the document and the signatures. We compared the JNW algorithm with the RF algorithm, which is a fast and accurate categorization algorithm, and found that the accuracy of the algorithm is superior to that of the RF algorithm. Thus, we selected the above two coefficients for the algorithm.

4.1 Indices for Categorization Algorithm

In this subsection, we present two indices for the categorization algorithm.
Amended Jaccard Coefficient. The Jaccard coefficient [23] is defined as the size of the intersection divided by the size of the union of the sample sets. The Jaccard coefficient is generally calculated as $\frac{|m \cap m_i|}{|m \cup m_i|}$, where m is a set of words included in a target document and m_i is a set of words included in a document i of group C_j. The formula $|x|$ denotes the number of elements in x. We obtain the Jaccard coefficient to calculate the ratio of the number of words included in both the target document and the document of group C_j. The Jaccard coefficient tends to include errors depending on the size of group C_j and needs to amend a normalized number of elements, n. We use $\log(n) \cdot \langle \frac{|m \cap m_i|}{|m \cup m_i|} \rangle_{m_i \in C_j}$ as a new amended Jaccard coefficient obtained by the following theorem, where $E(x)$ is the expectation value of x.

Theorem 1. If it is assumed that the probability density function $f(x)$ for each X_i is an exponential distribution function, the expectation value of the maximum Jaccard coefficient value X_{max} is estimated as $\langle X \rangle \mathbf{log}(n)$ for a sufficiently large n, where $\langle X \rangle$ is the expectation value of Jaccard coefficient values.

Proof. We define $F_{max}(x)$ as the probability distribution function of X_{max} and $X_{max} = max[X_1, X_2, ..., X_n]$.

$$F_{max}(x) = Pr[X_{max} \leq x]$$
$$= Pr[(X_1 \leq x) \wedge (X_2 \leq x) \wedge ... \wedge (X_n \leq x)]$$

It can be assumed that each event is independent and it is calculated using $F(x)$, which is the probability distribution function for each X_i as follows:

$$F_{max}(x) = F(x)^n$$

Now, we assume $f(x)$ is the exponential distribution function, that is $f(x) = \lambda e^{-\lambda x}$ $(x > 0)$. $F(x)$ is calculated as $F(x) = 1 - e^{-\lambda x}$ $(x > 0)$. From [24], we obtain

$$X_{max} = \sum_{t=1}^{n} = \frac{1}{\lambda t}$$

Thus,

$$X_{max} = \frac{1}{\lambda k} \approx \frac{1}{\lambda}\mathbf{log}(n) = \langle X \rangle \mathbf{log}(n) \quad \square$$

Note that, if the calculated value for the amended Jaccard coefficient was larger than 1, we used 1 as the value for the amended Jaccard coefficient. Furthermore, we decided that no amendment was required and used the original Jaccard coefficient value, where $\frac{|m \cap m_i|}{|m \cup m_i|} > \log(n) \cdot \langle \frac{|m \cap m_i|}{|m \cup m_i|} \rangle_{m_i \in C_j}$. Thus, we calculate the amended Jaccard coefficient $J_n(C_j)$ of the group C_j as follows:

$$J_n(C_j) = max\left[min\left[1, \log(n) \cdot \left\langle \frac{|m \cap m_i|}{|m \cup m_i|} \right\rangle_{m_i \in C_j}\right],\right.$$

$$\left. max\left[\frac{|m \cap m_i|}{|m \cup m_i|}\right]_{m_i \in C_j}\right]$$

New-Word-Based Coefficient. The Jaccard coefficient evaluates the overlaps of words in the target document and the words in documents of the group. On the other hand, the new-word-based coefficient calculates the differences between the sets of words in the target document and the sets of words in the document of the group. Thus, the coefficient carries out a function that complements that of the Jaccard coefficient. The new-word-based coefficient counts the number of words included in the target document and not included in group C_j. The algorithm decides whether the document is in group C_j to find group C_j, which has the minimum value of the following coefficient:

$$W_n(C_j) = r_{C_j} \log(|C_j|)$$

where r_{C_j} is the number of words included in the target document and not included in group C_j.

4.2 Two-stage JNW Algorithm

The algorithm decides whether the target document is categorized into group C_j or not by using the two indices explained in 4.1. The algorithm consists of two steps as follows;

1. Let $\overline{C_j}$ be the complement set of C_j ($\overline{C_j} = \bigcup_{k \neq j} C_k$) and n is $max[|C_j|, |\overline{C_j}|]$. The algorithm compares the NW coefficients $W_n(C_j)$ for C_j with the coefficients $W_n(\overline{C_j})$for $\overline{C_j}$. The algorithm counts the number of group such that $W_n(C_j) = W_n(\overline{C_j})$. The algorithm outputs C_j where the number is 1, otherwise the algorithm executes *step 2*.

2. The algorithm outputs a group C_j that has the maximum value of $J_n(C_j)$.

Figure 4 shows a diagram of the two-stage JNW algorithm.

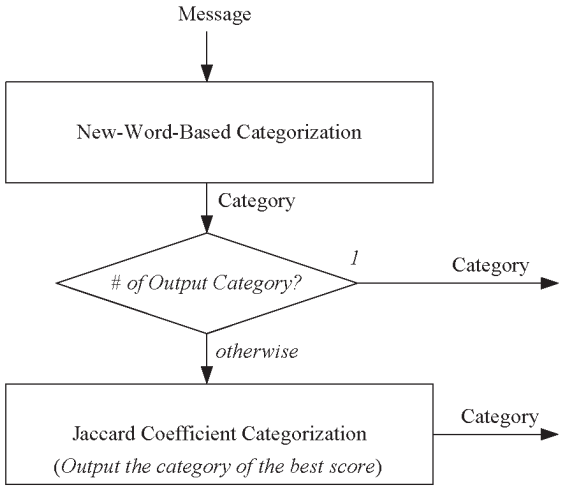

Figure 4: Two-Stage JNW Algorithm

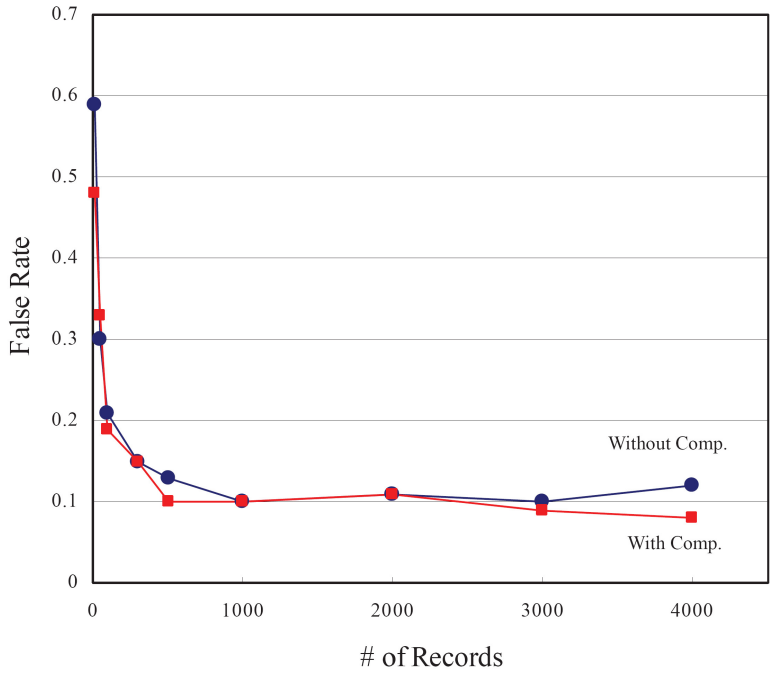

Figure 5: False Rate of Categorization Algorithm

4.3 Training and Updating Signatures.

The algorithm requires signatures to determine the appropriate category. The signatures are made using known labeled documents where the categories are given. Words are obtained from the given documents in order to create the signature. The N-gram model is used for obtaining the words from the documents, when the number of records in the signature database is small. We stored 1-gram, 2-gram and 3-gram words in additon to words from a morphological analysis, when the number og records is lower than 1000.

We assume that a governmental organization preliminarily generates a signature for each category from the labeled documents, and an initial signature database is distributed to all user terminals. The signature is also updated according to the current send/receive data, because the signature should be customized for each user. The signature data for the JNW algorithm consists of data records where each record is defined for each labeled document belonging to the category. Words selected from the labeled document are stored in the record. The signature data for each category consist of at most v records generated from each document belonging to the category. The record includes an l column that consists of a word m_i selected from document C_i and its term frequency $TF(C_i, m_i)$. Each record has data of l words in the document C_i. The term frequency $TF(C_i, m_i)$ is calculated as $d_i / \sum_i d_i$, where d_i is the number of words m_i in document C_i.

In an initial phase, the algorithm uses the signatures generated from the labeled documents to categorize the send/received data, and after categorizing the data, the algorithm adds the data to the signatures according to the categorization results and/or the user feedback. By frequent execution, the size of the signatures increases, even though the accuracy of the categorizing improves. Thus, a signature update process is needed to keep the size of the signatures constant.

A compression function is used for the signature update. Our data compression for the signatures consists of two steps: combining some records and reducing the number of words in the combined record. The compression function is executed for a signature when the number of records in the signature is more than v and consists of the four steps as follows:

1. The function calculates all amended Jaccard coefficients for any pair of two records, and finds a pair of two records with the maximum value of the coefficient.

2. The function creates a combined record from the two records to merge all words in both records.

3. If the combined record has more than l words, the function refers to the TF values of all words in the combined record and removes any words with the lowest TF value until the number of the words in the record is less than l.

4. The function executes steps from 1 to 3 until the number of records in the signature is less than v.

5 Evaluation

We implemented the prototype system on a smart phone (1.5GHz CPU) and evaluated the performance and accuracy of the categorization. it is assumed that a short message such as the *Twitter* [25] service is used in the system. We examined the accuracy of the categorization method using dataset of one sentences collected from messages on the Web. We picked up sentences less than 140 characters from news articles of three different categories on the Web: science(computer) news, economics, and sport. The sentences were divided into two data: data for the training, and data for the experiments. All results are average values of 1000 trials. Figure 5 shows the false rates for the categorization with/without the signature compression. In the case where 1000 records were stored in the database, the algorithms had a false rate of 1.0 %. The difference in the false rates between the no compression case and the compression case of the update algorithm is not remarkable. In the experiment, the compression process was started when over 1000 records has been included in the signatures. The compression did not seriously affect the false rate; which remained at about 1.0 % when we used compressed signatures from 1000 records.

Transaction time for the categorization with/without the signature compression is shown in Figure 6. When the number of records exceeded 1000, the update of the signatures was started. The transaction

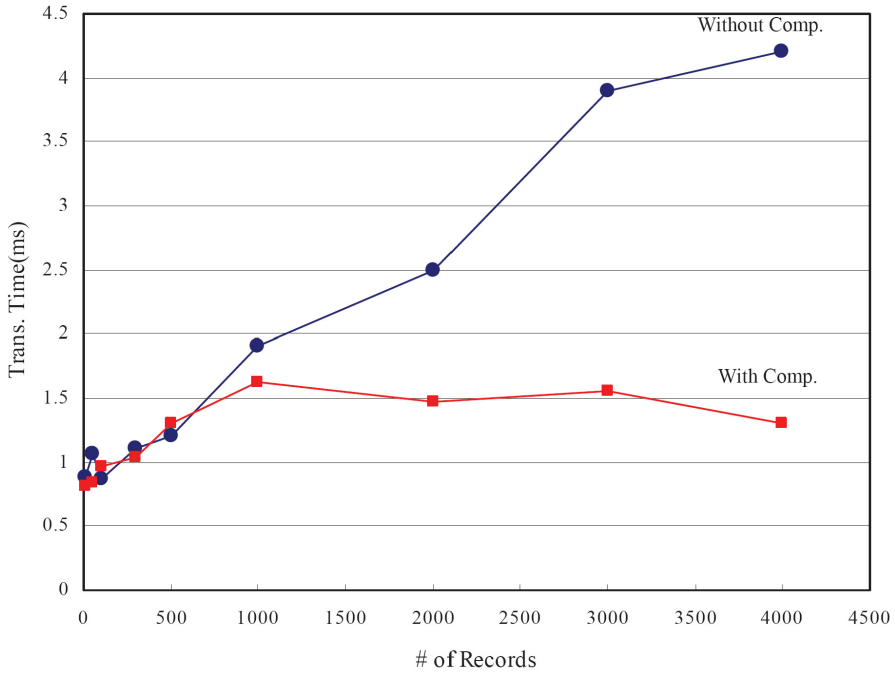

Figure 6: Transaction Time of Categorization

time of categorization is below 5.0 msec for 4000 records, and the transaction time can be constant where we use the signature compression. Transaction time of signature update is 134 msec when the database has 1000 records, and the transaction time was almost constant even when the number of the records increased. Thus, the proposed algorithm achieved feasible transaction times for categorization of data and update of the signatures, even though performance evaluation on mobile terminals is needed. Furthermore, to reduce the transaction time of the update process, we can execute an off-line batch update as a process that is independent from categorization.

Figure 7 denotes the data size of all signatures. Our update algorithm kept the size of all signatures below 3.0 Mbyte and the false rate was not increased by the update with signature compression. We also evaluated the accuracy of the categorization algorithm using real *Twitter* messages in a disaster, and confirmed that the results were similar.

6 Conclusion

In this paper, we discussed a disaster-information-gathering (DIG) system for use during large-scale disasters and proposed a categorization mechanism for the DIG system. By comparing data with signatures generated from old sent/received documents, the categorization algorithm determines whether data are important with a computation time of 5 msec and an accuracy more than 90 % on a smart phone. The update algorithm minimizes the size of signatures and the false rates of categorization.

This research is the first step towards efficient information collection in disasters and emergencies. In future research, we will evaluate the false rate to apply our algorithm to a real system and its transaction data.

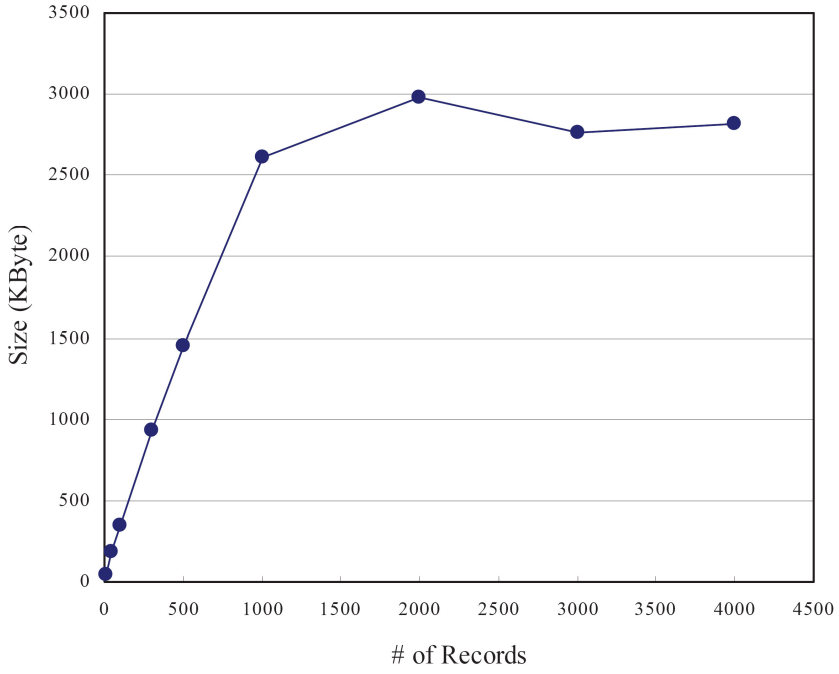

Figure 7: Size of Signature Database

Acknowledgment

This work has been supported by the Japanese Ministry of Internal Affairs and Communications funded project, “Study of Security Architecture for Cloud Computing in Disasters.”

References

[1] S. Kiyomoto, K. Fukushima, and Y. Miyake, “Security-and-privacy-related issues on it systems during disasters,” in *Proc. of the 2nd IFIP International Workshop on Security and Cognitive Informatics for Homeland Defense (SeCIHD’12), Prague, Czech, LNCS*, vol. 7465. Springer-Verlag, March 2012, pp. 445–459.

[2] T. Wood, E. Cecchet, K. K. Ramakrishnan, P. Shenoy, J. van der Merwe, and A. Venkataramani, “Disaster recovery as a cloud service: economic benefits & deployment challenges,” in *Proc. of the 2nd USENIX conference on Hot topics in cloud computing (HotCloud’10), Boston, USA*, June 2010.

[3] A. S. Atteih, S. A. Algahtani, and A. Nazmy, “Emergency management information system: Case study,” in *Proc. of the 4th International Conference on Cartography and GIS (GIS’12), Albena, Bulgaria*, June 2012, pp. 1–6.

[4] R. W. Perry, “Incident management systems in disaster management,” *Journal of Disaster Prevention and Management*, vol. 12, no. 5, pp. 405–412, 2003.

[5] X. Yao, M. Turoff, and R. Hiltz, “A field trial of a collaborative online scenario creation system for emergency management,” in *Proc. of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM’10), Seattle, Washington, D.C, USA*, May 2010.

[6] T. R. de Lanerolle, W. Anderson, S. DeFabbia-Kane, E. Fox-Epstein, D. Gochev, and R. Morelli, “Development of a virtual dashboard for event coordination between multiple groups,” in *Proc. of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM’10), Seattle, USA*, May 2010, pp. 35–35.

[7] G. Wickler, S. Potter, A. Tate, and J. Hansberger, "The virtual collaboration environment: New media for crisis response," in *Proc. of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM'11), Lisbon, Portugal*, May 2011.

[8] I. Shklovski, L. Palen, and J. Sutton, "Finding community through information and communication technology in disaster response," in *Proc. of the 2008 ACM conference on Computer Supported Cooperative Work (CSCW'08), San Diego, California, USA*. ACM, November 2008, pp. 127–136.

[9] Y.-N. Lien, H.-C. Jang, and T.-C. Tsai, "A MANET based emergency communication and information system for catastrophic natural disasters," in *Proc. of the 29th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS'09), Montreal, Quebec, Canada*. IEEE, June 2009, pp. 412–417.

[10] R. B. Dilmaghani and R. R. Rao, "A systematic approach to improve communication for emergency response," in *Proc. of the 42nd Hawaii International Conference on System Sciences (HICSS'09), Hawaii, USA*. IEEE, January 2009, pp. 1–8.

[11] J. T. B. Fajardo and C. M. Oppus, "A mobile disaster management system using the android technology," *International Journal of Communications*, vol. 3, no. 3, pp. 77–86, June 2009.

[12] Q.-A. Zeng, H. Wei, and V. Joshi, "An efficient communication system for disaster detection and coordinated emergency evacuation," in *Proc. of the 2008 Wireless Telecommunications Symposium (WTS'08), California, USA*. IEEE, April 2008, pp. 329–333.

[13] H. Stern, "A survey of modern spam tools," in *Proc. of the 5th Conference on Email and Anti-Spam (CEAS'08), California, USA*, August 2008.

[14] E. Kirda and C. Krügel, "Protecting users against phishing attack," *Computer Journal*, vol. 49, no. 5, pp. 554–561, 2006.

[15] G. Robinson, "A statistical approach to the spam problem: Using bayesian statistics to detect an e-mail's spamminess," Linux Journal, March 2003, http://www.linuxjournal.com/article/6467.

[16] R. A. Maxion and T. N. Townsend, "Masquerade detection using truncated command lines," in *Proc. of the 2002 International Conference on Dependable Systems & Networks (DSN'02), Bethesda, Maryland*. IEEE, June 2002, pp. 219–228.

[17] K. Julisch and M. Dacier, "Mining intrusion detection alarms for actionable knowledge," in *Proc. of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD'02), Alberta, Canada*. ACM, July 2002, pp. 366–375.

[18] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Transactions on Information and System Security (TISSEC)*, vol. 3, no. 4, pp. 227–261, 2000.

[19] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275–300, May 2004.

[20] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Proc. of the 15th Workshop on Learning for Text Categorization (AAAI'98), Madison, Wisconsin, USA*, July 1998, pp. 41–48.

[21] C. D. Manning and H. Schutze, *Foundations of statistical natural language processing*. The MIT Press, 1999.

[22] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, July 1945.

[23] P. Jaccard, "Etude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 37, no. 1, pp. 547–579, 1901.

[24] B. Eisenberg, "On the expectation of the maximum of IID geometric random variables," *Statistics & Probability Letters*, vol. 78, no. 2, pp. 135–143, February 2008.

[25] I. Twitter, "Twitter, Inc." https://twitter.com/.

Shinsaku Kiyomoto received his B.E. in Engineering Sciences and his M.E. in Materials Science from Tsukuba University, Japan, in 1998 and 2000, respectively. He joined KDD (now KDDI) and has been engaged in research on stream ciphers, cryptographic protocols, and mobile security. He is currently a senior researcher at the Information Security Lab. of KDDI R&D Laboratories Inc. He was a visiting researcher of the Information Security Group, Royal Holloway University of London from 2008 to 2009. He received his Doctorate in Engineering from Kyushu University in 2006. He received the IEICE Young Engineer Award in 2004. He is a member of JPS.

Kazuhide Fukushima received his M.E. in Information Engineering from Kyushu University, Japan, in 2004. He joined KDDI and has been engaged in the research on digital rights management technologies, including software obfuscation and key-management schemes. He is currently a researcher at the Information Security Lab. of KDDI R&D Laboratories Inc. He received his Doctorate in Engineering from Kyushu University in 2009. He received the IEICE Young Engineer Award in 2012. He is a member of Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, and ACM.

Yutaka Miyake received the B.E. and M.E. degrees of Electrical Engineering from Keio University, Japan, in 1988 and 1990, respectively. He joined KDD (now KDDI) in 1990, and has been engaged in the research on high-speed communication protocol and secure communication system. He received the Dr. degree in engineering from the University of Electro-Communications, Japan, in 2009. He is currently a senior manager of Information Security Laboratory in KDDI R&D Laboratories Inc. He received IPSJ Convention Award in 1995 and the Meritorious Award on Radio of ARIB in 2003.