Multilingual Rules for Spam Detection

Minh Tuan Vu[1], Quang Anh Tran[1], Frank Jiang[2] and Van Quan Tran[1]

[1]*Faculty of Information Technology, Hanoi University, Hanoi, Vietnam*
[2]*School of Engineering and IT, University of New South Wales, Canberra, Australia*

Received 15 April 2013; Accepted 23 May 2014
Publication 4 August 2014

Abstract

In this paper, we introduced a statistical rule-based method to create rules for SpamAssassin to detect spams in different languages. The theoretical framework of generating and maintaining multilingual rules were also illustrated. The experiments were conducted against the dataset of three languages including Chinese, Vietnamese and English. The detecting achievement of multilingual rule was 89.5% for the true detection and only 3.8% for the failed alarm at the threshold of 2 while the true detection rate of single language rule was not over 61% and the failed alarm rate was up to 4.9%.

Keywords: Spam detection, multilingual rules, SpamAssassin, spam, ham.

1 Introduction

In recent years, the battle against spam e-mail is extremely fierce. Despite the anti-spam technology development, spammers keep working hard to find new strategies which help deliver unwanted messages to email users all around the world. One of these tricks is sending spam emails in different languages beside users' vernaculars. According to Message Labs' July 2009 Intelligence Report [1], in France, Netherland and Germany, spammers used spam translation technique to generate spam at 53%, 25% and 46% respectively. In China and Japan, the rates of non-English spam were even up to 63.3% and 54.7%.

Journal of Machine to Machine Communications, Vol. 1 , 107–122.
doi: 10.13052/jmmc2246-137X.122

The trick has worked relatively well because spammers dig deep into the flaw of current spam-filtering machines detecting spam based on the wordlist which is not good at dealing with multilingual emails. The report [1] also explained that by making full use of auto-translation tools, spammers have created different language spam and causes a 13% rise in overall spam in mentioned countries above.

The development of automated translation tools is natural and necessary. In order to solve this problem, the multilingual rules for spam-filtering machines should be proposed. In a recent paper, Quang-Anh Tran et al. [2] introduced a method to create Chinese rules for SpamAssassin. Although this set of rules has done a good job and been shared by thousands of email servers all around the world, it could detect the spam email in Chinese only. To surmount this ruse, we level up the method and make it multilingual. In other words, we created a system that could generate the set of anti-spam rules for different languages. The experiments were conducted with the same dataset for every language (Chinese, Vietnamese and English) and mixed type of these three ones.

The paper is structured as follows: In section II, we reviewed some approaches to filer spams in specific languages and related knowledge. Section III follows with the theoretical framework of our method. Next, the experiments are conducted and the results are compared in section IV. Finally, section V concludes the paper and discusses the future of work.

2 Related Works

2.1 SpamAssassin Rules

SpamAssassin is one of the most popular for deciding how likely an email message is spam. It filters spam based on content-matching rules. Each rule has its own score. If an email message gains enough scores (over the pre-defined threshold), it will be marked as spam.

Here is the sample of a SpamAssassin rule:

Figure 1 is an example of complete rule definition. The rule named FROM_START_WITH_NUMS checks to see if an email's FROM header starts with at least two numbers against the regular expression. It adds a score to the email's spam score if the email matches the rule. An anatomy of a rule was described in details by Schwartz (2004) [15]. In order to catch the spam effectively in specific languages, the rules should be generated based on the characteristic of those languages. That is the reason why we are aiming to build a multilingual rule set for an international environment.

```
header FROM_STARTS_WITH_NUMS    From =~ /^\d\d/
describe FROM_STARTS_WITH_NUMS  From: starts with nums

score FROM_STARTS_WITH_NUMS     0.390 1.574 1.044 0.579
```

Figure 1 SpamAssassin rule sample

2.2 Researches on SpamAssassin Rules for Specific Language

Quang-Anh Tran and his partners [2] explained in their paper that spam detections fall into two categories: rule-based and statistical-based. The first one refers to the detection performed by searching for the spam-liked pattern in the email. SpamAssissin is known as the most popular representative of ruled-based spam detection machines. The latter, on the other hand, manages to deal with a two-class categorization problem; the dataset of spam and ham is used to train the detector. Bayesian algorithms are most widely used statistical-based method for detecting spam. Androutsopoulos (2000) [4] and Graham (2002) [5] had typical works on this subject. Besides, other statistical-based methods are proposed such as Neural Network [6], Support Vector Machines [7] for spam detection.

However, each method (rule-based or statistical-based) has its disadvantages. The rule-based method is easy to share among servers (or users) but it is built manually. Thus, it is difficult to keep up-to-date with the quick changes of spam. Whereas, with the statistical-based method, it is easier to retrain the spam detector as long as the training dataset is up-to-date. However, it is impossible to share the knowledge of the detector. Therefore, they propose a hybrid method which is a trade-off rule-based and statistical-based to create the rules for detecting spam in Chinese. This method has all advantages of rule-based and statistical-based method: the quick-training for the detector and easy to share between servers.

Nguyen T.A et al. [8] showed an approach to detect Vietnamese spam based on language classification. They aimed to introduce a Vietnamese segmentation for using token selection for building a Vietnamese spam filter based on language classification and Bayesian combination to sufficiently support Vietnamese. The results on spam detection between their Vietnamese segmentation and space token segmentation were compared. Their spam detection rate is about 9% higher and the ham error rate is 3% lower.

Although both methods proposed in [2] and [8] achieve positive results, they only focused on a specific language. The question, here, is how these methods deal with the real circumstance that users receive emails in more than

one languages every day and spammers keep sending multilingual spams to email users around the world.

3 Theoretical Framework

The Figure 2 illustrates how our multilingual rules are generated and maintained.

The email from different sources are classified and saved into the Spam & Ham database. The classification is carried out by email users and researchers. Because this is a kind of content-based approach, all we need of an email are the subject and body. After decoding the encoded content and strip the entire html tags attached with the email, we use Google API to detect the language of each email. For each language (only three languages including Chinese, Vietnamese and English are used in this paper), a suitable segmentation method will be called. The product of this step is a meaningful wordlist which are the output for next step. We reuse the algorithms in [2] for the rest of process which are discussed in the next part.

The multilingual rules set are generated automatically by three steps: Pattern retrieval, Pattern Selection and Score Assignment.

3.1 Pattern Retrieval

As we mentioned above, each language has its own way to split sentences into meaningful words. For some languages such as English, French or Germany, words can be identified easily by the space. However, with Vietnamese,

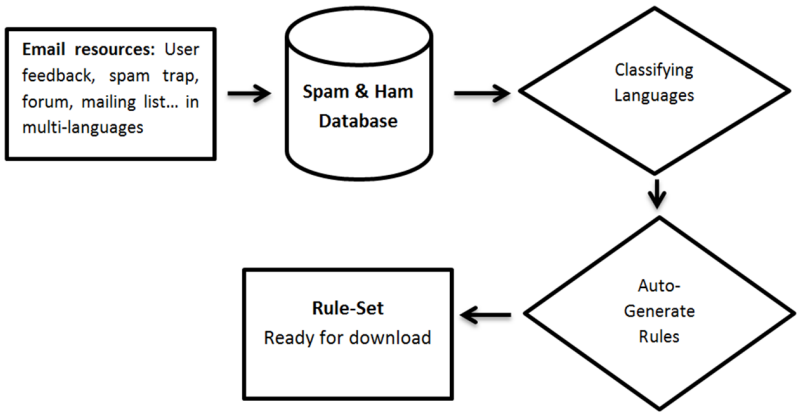

Figure 2 Process of generating multilingual rules

Chinese or some other Asian languages, it is impossible because they have special linguistic united known as syllable ("Tiếng" in Vietnamese or "hanzi" in Chinese).

In order to achieve the highest effectiveness when processing the email content, we used Google Translate API [9] to detect the language of the email. Although the API works well and is easy to use, it is not free. Then, we only used for the experimental period. For further usage, we consider some other solutions such as Lingua::Identify available at [10] and Guess-Language [11]. In this paper, we only implemented the segmentation for three languages: Chinese, Vietnamese and English. However, we are working hard to propose a multilingual word segmentation method as Guo-Wei Lee mentions in his research [12].

With Chinese emails, we applied exactly the Chinese segmentation technique used in [2] which is based on methods: Dictionary-based, Maximum Matching; and from left to right.

With Vietnamese emails, we dealed with the word segmentation by a program proposed by Phuong Le-Hong [13]. This program works on the Vietnamese text file or folder and exports the meaningful wordlist to the XML format. It is quite straightforward to read the wordlist from this XML file.

It is much simpler to split words in English emails because words are separated by spaces. We just found and replaced the space character with the new line character and eliminated all punctuations in the sentence.

3.2 Pattern Selection

After classifying the email by languages and extracting the meaningful words, we applied some pattern selection methods to select good patterns for subject rules and body rules, individually. We could not find any difference among selecting pattern of Chinese, Vietnamese and English emails; thus, we once again reused the pattern selection algorithms in [2].

In spite of being based on the traditional pattern selection method by Yang [14], there are some changes in the approach. Only the spam-liked patterns were used to detect spam. As a result, the formula for selecting pattern was modified. The V_{ts} and V_{th} are computed as follows according to Conditional Probabilities and Bayes's Theorem:

$$V_{ts} = P(E \mid H) = \frac{P(E \wedge H)}{P(H)} \tag{1}$$

$$V_{th} = P\left(\overline{E} \mid H\right) = \frac{P\left(\overline{E} \wedge H\right)}{P\left(\overline{H}\right)} \tag{2}$$

In which:

- V_{ts} and V_{th} can best evaluate the connection between pattern *t* and spam, pattern *t* and ham, namely.
- Top N pattern that have highest value of ratio Rt = V_{ts}/V_{th} are chosen.
- N is the size of the rule set, which is a factor that control the performance of the rule set.
- E is a hypothesis that a message occurs as spam.
- H is a hypothesis that a message occurs as ham.

Given a spam and ham datasets, for a pattern t, A and B are the number of times that spam and ham messages contain *t*, respectively; C and D are the numbers of times spam and ham messages do not contain *t*, respectively. The values of the probabilities in (1) and (2) are computed as follows.

$$P\left(E\right) = \frac{A+C}{A+B+C+D} \tag{3}$$

$$P\left(\overline{E}\right) = \frac{B+D}{A+B+C+D} \tag{4}$$

$$P\left(H\right) = \frac{A+B}{A+B+C+D} \tag{5}$$

$$P\left(E \wedge H\right) = \frac{A}{A+B+C+D} \tag{6}$$

$$P\left(\overline{E} \wedge H\right) = \frac{B}{A+B+C+D} \tag{7}$$

3.3 Score Assignment

The rules are created on the basis of the selected set of spam-liked patterns. There are two types of rules: Body rule and Subject rule. The Fast SpamAssassin Score Learning Tool by Henry Stern [15] is used to assign the score to each rule.

According to the illustration of Quang-Anh Tran et al [2], The "Stochastic Gradient Descent" method of training a neutral network was implemented. The program uses a single perceptron and (8) a logsig activation function (9) to map the weights to SpamAssassin score space.

$$f\left(x\right) = \int_{i=1}^{N} w_i x_i \tag{8}$$

$$y(x) = \frac{1}{1 + e^{-f(x)}} \tag{9}$$

where w_i represents the score for rule i and x_i describes whether a given message activates rule i or not, the transfer function (8) returns the message's score. The gradient descent is employed to train the neural network. The parameter of the network is tuned iteratively to ensure that the rate of mean error always decreases. Without getting into calculus, the error gradient for a perceptron with a linear transfer function, logsig activation function and mean squared error function is as follows:

$$E(x) = y(x)(1 - y(x))(y_{exp} - y(x)) \tag{10}$$

And the weights are updated using the function:

$$w_i = w_i + \alpha E(x) x_i \tag{11}$$

In which, α is a learning rate. The implementation uses the so-called "Stochastic gradient descent" method which does incremental updates by walking through the training set randomly rather than doing one batch update per epoch because the SpamAssassin rule hits are spares.

4 Experiments

4.1 Dataset

The data we used to conduct the experiments is divided into 4 groups.

E-mails come from email users' personal inboxes and are classified as spam and ham manually by authors. We store all emails in the MySQL database. The spam and ham in each language are saved in separated table with the same structure (ID (PK), Subject, Body, Status, Date), then, there are eight tables serving the experiments.

Firstly, the experiments are conducted with three first groups (Group 1, 2 and 3) to create the rule for corresponding language. The rule set is tested based on single language dataset only. The results are saved for the

Table 1 Dataset description

Group	Num. of Spams	Num. of Hams	Language
1	200	200	Chinese
2	231	251	Vietnamese
3	274	202	English
4	705	653	Multi-languages

comparison (1). Next, the single language rule sets are tested based on data group 4 (multi-language emails) to evaluate the efficiency (2). Finally, the mixed languages rule set is generated based on data group 4. The effectiveness of this rule set is recorded and compared with the result (1) and (2).

4.2 Single-language Rule Set Creation

The procedures of generating rule set for specific are applied exactly mentioned in section 3.

For Chinese rules, the experiment is based on the data group 1 with 200 hams and 200 spams. The spam detection rate (Spam Recall) and the failed alarm rate (Ham Error) are illustrated in Table 2.

The Chinese rule gives the best result with the threshold equal to 2.5 at which the positive true rate is 91.5% and the failed alarm rate is eliminated.

The experiment conducted based on the Vietnamese (generating rules and testing rules totally with Vietnamese dataset – group 2) also brings positive results.

Table 2 Performance of Chinese rule with Chinese dataset

Threshold	Spam Recall	Ham Error
0.5	93.5%	30.5%
1	91.5%	9.0%
1.5	91.5%	5.5%
2	91.5%	4.0%
2.5	91.5%	0.0%
3	91.5%	0.0%
3.5	85.0%	0.0%
4	75.5%	0.0%
4.5	71.0%	0.0%

Table 3 Performance of Vietnamese rule with Vietnamese dataset

Threshold	Spam Recall	Ham Error
0.5	90.5%	34.7%
1	87.4%	27.9%
1.5	83.1%	11.2%
2	81.4%	2.4%
2.5	81.4%	0.0%
3	78.4%	0.0%
3.5	73.6%	0.0%
4	66.2%	0.0%
4.5	59.3%	0.0%

Table 4 Performance of English rule with English dataset

Threshold	Spam Recall	Ham Error
0.5	98.5%	81.2%
1	97.1%	50.5%
1.5	96.0%	24.3%
2	95.6%	5.0%
2.5	95.3%	0.0%
3	93.1%	0.0%
3.5	87.6%	0.0%
4	82.8%	0.0%
4.5	60.2%	0.0%

At threshold 0.5, the spam recall rate is really high (90.5%) but the ham error rate is unacceptable (up to 34.7%). However, when we increase the threshold, the result is better and better. Especially, the ham error rate falls significantly at the threshold 2 (2.4%) while the spam recall stay unchanged in comparing to the previous threshold (81.4%).

We did the same thing to generate the English rule set and then recorded the result after testing the rule based on English emails only. The results are displayed in the Table 4.

The English rule set works extremely effectively at the threshold of 2.5. At this point, the positive true rate stays high over 95% while the ham error is totally eliminated.

On finishing the experiment to generate the SpamAssassin rule and to test these rule set with the corresponding language, we gained really positive results on true spam detection rate and failed alarm rate. However, whether these rule sets still work well with multi-language dataset? The answer is coming with the next experiment.

4.3 Single-language Rule Set Tested with Multi-language Emails

Three sets of rule in Chinese, Vietnamese and English are tested with the data group 4 which contains 705 spams and 653 hams in multi-languages in order to evaluate the efficiency in a multilingual environment. Table 5 shows the result of how Chinese, Vietnamese and English rule sets works.

The statistic shows obviously that when working with the multilingual email dataset, all sets of rules give very poor performance in the true positive rate, especially, the Chinese rule which detects only 24.5% at threshold 0.5 in comparing to over 93% of the last experiment. English rule and Vietnamese are better at spam detecting but the result is far lower than those when working

Table 5 Performance of single language rule with multilingual dataset

Threshold	Chinese		Vietnamese		English	
	Spam Recall	Ham Error	Spam Recall	Ham Error	Spam Recall	Ham Error
0.5	24.5%	0.2%	60.6%	3.5%	51.9%	4.9%
1	21.8%	0.2%	59.0%	2.0%	49.9%	1.8%
1.5	20.7%	0.0%	54.8%	0.2%	44.3%	1.1%
2	19.0%	0.0%	53.6%	0.2%	42.4%	0.3%
2.5	18.2%	0.0%	49.4%	0.0%	41.6%	0.0%
3	16.9%	0.0%	46.8%	0.0%	40.7%	0.0%
3.5	16.7%	0.0%	38.4%	0.0%	39.9%	0.0%
4	16.0%	0.0%	20.7%	0.0%	32.5%	0.0%
4.5	15.6%	0.0%	14.8%	0.0%	22.1%	0.0%

with single language dataset. The explanation for the fall is quite clear and straightforward. The rule generated from specific language can detect the spam effectively in that language only. Therefore, we are expecting a multilingual rule set that can detect spam effectively among ton of multi-language emails.

4.4 Multilingual Rule Set

The final experiment is to generate and to test the rule set from the data group 4 which contains the email in three languages Chinese, Vietnamese and English.

After classifying the language of each email, we did the word segmentation for the set of emails in the same language. The pattern selection chooses the best pattern for evaluating the whether a pattern is spam-liked or not. Based on the selected pattern the rule set is generated automatically. The Fast SpamAssassin Score Learning Tool will handle the rest by assigning the score for each rule. Applying these steps on the multi-language dataset, we gained a set of multilingual rule for detecting the spam.

Table 6 illustrates the result of the test detecting spam based on the multilingual rule set. At the first level of the threshold, although the spam recall rate is highest (94.17%), the ham error rate is up to 48.80%. It is intolerant for a set of SpamAssassin rules. However, the result is much better when the threshold increase to 2.5. The true positive rate is 89.40% and the false positive rate is 0%. At this threshold, with the same multi-language dataset, the performance of Chinese rule, Vietnamese rule and English rule are 18.2%, 49.4% and 41.6% namely. This comparison proves that the rule generated from the multilingual dataset works much more effectively than the one generated based on the single language only.

Table 6 Performance of multilingual rule with multilingual dataset

Threshold	Spam Recall	Ham Error
0.5	94.17%	48.80%
1	92.00%	29.13%
1.5	90.20%	13.67%
2	89.50%	3.80%
2.5	89.40%	0.00%
3	87.67%	0.00%
3.5	82.07%	0.00%
4	74.83%	0.00%
4.5	63.50%	0.00%

5 Remarks

With a number of experiments carried out above, three sets of single-language rules are generated. The results in the Tables 3, 4 and 5 show that these sets of rules work effectively when dealing with the set of single-language dataset (Most of emails are in only one language). At the same threshold 2.5, the Chinese rules can detect up to 91.5% spam, the Vietnamese rules detect 81.4% spam and the result of English rules are 95.3% while the failed alarm is 0%.

However, when applying these sets of single-language rules in detecting multilingual spam, the results are not good at all. In detail, at the threshold 2.5, the percentage of spam detecting of Chinese, Vietnamese and English rules are 18.2%, 49.4% and 41.6%. The reason for this drop is clear. Each set of single-language rules is generated based on the corresponding language dataset. It means the rule can deal with spam in that language only. Therefore, a set of multilingual rule is considered and evaluated. This set of rules is built from the multilingual dataset including Chinese, Vietnamese and English.

An experiment is run to evaluate the performance of multilingual set of rules. The result is positive and promising. At threshold of 2.5, the rate of spam detecting is 89.40% while the ham error is eliminated. From these findings, it is shown that that effectiveness of a multilingual set of rule in detecting spams when applying in an international working environment.

6 Conclusion

Generating the rule for spam detection based on a specific language is a proper approach to fight against spammers. However, in order to deal with an email server receiving emails in more than one language, we need an extended solution. Therefore, we upgraded the method proposed in [2] to implement

the system that is able to generate automatically the multilingual rule based on the multi-language dataset. The experiment results show that these rules help SpamAssassin detect spam more exactly in comparison with the ones generated based on single language dataset.

Despite of the positive results achieved, there are some issues we need to deal with in the future. Firstly, a new method to detect the language of the email should be analysed. The cost for the current one is so high. Secondly, we are expecting a better algorithm to retrieve the pattern of raw emails. Finally, it would be a big problem for the word segmentation if the system faces up to a large number of languages due to the lack of a common segmentation method as mentioned in [12].

Acknowledgments

This research was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under project number 102.01-2010.09. This work received tremendous support from the Vice-chancellors' research initiatives from the University of New South Wales (UNSW).

References

[1] Multi-language Spam-A New Trend Among Spammers. Available: http://www.spamfighter.com/News-12908-Multi-language-Spam-A-New-Trend-Among-Spammers.htm

[2] Tran, Q. A., Duan, H. X. Li, X., 'Real-time statistical rules for spam detection' IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.2B, pp 178–184, February 2006.

[3] Androutsopoulos I., Koutsias, J., Chandrinos, K.V., Paliouras, G., Spyropoulos, C.D., 'An evaluation of Naive Bayesian anti-Spam filtering', Proceedings of the Workshop on Machine Learning in the New Information Age, pp 9–17, 11th European Conference on Machine Learning, Barcelona, Spain, 2000.

[4] Graham, P., 'A plan for spam. Web document' (2002). Available: http://www.paulgraham.com/spam.html.

[5] Drucker, H., Wu, D., Vapnik V., 'Support Vector Machines for spam categorization', IEEE Transaction on Neural Networks.10(5), 1048–1054, 1999.

[6] Özgür, L., Güngör, T., Gürgen, F., 'Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish', Pattern Recognition Letters, 1819–1831 25. 2004.
[7] Nguyen T.A., Tran Q.A., Nguyen N.B., 'Vietnamese spam detection based on language classification', HUT-ICCE 2008 - 2nd International Conference on Communications and Electronics, Hoi An, Vietnam, 2008.
[8] Google Translate API. Available: https://developers.google.com/translate/
[9] Lingua::Identify. Available: http://search.cpan.org/ambs/Lingua-Identify-0.51/lib/Lingua/Identify.pm
[10] Guess-Language. Available at http://code.google.com/p/guess-language/
[11] Guo-Wei Lee, 'A Mechanism for Filtering Multilingual Spam Mail based on Decision Tree and Integrated Feature Selection Algorithm', 1997.
[12] Phuong Le-Hong, et al, 'A hybrid approach to word segmentation of Vietnamese texts', Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, LATA , Springer LNCS 5196, Tarragona, Spain, 2008.
[13] Yang, Y., Pedersen, J.O., 'A comparative study on feature selection in text categorization', Proceedings of the 14th International Conference on Machine Learning, pp 412–420, 1997.
[14] The Fast SpamAssassin Score Learning Tool. Available: http://spamassassin.apache.org
[15] Schwartz, 'SpamAssassin', 2004, O'Reilly.

Biographies

Vu Minh Tuan is a lecturer and a research coordinator of the Faculty of Information Technology at Hanoi University. Currently, he also is working as the project manager at the software department of IP Communications, JSC.

In 2012 he started his Master of Science from University of Central Lancashire, UK. He was a student of Hanoi University from 2006 to 2010. He has been working on Vietnamese rules for SpamAssassin and some projects in the field of AntiSpam, email system and data mining.

Tran Quang Anh is an associate professor of information technology from Hanoi University. He obtained Ph.D and Master degrees at Tsinghua University in 2003 and 2001 respectively. He finished his bachelor at Huazhong University of Science and Technology in 1997. His research interests include network security evolutionary algorithms and field-programmable gates array.

Dr. Jiang received his B.Sc. degree in System & Control Engineering and M.Sc. degree (by research) in Computer Science Engineering on 1997 and 1999 respectively in China and Australia. With a success of holding an Australian Postgraduate Award (APA) scholarship, he completed his PhD degree in communication engineering and software engineering at University of Technology, Sydney (UTS) in 2008. Prior to joining into UTS, Dr. Jiang

has 5 years' hardware and software working experience in the VoIP industry in Sydney, Australia from 1999 to 2003. After his PhD, he was employed as a research assistant and later a research fellow for overall 3 years in Faculty of Engineering and IT (FEIT), UTS. Additionally, he has 5 years of teaching experiences as a lecturer and 2 years of subject coordinator in UTS. Currently, he works in the University of New South Wales (UNSW) as a lecturer and a full-time UNSW Vice-Chancellor appointed Research Fellow. He has published over 60 international journal and conference papers in the field of computational intelligence and its applications. His current research interests include data analytics, bio-inspired algorithms and metaheursitics, Underwater communication, Network Security, Autonomic communication networks, Intelligent and mobile agents, Network Protocols.

Van Quan, Tran Quan graduated from Hanoi University (HANU) on 2012 with the B.Sc. degree in Information technology and currently (2014) is an Information System Design master student at University of Central Lancashire, UK. He has teaching experience as a teacher assistant at HANU and more than 2 years' experience working as a developer. His current research interests include: text mining, voice recognition and human-computer interaction.