World Applied Sciences Journal 10 (8): 879-885, 2010
ISSN 1818-4952

A New Statistical Approach to Detect Invalid Email Address

Morteza Zahedi and Mohsen Ameri

Computer Engineering and IT Department, Shahrood University of Technology, Shahrood, Iran

Abstract: The validity of user's email addresses is important for all websites and online portals in order to recognize the users being a human and also keeping contact with the users who are the costumers of an online shop or any other online service provider. Thus the site developers try to test input emails by regular expressions and then sending a verification email to the email owners. Although these methods seem to be sufficient for recognizing and keeping contact with the users, these methods face to some problems. The main problem is sending many emails to unverified email addresses. This makes spam detection robots unsure about our servers and may detect us as a spam sender. In this paper, we introduce a statistical approach based on hidden Markov model in order to verify the validity of the emails in first stage. This method detects 87.6% of the invalid emails correctly. In second step, the verified emails can be evaluated by using conventional methods with more confidence.

Key words: Invalid email address detection · Hidden markov model · Language model · Valid internet communication · Email address recognition · Email Address

INTRODUCTION

User authentication is important for every website. Websites authenticate users to give them more services and give them better instructions to use their website by sending them newsletters and other suggestions and coupons. This method is very useful because it is not expensive and consumes a little time from the employees and servers. There are many ways to authenticate users some just check that user is a human and not a program or machine like the systems which use captchas to authenticate the users. Although this method is very popular and useful, using captchas in signup process is not comfortable and easy to use for users especially the very old and very young ones and this causes users not to sign up in our website. Captchas have other problems which are not suitable for websites with mass users, because all people are not capable to enter captchas correctly due to some problems like not seeing very well or other things. So for above reasons captchas are not trustworthy. Another way to authenticate users is to ask their email addresses to authenticate them. However in this case you are not sure about validation of their email address because some people may try to just type some characters in format of a valid email address like "xzxczcxzcx@somthing.com". It is obvious that this email address is not valid because its combination of letters is not known. This email address will pass any regular expression to check email address format validation but we know that this is not a valid email address. So developers use another way to authenticate users. First they check email address format by regular expressions and then send a validation email to check user inputs validation, but this time there are some possibilities like not delivering the email to its destination on time or never delivered. So there will be some problems like missing some good users or this may introduce the servers as a spam sender after a while and this cause our site not working well. Since our web site needs to communicate with its users, emails are free and powerful enough to give a huge terrific must be used in a correct way. Using email addresses helps to offer some services like sending a forgotten password or any case like this. Thus user's valid email addresses are extraordinary important for any website.

In this paper a statistical approach is introduced in order to find out validation of an email address by using hidden Markov model which is based on language models of the email addresses collected and annotated before. The purpose is to detect validation of email addresses at the time a user is entering his email address in the application. It is done by analyzing order of the letters in

Correspondnig Author: Morteza Zahedi, Computer Engineering and IT Department, Shahrood University of Technology, Shahrood, Iran, Tel: +98-21-44234255.

an email addresses and the occurrence probability of a letter in each position and then deciding whether it is valid or not. To perform such a task first a database of valid and invalid email addresses must be create and then the hidden Markov model trained to automatically detect valid and invalid email addresses. To optimize the results hidden Markov model parameters can be changed and then compare them to ex-results. In this paper the hidden Markov model parameters discussed to adjust the program with highest performance. Later this approach and its unique benefits will be discussed to assure this method is working well.

The following part contains list of related works which use hidden Markov model to model sequence of features extracted from different kinds of data. Then continued by introducing the system overview and hidden Markov model and finally ended up by presentation and analytics of experimental results, conclusion and outlook.

State of the Art: Hidden Markov Model was successfully used in Cryptanalysis. Cryptanalysis is the study of methods for obtaining the meaning of encrypted information, without access to the secret information which is normally required to do so. Typically, this involves finding a secret key. Hidden Markov model is used in different research issues which try to model hidden rules and information in a sequence of data or signal like speech recognition [1-4] (speech recognition converts spoken words to machine-readable input), language processing and machine translation [5-9] (machine translation, sometimes referred to by the abbreviation MT, is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another) and gesture recognition [10-12] and sign language recognition [13-17].

Since most of these approaches work on word level for recognition and translation, they use hidden Markov model and language model for words. However this approach is used for letters and their combinations which is due to the application of email address validity detection. Since an email address is a stream of letters making a word or a concatenated combinational phrase, we can model an email address byhidden Markov model which is used in the applications listed above for speech recognition, machine translation, sign language recognition and gesture recognition.

Database: As statistical frameworks depend on data sets, some valid and invalid email addresses are needed to perform the project. Using of natural and real data is very important to make a statistical model very close to real word, so planned to make the database by employing the students of three universities located in a region by entering their email addresses. Thus expected the email addresses collected from the student with the same language and very similar habits and interests in a special region obeys some hidden rules which is the language model.

A data set was produced for the project with 14000 valid and exact the same number of invalid email addresses. This number of email addresses divided into 3 sets, 10000 of each group of emails in train set and 2000 of them in two groups of development and test sets.

To collect valid email addresses university students were asked to write their email addresses and to collect invalid email addresses there were two options, first to generate them by computer and using some random algorithms and second to manually generate them. Second method was chosen, because in this way we have exactly the same email addresses which somebody may enter as an invalid email address in one application is available. University students were asked to write some invalid email addresses which they may enter in a webpage. Some examples of valid and invalid email addresses are shown in the Table 1.

These emails were put into 3 groups absolutely randomly. The training set data is used to train the statistical model which is based on hidden Markov model. After training the parameters were optimized on the development set by tuning the parameters in order to get the best result on the development set. Finally the model was evaluated on the test set to calculate the recognition rate of the recognition system which clusters the valid and invalid emails.

Table 1: Examples of valid and invalid emails in the data set

Valid emails	Invalid emails
boy_love_008@ymail.com	vbcvjkcx@fhhfgh.com
persianblog21@yahoo.com	hdvdbvsd@dsdsds.cs
Master_rock60@yahoo.com	dsfdvcbd.wd@dwewq.dsa
iraniam58@gmail.com	adfjdsfdsf@ddsd.fed
rainboy_2002@msn.com	fdfdjfgie@dwsds.fdf
shining2003r@gmail.com	fjsgwrw@dsds.com
weblog_open@yahoo.com	wr35y3r@dsds.com
a_lone_girl@googlemail.com	hjdagfi3l@htgr.vom

Fig. 1: The topology of employed HMM

Hidden Markov Model: The decision making of this system employs continuous density hidden Markov models (abbreviated to HMM) in order to recognize the valid and invalid email addresses. The recognition of invalid email address is similar to spoken word recognition in the modeling of sequential samples. The topology of the HMM is shown in Figure 1.

There is a transition loop at each state and the maximum allowable transition which is set to two. One HMM considered for each word w = 1, ..., W. The basic decision rule used for the classification of $x_1^T = x_1, \ldots, x_t, \ldots x_T$ is:

$$r(x_1^T) = \text{argmax}_w P_r(w|x_t) = argmax_w (p_r(w) \times p(x|w)) \quad (1)$$

Where Pr(w) is the prior probability of class w and Pr(x_t|w) is the class conditional probability of x given class w. the Pr(x_t|w) is defined by:

$$p_r(x_t \mid w) = \max_{s_1^T} \prod_{t=1}^{T} pr(s_t \mid s_t - 1, w) \times pr(x_t \mid s_t, w) \quad (2)$$

Where S_1^T is the sequence of states $pr(s_t|s_{t-1},w)$ and $pr(x_t|s_t,w)$ are the transition probability and emission probability, respectively. The transition probability is calculated by simple counting. The Gaussian and Laplace functions have been implemented as emission probability distributions $pr(x_t|s_t,w)$ in the states where the experiments show that Gaussian functions for emission probability yields better results. To estimate $pr(s_t|s_{t-1},w)$ the maximum likelihood estimation method for the Gaussian and Laplace functions is used, i.e. standard deviation and mean deviation estimation, respectively. The number of states for the HMM of each word can be determined in two ways: minimum and average sequence length of the training samples. Mixture densities with a maximum number of five densities are used in each state. Viterbi algorithm is used to find the sequence of the HMM. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states - called the Viterbi path - that results in a sequence of observed events, especially in the context of Markov information sources and generally, hidden Markov models. The algorithm makes a number of assumptions:

- Both the observed events and hidden events must be in a sequence. This sequence often corresponds to time.
- These two sequences need to be aligned and an instance of an observed event needs to correspond to exactly one instance of a hidden event.
- Computing the most likely hidden sequence up to a certain point t must depend only on the observed event at point t and the most likely sequence at point t - 1.

These assumptions are all satisfied in a first-order hidden Markov model. In addition to the density-dependent estimation of the variances, we use pooling during the training of the HMM which means that we do not estimate variances for each density of the HMM, but instead we estimate one set of variances for all densities in each state of the model (state-dependent pooling) or for all densities in the complete model (word-dependent pooling).

Language Model: The language model models syntax, semantics and pragmatics of the email at the character level. The probability of Pr (w_1^N) as an a-priori probability of a character sequence w_1^N is provided by a stochastic model. This probability concerns only the constraints of written glosses from the email address. To estimate the language model probability we assume that an email address character sequence follows an (m - 1)-th order Markov chain. Therefore the probability of observing a character sequence w_1^N is calculated by:

$$pr(w_1^N) = \prod_{n=1}^{N} pr(w_n \mid w_1^{n-1}) = \prod_{n=1}^{N} pr(w_n \mid w_{n-m+1}^{n-1}) \quad (3)$$

The character sequence w_{n-m+1}^{n-1} which is denoted by h_n is named history of the character w_n with a history length of m - 1. In other words if the history length is equal to m, a character w_n depends on its history h_n which is named m-gram language model [18]. In particular, the language models with a history length of one, two and three are calledunigram, bigram and trigram respectively. This definition needs some assumption to workproperly:

- If the index n - m + 1 in w_{n+m-1}^{n+1} is smaller than one, it is equal to one.
- If the upper index n - 1 is smaller than the lower index n - m + 1, the w_{n+m-1}^{n+1} is an empty sequence and the language probability is equal to Pr(w1).

The maximum-likelihood principle is used for estimation and evaluation of the language model. We define an equivalent criterion of language model is named perplexity (PP) [18]. In the following, the perplexity of a character sequence w_1^N is defined by

$$PP = pr(w_1^N)^{-\frac{1}{N}} = \left[\prod_{n=1}^{N} p_r(w_n \mid h_n)\right]^{-\frac{1}{N}} \quad (4)$$

Which is the inverse geometric meaning of the product of the conditional probability $Pr(w_n|h_n)$ of all characters of the whole sequence. The perplexity is the average number of possible characters at each position of the entire text. Therefore the perplexity as a criterion in the training process has to be decreased as much as possible. The language model score is calculated by the negative logarithm of the language model probability:

$$\log PP = -\frac{1}{N}\sum_{n=1}^{N} \log p_r(w_n \mid h_n) \quad (5)$$

In this paper the explained language model is employed as baseline of the recognition system. The efficiency of language model is shown in speech recognition and sign language recognition [10,11] very well. However it is always used along with other features (acoustic features in speech recognition and visual features in gesture and sign language recognition). Since the language model which shows the hidden rules in order of characters in an email address is the only feature used for recognition, it is not expect to achieve the best result by using language model. Thus we use the language model as baseline of the experiments and then try to optimize hidden Markov models which are the base of employed language model.

Experimental Results: Here, the experimental results obtained by using the method on the database are denoted. The hidden Markov models are implemented for the email addresses by using LTI-LIB which is an object oriented library with a bunch of algorithms and data structures frequently used in the field of image processing, computer vision and pattern recognition.

First the hidden Markov models are trained for valid and invalid emails with our training data. As the email addresses contain sequence of English characters first a value for any character should be assigned. If the ASCII value of any character assigned to its corresponding letter, it means the neighborhood of the characters is meaningless and cannot be used a model which is based on neighborhood of the letters and changing the values assigned to the characters continuously. However, experiments are started by assigning the ASCII values for the characters and special restrictions which simplifies the hidden Markov model to a language model. After training the models, they are tested with development data which leads to a recognition rate of 61.5% the experiments are conducted again by using the repeating percentage of any character in English language. This method improved the results and recognition rate is improved to 76.2%. This improvement in the result made us to calculate the repeating percentage of any character in the training valid email addresses. All values calculated and the model trained with the value of frequency of any letter in the training data for each character. As the language model depends on the people living in a special region, the recognition system expected to recognizes the invalid emails more correctly. Recognition rate 78.3% presents the improvement in recognition system that relied on the data collected in a closed region with a determined language model. The results which are obtained by using different assignment methods are summarized in the Table 3. Although we expect improving the recognition rate by changing the proposed method for assignment of the characters into a value to be used in hidden Markov model, the concentration is more on other parameters of hidden Markov model which are not still optimized for this application.

Table 2: Email groups and number of them in the training, development and test parts

Emails Group	Valid emails	Invalid emails
Training set	10,000	10,000
Development set	2,000	2,000
Test set	2,000	2,000
Total	14,000	14,000

Table 3: Different methods for assignment function which maps a character into a value

Assignment method	Recognition rate
ASCII values	61.5%
Frequency in English language	76.2%
Frequency in the training set	78.3%

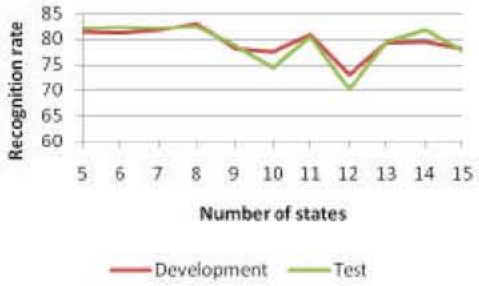

Fig. 2: The recognition rate with the changes in number of states

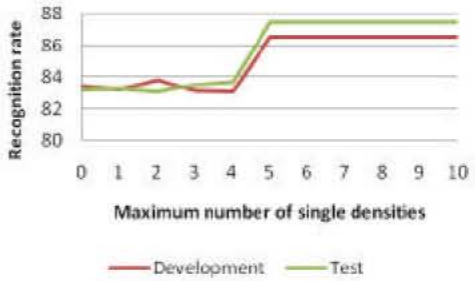

Fig. 3: Changes in maximum number of single densities for the emission probability of the states

After training phase hidden Markov model parameters on the development set are being optimized. First it is started with the hidden Markov model topology. The experiments are performed with different topologies of hidden Markov model to find the best one for the application. The topology of 0-1-2 is very popular to be used for hidden Markov model that allows only 1 state skipping or in other words maximum number of two jumps with a transition loop on any state. The empirical experiments show the 0-1-2 is the best topology for the application that allows a character loops in any state and skips a character maximum in one state. However number of states considered for any hidden Markov model is very important especially where the data set is not big enough to be sufficient for training the probability distributions of the model. The experimental results presented in the Figure 2 shows the achievement of the best recognition rate when the number of states is set to eight which is the average number of emails length in the training set.

The mixture densities for the emission probability of the states are also used [17]. The maximum number of single densities is changed in order to improve the recognition rate. The Fgure 3 shows that when the maximum number of single densities is set to lower than 4 our result getting worse but with the number of 5 and more there is no change in our results. So the maximum number for single densities constructing the mixture densities in each state is 5.

Table 4: The results obtained by different pooling types

Pooling type	Recognition rate
No pooling	87.1%
Model pooling	85.5%
State pooling	85.9%

Table 5: Summarized results

Parameters	Development	Evaluation
Language model	61.5%	65.0%
HMM(0-1-2)	65.5%	67.3%
Frequency in English language	76.2%	78.1%
Frequency in the training set	78.3%	79.2%
Optimizing number of states	83.2%	83.0%
Max. number of single densities	86.6%	87.2%
Pooling type	87.1%	87.6%

As it is noticed before variance pooling employed over a state or on a model where distributing the variance targeted in the probability distribution. First the experiments are done with no pooling which means any variance value is calculated for any feature value in any state and for any single density, then with state pooling which is performed over the single densities of a mixture density in a state and finally model pooling in which the variances are pooled for all the densities existing in a hidden Markov model. The experimental results shown in the Table 4 indicates the best result is obtained with no pooling for the variance. The result is reasonable due to the data set which is not too big to need a pooling over the densities.

Now, the model is evaluated by using training set and optimized by tuning the parameters on the development set. It is concluded a 0-1-2 topology, eight states for each model, maximum number of 5 densities for any mixture density and no pooling over the states. The experiment is done on the test set introduced before and the results show the model optimized to obtain 87.1% of recognition rate on the test set. The experimental results on the development set and test set are summarized in the Table 5.

As it is shown in the table, using hidden Markov model instead of simple language model helps us to improve the recognition rate from 61.5% to 65.5%. Then we promote the recognition system by optimizing the parameters of hidden Markov model leading us to obtain the best recognition rate of 87.1%.

CONCLUTION AND OUTLOOK

In this paper a new statistical method was presented to detect invalid email addresses by using language model and hidden Markov model.

The recognition rate of 87.1% shows the ability and efficiency of the method in the field of invalid email detection. This method can also used to filter bad words or in any other applications. For example a program can be developed to detect some groups of words automatically in order to filter web sites with unwanted contents or a program to disable cheating in entering data in applications in any way like entering letter by letter or with some spaces in between or adding a bad word exactly after another word or before that.

Although the hidden Markov model parameters were optimized for this application, it seems it is more useful to optimize the assigning method which is discussed in this paper. Furthermore we expect better results if we change the distance function along changing the assignment method. The distance function is used to calculate the difference between a character and another one. Currently the distance function is set to calculate subtract of frequency of occurrence of the characters in the training data but it can be optimized more correctly by using the key maps on the keyboard and since computer users use the keys in a close neighborhood the key maps should be considered according to users habits.

In addition, as each character is modeled by a continuous density HMM and the feature is ASCII code of each character or letter, where the range of standard ASCII code is limited (0-256), thus quantization error is not important in this case. Although, this application is carried out in real time in current state of the database, for larger databases the character can be modeled by discrete density HMM (DDHMM) which works faster with the same results.

ACKNOWLEDGMENTS

We are very thankful for the Information and Communication Treasure Company (ICT Co.) and Shahrood University of Technology (SUT) funding this research. Also we appreciate the efforts of the students of SUT who helped us very much producing the training data set.

REFERENCES

1. Rabiner, L.R., 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition., Proceedings of the IEEE, 77(2): 257-286, February 1989.
2. Jelink, F., 1998. Statistical Methods for Speech Recognition, MIT Press, Cambridge, Massachusetts, January 1998.
3. Kanthak, S., S. Molau, A. Sixtus, R. Schluter and H. Ney, 2000. The RWTH Large Vocabulary Speech Recognition System for Spontaneous Speech., Proceeding of the Konvens 2000, pp: 249-254. Ilmenau, Germany, 2000.
4. Lööf, J., M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter and H. Ney, 2006. *The 2006 RWTH parliamentary speeches transcription system*, In Proceeding of ICSLP, Pittsburgh, PA, USA, September 2006.
5. Huenerfauth, M., 2004. A Multi-path Architecture for Machine Translation of English text into American Sign Language Animation. ", Proc. of Student Workshop at Human Language Technologies Conference HLT-NAACL, Boston, MA, USA, May 2004.
6. Chiu, Y.H., C.H. Wu, H.Y. Su, C.J. Cheng, 2007. Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis. IEEE Trans. Pattern Analysis and Machine Intelligence, 1(28): 208-309, Jan. 2007.
7. S´af´ar, E. and I. Marshall, 2002. Sign Language Generation Using HPSG., Proc. of 9th International Conference on Theoretical and Methodological Issues in Machine Translation, pp: 105-114, TMI, Japan, March 2002.
8. Morrissey, S. and A. Way, An Example-Based Approach to Translating Sign Language. Workshop Example-Based Machine Translation (MT X–05), pp: 109-116, Phuket, Thailand, September 2005.
9. Yonggang Deng Byrne, W., 2008. HMM Word and Phrase Alignment for Statistical Machine Translation", Audio, Speech and Language Processing, IEEE Transactions on March 2008 16(3): 494-507.
10. Cui, Y., D. Swets and J. Weng, 1995. Learning-Based Hand Sign Recognition Using SHOSLIF-M., Proceeding of Int. Workshop on Automatic Face and Gesture Recognition, pp: 201-206, Zurich, 1995.
11. Quek, M.Z.F., 1996. Inductive Learning in Hand Pose Recognition., Proc. of Second IEEE Int. Conf. on Automatic Face and Gesture Recognition, Washington, DC, USA, IEEE.
12. Triesch, J., C. von der Malsburg, A System for Person- Independent Hand Posture Recognition against Complex Backgrounds., IEEE Trans. Pattern Analysis and Machine Intelligence, 23(12): 1449-1453. Dec., 2001.

13. Bauer, B. and H. Hienz, 2000. Relevant Features for Video-based Continuous Sign Language Recognition, in Proceedings of the 4th International Conference Automatic Face and Gesture Recognition Grenoble, France, pp: 440-445.
14. Vogler, C. and D. Metaxas, Adapting Hidden Markov Models for ASL Recognition by Using Three-dimentional Computer Vision Methods, in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Orlando, FL., pp: 156-161.
15. Starner, T., J. Weaver and A. Pentland, Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video, in Transaction of Pattern Analysis and Machine Intelligence, 20(2): 1371-1375.
16. Bowden, R., D. Windridge, T. Kabir, A. Zisserman and M. Bardy, 2004. A Linguaistic Feature Vector for the Visual Interpretation of Sign Language, in Proceedings of ECCV 2004, the 8th European Conference on Computer Vision, Prague, Czech Republic.
17. Dreuw, P., D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth and H. Ney. 2008. Spoken Language Processing Techniques for Sign Language Recognition and Translation.", Technology and Disability Journal, 20(2): 121-133, Amsterdam, The Netherlands, June 2008.
18. Bahl, L.R., F. Jelinek and R.L. Mercer, 1983. A Maximum Likelihood Approach to Continuous Speech Recognition., IEEE Transactions on Pattern Analysis and Machine Intelligence, 5: 179-190, March 1983.