Privacy Preferences for Geo-Calendar Based SMS Using Intelligent Text Configurator

K. B. Priya Iyer and V. Shanthi

***Abstract*—The mobile technology revolutionizes the world of communications opening up new possibilities for mobile commerce through Location based Advertising. Disclosing user location or personal information is major privacy issue in these applications. In this paper, we introduce a new User Privacy Preferences framework for Geo-Calendar based SMS. SMS block is based on Textese Vocabulary, Current Location, and Schedule. The algorithm consists of four phases namely Rule Analyzer, Intelligent Text Miner, Log Filer, SMS dispatcher. Simulated experiments were conducted by collecting large number of training data sets for text classification. The results prove the prediction accuracy of text classifier and also provide high location privacy for user based on both content and location based attacks.**

***Index Terms*—Location based SMS, privacy preferences, SMS classification, content based mining, location based advertising, GIS, and location based services, GPS.**

I. Introduction

With the advancements in mobile technologies and positioning systems, the users interest in searching for location based services (public servers for retrieving point of interest) are increased to a great extent. In accessing this information, the users are leaking their location or personal information to the public servers. Marketers also use the mobile way of advertising for marketing their products which opens a new way of mobile commerce called Location based Advertising (LBA).

Fig. 1. Location based POI access – user privacy

Manuscript received July 23, 2012; revised September 17, 2012.

K. B. P. Iyer is with the Research Scholar, Sathyabama University. Associate Professor, M.O.P Vaishnav College for Women (Autonomous), Chennai, India (e-mail: priya_balu_2002@yahoo.co.in).

V. Shanthi is with the St. Joseph College of Engineering, Chennai, India (e-mail: drvshanthi@yahoo.co.in).

Fig. 2. SMS spam, lba – user privacy

Many privacy enhancing techniques are available for hiding user personal information. Still users feel their privacy is in jeopardize because of the SMS spam, Location based advertising and unwanted SMS from known and unknown members. The Figure 1 and Figure 2 represents two examples situations where user privacy policy has to be improved in order to safe guard the user sensitive information.

To sum up we make the following contributions:

1) Developed an algorithm for blocking SMS based on keyword, date-wise, time-wise and location-wise. The algorithm consists of four phases namely Rule Analyzer, Intelligent Text Miner, Log Filer and SMS dispatcher. Textese content based filtering of SMS is done using both keyword and Textese vocabulary by Intelligent Text Miner (ITM).
2) Privacy preferences can be set based on the date, time and locations for types of the SMS messages with reactive words, Logograms, Emoticons, Regional variants etc. The blocked SMS can be retrieved later using Log Filer. The SMS can be retrieved based on user accessing such as date-wise, location-wise, keyword-wise etc.
3) SMS is classified using Intelligent Text Miner (ITM). The Phases in ITM are Parser, Tokenizer, Text Classifier, Relevance Scorer, and Accuracy Predictor.
4) The SMS dispatcher calculates the relevancy score of the Overall SMS content and decides whether to Blocks/Unblocks the message.

The reminder of this paper is organized as follows. In Section 2, we review the related work on both privacy related studies and text classification techniques. In Section 3, we formally define System model for setting up User privacy preference. In Section 4, we introduce the algorithm for Location based SMS blocking through four phases namely

Rule Analyzer, Intelligent Text Miner, Log Filer and SMS dispatcher. Section 5 presents the results of our experimental evaluation of our proposed model using training data sets. Finally section 6 concludes the paper with future research.

II. Related Work

A. Location Based Privacy and User Privacy Preferences /Policies

Location based applications [6], [7], [8], [9], [10], [11] are gaining popularity with millions of users accessing goe-spatial points from anywhere and at anytime. Location-based services [12] offer users to connect with friends, to discover their environment or to optimize their mobility. In most services, users share their location episodically when connecting to the service. Some services such as road-traffic monitoring systems require users to continuously share their location. Golle and Partridge [4] identify location-based quasi-identifiers by showing that home and work locations uniquely identify most of the US population. In [3], personalization privacy technologies were introduced.

B. SMS based Text Classification Studies

In [5], a new algorithm for text classification is introduced where instead of using words, word relation i.e. association rules is used to derive feature set from pre-classified text documents. The concept of Na ïve Bayes Classifier is then used on derived features and finally a concept of Genetic Algorithm has been added for final classification. In [2], mobile text classification is made up of six main parts; SMS collection, preprocessing, feature selection, term weighting, re-parameterization using PCA and Neural Network classification. In [13], an anti-spam technique based on Artificial Immune System (AIS) for filtering SMS spam messages. The proposed technique utilizes a set of some features that can be used as inputs to spam detection model. The idea is to classify message using trained dataset that contains Phone Numbers, Spam Words, and Detectors. The [1], explores the potential application of sentiment mining for analyzing short message service (SMS) texts in teaching evaluation.

III. Algorithm

The algorithm consists of four phases namely Rule Analyzer, Intelligent Text Miner, Log Filer, SMS dispatcher.

A. Phase I: Rule Analyzer

The incoming SMS is filter based on the priority preference rule set defined by the Rule Analyser. The incoming SMS is blocked based on the following Rule Set defined by Rule Analyser:

The BlockListDB maintains the list of sender numbers to be blocked, sender names to be blocked. The Location based Content block list and Content block list is maintained by the LocBlockListDB. The incoming SMS is first checked with the BlockListDB. If the sender number/ sender name matches with the list then the SMS is blocked without knowing to the user and it is forwarded to the Logfiler. If the user wishes, he can access the blocked SMS at a later time or he can set automatic deletion of block messages.

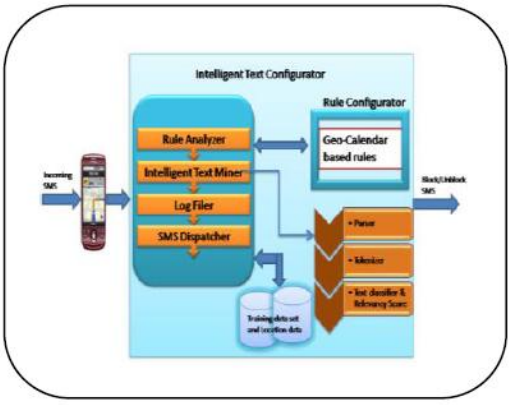

Fig. 3. Block diagram of geo-calendar based sms

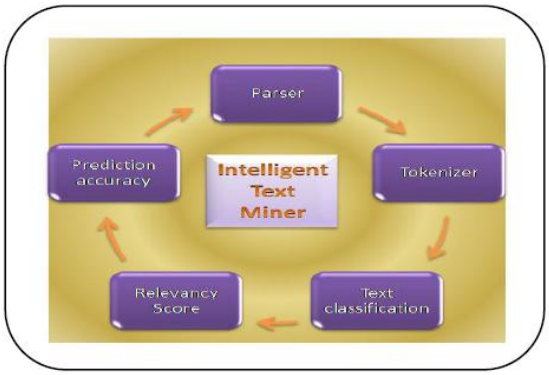

Fig. 4. Intelligent text confgurator stages

B. Phase II: Intelligent Text Miner (ITM)

During this phase, if the incoming SMS is not in BlockListDB then it is forwarded to ITM. Here the Geo-Calendar based Rule List and Content based Rule List is considered for processing. The rules defined by the user in the LocBlockListDB are taken by the Intelligent Text Miner for processing. Now the SMS text is classified basing on the following phases.

1) *Parser:* The Message is first scanned for spams. The entire SMS is read and the text is ready for classification.
2) *Tokenizer*: The SMS is divided into components or tokens. These tokens are separated as Stop words, Stemming, Acronyms, Logograms, Regional variants etc.
3) *Text Classification*: The text is classified by giving weight to each word.
4) *Relevancy score:* The score of each text is measured.
5) *Prediction Accuracy*: After text is classified, the accuracy of classification is predicted.

Phase 4.3: Log Filer

The ITM calculates the relevancy of content and messages are decided whether to block or not. If messages are blocked then it is forwarded to the LogFiler. The Log Filer maintains the list of blocked messages. The blocked messages can accessed the user at any time. The list can be accessed Location-wise, Number-wise, Name-wise or Content-wise.

Phase 4.4: SMS dispatcher

The SMS despatcher sends the related SMS to inbox of the user.

C. Algorithm

Algorithm 1 : Privacy Preferences

PPSMS (uloc,msg,sno,BlockListDB,LocBlockListDB)

/* uloc: user location (latitude,longitude) , msg: SMSmessage, sno:sender number, BlockListDB: block list of number or names, LocBlockListDB: location based content block list and content block list. */

1) Retrieve the sender phone number and check for blocking
   No_bstatus ← CHK_NUMBLOCK(sno,BlockListDB)
2) If (No_bstatus = true) then
3) LF=STORE_LOGFILE(sno,msg) /* block the message and store to LogFiler */
4) exit
5) End if
6) Name_bstatus ← CHK_NAMEBLOCK(sname,BlockListDB) /* retrieve the sender name and check with list */
7) If (Name_bstatus = true) then
8) LF=STORE_LOGFILE(sname,msg) /* block the message and store to LogFiler */
9) exit
10) End if
11) LocRule ←RET_LOCRULE(uloc,LocBlockListDB) /* retrieves the list of content block for user current location and Locstatus */
12) If (Locstatus= true) then
13) Relevance_score← ITM(LocRule,msg)
14) Else
15) Relevance_score← ITM(msg)
16) End if
17) If (Relevance_score) then
18) LF=STORE_LOGFILE(msg) /* block the message and store to LogFiler */
19) Exit
20) End if

IV. Experimental Evaluation

The dataset is collected from Department of Computer Science, National University of Singapore [32]. A collection of about 10117 SMS messages collected for research from NUS SMS Corpus (NSC).

A. Preprocessing

The first step of text mining process is text preprocessing in which the SMS Messages is analyzed syntactically or semantically. The text message is considered as a bag of words because the words and its occurrences are used to represent the SMS (Table II, Table III).

The algorithms applied at this stage are stemming and stop word removal, number removal and strip whitespaces. Tokenization is the task of slicing it up into pieces, called tokens and at the same time throwing away certain characters, such as punctuation. Stop words are words that do not add meaningful content to the SMS. Stemming reduces the frequency of unique words. Term frequency is the total occurrence of a word in a particular SMS (Table I).

TABLE I: Training Data Set

Words	Words Frequency in entire dataset	Words	Words Frequency in entire dataset
I	57138	Library	12
For	1042	Movie	83
So	22234	Restauran	13
How	10605	School	82
You	1370	Canteen	33
But	425	Picture	4
Just	190	Trip	31
On	14045	Matches	30
Me	24665	Email	65
U	56047	Shopping	45

B. Data Collection Statistics

TABLE II: Sample Dictionary

Stop words	Stemming Words	Texese Words	Categories chosen for Test
Sample : About, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, was, wasn't, we, we'd, we'll, we're, we've,	*Sample* : Division – divide, dividing, divided, divisor Fish – fishing, fisher Travel – travelling, traveler Apply – applying, applies, applied	*Sample :* Love- LUV Great – gr8 At - @ Before – b4 Thanks – thx Today - 2day	Education Movie Health Travel Sport Home Others

TABLE III: Sample SMS

1. I'm in my room... Later when you are done you call me. I meet you at central library?
2. I at home now.. Later after shopping den i go fetch u all lor. And i think u all will shop until quite late rite.
3. the lecturer like havent finish teaching lor how to do.
4. have booked table at angel restaurant at faculty club. Meet at twelve noon at outside the restaurant
5. Esplanade... An italian restaurant... Just finished e cake.. Later wanna come fetch me?

TABLE IV: User – Location based Preferences

Location Name (user significance of place)	Category to be blocked	Date-wise	Time-wise
OOty (when user at Hill station)	Education	Jun 1st 2012	6.00 A.M. TO 9.00 A.M.
Adyar, Chennai (when user at College)	Travel, Movie	Jun 15th 2012	10.00 A.M. TO4.00 P.M.
Mylapore, Chennai (when user at Theater)	Health, Home	Jun 16th 2012	7.00 P.M. TO 10.00 P.M.
Chetpet, Chennai (when user at Sports ground)	Education, Home	Jun 18th 2012	11.00 A.M. TO 3.00 P.M.

TABLE V: CLASSIFICATION EVALUATION FORMULA DEFINITION

Measure	Formula	Intuitive Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct.

TABLE VI: RESULTS OF CLASSIFICATION FOR GEO-CALENDAR BASED SMS BLOCK

Preference List	Block category	TP	FP	FN	TN	P	R	A
OOty (Hill station)	Education	75	5	1	1	93%	98%	92%
Adyar, Chennai (College)	Travel, Movie	110	15	5	25	88%	95%	87%
Mylapore, Chennai (Theater)	Health, Home	102	3	19	5	97%	84%	82%
Chetpet, Chennai (Sports	Education, Home	99	25	15	3	79%	86%	71%

To analyze the detection capability of the technique, we employ evaluation measures that include True Positive,

True negative, False Positive, False Negative. The results are shown in table VI.

V. CONCLUSION

In this paper, we propose a User Privacy Policy based on Geo-Calendar based SMS block using Textese vocabulary. SMS classification is based on tokens such as Acronyms, Logograms, and regional variants. The goe-calendar based privacy preferences are maintained by Rule Configurator. The user can view the blocked messages at later time by visiting the LogFiler. Text classification algorithms are used and predictor accuracy is calculated. The experimental results showed high privacy for content based block. Additional future works includes SMS block for partial numbers, day significance and seasonal variations.

REFERENCES

[1] C. K. Leong, Y. H. Lee, and W. K. Mak, "Mining sentiments in SMS texts for teaching evaluation," *Expert Systems with Applications* vol. 39, pp. 2584–2589, 2012.

[2] D. Patel and M. Bhatnagar, "Mobile SMS classification an application of text classification," *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307 (Online), vol. I, Issue-I, March 2011.

[3] E. Toch, Y. Wang, and L. F. Cranor, "Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems," *User Model User-Adap Inter*, vol. 22, pp. 203–220, 2012.

[4] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," *In Pervasive*, 2009.

[5] S. M. Kamruzzaman and F. Haider, "A hybrid learning algorithm for text classification," *3rd International Conference on Electrical and Computer Engineering ICECE 2004*, 28-30 December 2004, Dhaka, Bangladesh.

[6] K. B. Priya Iyer and V. Shanthi, "Goal directed relative skyline queries in time dependent road networks," *IJDMS*, vol.4, no.2, April 2012.

[7] K. B. Priya Iyer and V. Shanthi, "Geo-calendar based Predictive Skyline Queries using Fuzzy Inference Engine," *IJCA*, vol. 45, Number 15, May 2012.

[8] K. B. Priya Iyer and V. Shanthi, "Spatial Boolean Skyline Boundary Queries in Road Networks," *Proceedings of ICCCNT*, July 2012, India.

[9] K. B. Priya Iyer and V. Shanthi, Location based emergency response management system through satellites in road networks," *Proceedings of EEC* 2012, July 2012.

[10] K. B. Priya Iyer and V. Shanthi, "Location based disaster management system through smart phones," *Proceedings of ACCT*, July 2012, India.

[11] K. B. Priya Iyer and V. Shanthi, "Intelligent path finder for goal directed queries in road networks," *Proceedings of MNC Apps, AIEEE*, August 2012.

[12] T. Hashem and L. Kulik, "Don't trust anyone: Privacy protection for location-based services," *Pervasive and Mobile Computing* vol. 7, pp. 44–59, 2011.

[13] T. M. Mahmoud and A. M. Mahfouz, "SMS spam filtering technique based on artificial immune system," *IJCSI International Journal of Computer Science Issues*, vol. 9, Issue 2, no 1, March 2012.

[14] T. Xu and Y. Cai, "Location anonymity in continuous location-based services," in *Proceedings of the 15th Annual ACM international Symposium on Advances in Geographic information Systems*, pp. 1–8, New York, NY, USA, 2007.

[15] G. Zhong and U. Hengartner, "A distributed k-anonymity protocol for location privacy," in *Proceedings of Seventh IEEE International Conference on Pervasive Computing and Communication (PerCom 2009)*. Galveston, TX, pp. 253–262, 2009.

[16] G. M. Blog, Finding places "near me now" is easier and faster than ever, 2010. [Online]. Available: http://googlemobile.blogspot.com/2010/01/finding-places-near-me-now-is-easier.html.

[17] Collecting SMS Messages for a Public Research Corpus. [Online]. Available: http://wing.comp.nus.edu.sg:8080/SMSCorpus/