2C4-IOS-3c-5

A Collective Intelligence Approach to Detecting IDN Phishing

SHIAN-SHYONG TSENG[1,2], AI-CHIN LU[2], CHING-HENG KU[2], GUANG-GANG GENG[3]
[1]Dept. of Applied Informatics and Multimedia, Asia University, Taichung, Taiwan
[2]Taiwan Network Information Center, Taipei, Taiwan
[3]China Internet Network Information Center, Computer Network Information Center,
Chinese Academy of Sciences, Beijing, China
{sstseng, aclu, chku}@twnic.net.tw, gengguanggang@cnnic.cn

In recent years, with the rapid growth of the Internet applications and services, phishing becomes one of the most severe threats on the Internet. The advent of internationalized domain name (IDN) has introduced a new threat with the non-English character sets allowing visual mimicry of common domain names.

The IDN homograph attack is a way that a malicious party may deceive computer users, especially, in the Chinese domain name related to the Chinese-homograph, denoting a group of different Chinese characters within the similar shape but different meanings, and the Chinese synonyms, denoting a group of the different words or phrases within the same meaning as another. Both of them can easily cause user confusion, resulting in the possibility of the phishing, for example, “栓” v.s. “拴”, “李” v.s. “季”, “未” v.s. “末”.

Our idea is to apply the collective intelligence approach to construct a Chinese-homograph and Chinese synonym database by Internet crowd collectively, so that the IDN phishing can be easily detected by consulting the database. A website is created to collect the Chinese-homograph and Chinese synonyms that include abbreviations and reversed words. Besides, the data validation has also been implemented by the crowdvoting method to increase the trustworthiness of our database.

By our approach, the detection of Chinese IDN phishing consists of three stages: suspect detection, website checking, and confirmation of phishing website. In the experimental result, the database is efficiently and effectively constructed, where 881 items of the Chinese-homograph and 3552 items of the Chinese synonyms have been created. In the future, the database will be used in the Internet browser or email client to achieve Chinese-homograph identification or blocking.

Key-Words: - Collective intelligence, IDN, Phishing, Chinese-Homograph, Chinese Synonym

1. Introduction

1.1 Background and Motivation

Phishing refers to the attacker's use of deceptive e-mail and web site for fraud. The victims often divulge their personal information and financial data, including the technical data, personal contact, e-mail, bank account number, password, etc. The information is used for future target advertisements or theft attacks (e.g., transfer money from victims' bank account) [1]. According to the report of Anti-phishing Working Group (APWG) [2], most phishing occurs on hacked or compromised web servers. In 2012, Anti-Phishing Alliance of China (APAC) [3] handled 24,535 phishing websites, where the distribution of phishing websites remains mainly in payment/transaction, finance/securities and media/communication websites or pages.

The approval of the Internationalized Domain Name (IDN) country code Top-Level Domain (ccTLD) Fast Track Process[4] by the ICANN Board in October 2009 enabled countries and territories to submit requests to ICANN for IDN ccTLDs representing their respective country or territory names in scripts, such as Arabic, Chinese, Russian, etc., other than US-ASCII characters. These are the domain names that contain one or more characters that do not belong to a Latin-based western language. Therefore, the IDN-enabled web application which may contain Chinese words displayed in the browser can benefit Chinese people to access the Internet.

Unfortunately, the IDN-based phishings are developed and deployed to attack the websites involving IDNs. In October 2009, Symantec [5] observed 10 phishing websites that contained IDNs. Anyone of these phishing Web sites was leveraging international characters resembling ASCII characters to spoof a western brand’s domain name. Besides, the IDN homograph attack [6] becomes a new way to deceive computer users. Especially, in the Chinese domain name, the Chinese-homograph, that is a group of different Chinese characters within the similar shape but different meanings, and the Chinese synonyms, that is defined as a group of the different word or phrase within the same meaning as another, can easily cause user confusion, resulting in the increase of the possibility of phishing; for example, the Chinese character “栓” v.s. “拴”, “李” v.s. “季”, “未” v.s. “末”, etc. This kind of potential threat is difficult to be resolved.

In this study, a Chinese-homograph and Chinese synonym database for IDN is proposed to cope with the above issue, so that the threat of the homograph attack can be easily detected by consulting the database. However, the construction and maintenance of the database needs a lot of experts and users to contribute their human expertise and user’s experience, where the more people participate the more it can identify confusing words. Hence, the idea of this paper is to use the collective intelligence approach [6-7] to construct the Chinese-homograph and Chinese synonym (including abbreviations and reversed

Contact: Ching-Heng Ku, Taiwan Network Information Center, 4F-2, No.9, Roosevelt Road, Sec.2, Taipei, Taiwan, Tel: +886-2-2341-1313, Fax: +886-2-23968871, Email: chku@twnic.net.tw

words) database, where the Internet crowd can collectively detect and report the existence of IDN Phishing. Besides, the data validation has also been implemented by the corwdvoting method to increase the trustworthiness of our database.

1.2 Related work

In multilingual computer systems, different logical characters may have almost identical appearances. The problem arises from the different treatment of the characters in the user's mind and the computer's programming. Internationalized domain names provide a backward-compatible way for domain names to use the full Unicode character set which is already widely supported.

According to the US-CERT report of technical trends in phishing attacks [8], International Domain Names in Applications (IDNA) uses an encoding syntax called Punycode [9] to represent Unicode characters in ASCII format. A web browser that supports IDNA would interpret this syntax to display the Unicode characters when appropriate. Users of web browsers that support IDNA could be susceptible to phishing via homograph attacks [10], where an attacker could register a domain that contains a Unicode character that appears identical to an ASCII character in a legitimate site. While a proof-of-concept of this type of attack was made, there is no public report of the IDNA abuse within a phishing scam.

The registration of homographic domain names is akin to typosquatting. The major difference is that in typosquatting the perpetrator relies on natural human typos, while in homograph spoofing [11-12] the perpetrator intentionally deceives the web surfer with visually similar names. An attacker could register a domain name that looks just like that of a legitimate website, but in which some of the letters have been replaced by homographs in another alphabet. Some homographs in internationalized domain names [13], such as Cyrillic, Greek, Armenian, and Hebrew, have been collected. These are only the most obvious and common. The possibilities are far more numerous than can be listed there.

Those homographs are based on the alphabet system, so they are substantially different from the language based on the symbol system, like Chinese. In Chinese, different logical characters may have identical or similar appearances. Currently, it still lacks a rich database for the Chinese-homograph and Chinese synonym to deal with the phishing attack. Hence, the collection the Chinese-homograph and Chinese synonym for the prevention of the phishing becomes an important issue.

2. Collection of Chinese-Homograph and Chinese Synonym by Collective Intelligence

Chinese words are based on the Chinese character that is totally different from those in the English system based on the composition of the alphabet. This characteristic causes the existence of many Chinese-homograph and Chinese synonym from the visual characteristics of Chinese words. For example, some different Chinese characters may have the same pronunciation and the similar shapes; some different Chinese characters may have similar pronunciation and similar shapes; some different Chinese characters within similar shapes may have different meanings; and some different words may have the same meaning. In this study, we are concerned with the problem that is how to efficiently and effectively collect the Chinese-homograph and Chinese synonym for the prevention of the user confusion in the Chinese domain name. Hence, the collective intelligence approach is proposed and described in the following.

2.1 Collective intelligence approach for Chinese-homograph and Chinese synonym database

According to Don Tapscott and Anthony D. Williams, the collective intelligence is the mass collaboration [14]. In order for this concept to happen, four principles need to exist; openness, peering, sharing, and acting globally. The proposed structure of the collective intelligence approach for the collection of the Chinese-homograph and Chinese synonym is composed of two directions, such as competencies and mechanics, as shown in Figure 1. The competencies are based on the organization's domain specific knowledge on the Chinese IDN. The mechanics are based on the culture norms on the Chinese words.

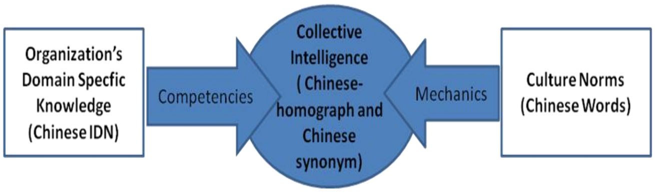

Organization's Domain Specfic Knowledge (Chinese IDN)
Competencies
Collective Intelligence (Chinese-homograph and Chinese synonym)
Mechanics
Culture Norms (Chinese Words)

Figure 1. The Structure of the Collective Intelligence Approach

In this study, the collective intelligence is used as a group intelligence that emerges from the collaboration of many individuals. The collection of the Chinese-homograph and Chinese synonym is inputted by experts or individuals. The Chinese language experts provide the data of Chinese words, including Chinese-homographs, similar words, and Chinese synonyms. The general public could input the idiom related to Chinese synonyms by the Chinese culture, habit, and norm. Besides, the registrants of the Chinese IDN also could input the Chinese synonym related to his IDN.

The schema of the database includes the following categories: Chinese-homograph or Chinese synonym, serial numbers of categories, the original word, the corresponding list of Chinese-homographs or Chinese synonyms, the time stamp of data input, the score of the crowdvoting, and the status that appears in the phishing website. For example, in the database, if the original Chinese word is the "已", the corresponding list of Chinese-homographs (similar words) are "己" and "巳". These words will be treated as the possible candidates that will be validated by the crowdvoting method described in the following.

2.2 Consensus Building for the validation of the Chinese-homograph and Chinese synonym by the Crowdvoting

In this study, the collective intelligence not only is used as the collection of the Chinese words and phrases, but also appears in the consensus decision making for the validation of the Chinese-homograph and Chinese synonym, as shown in Figure 2.

The crowdvoting approach comes from the web-based crowdsourcing efforts [15-16], where the crowdsourcing is an online, distributed problem-solving and has some common categories that can be used effectively in the commercial world. Crowdvoting occurs when a website gathers a large group's opinions and judgment on a certain topic. We use the crowdvoting method to validate the Chinese-homographs and Chinese synonyms. The score of the vote ranges from 1 to 5, where number 1 and number 5 represent the full disagreement and the full agreement, respectively. The more score presents the more degree of the agreement, and vice versa.

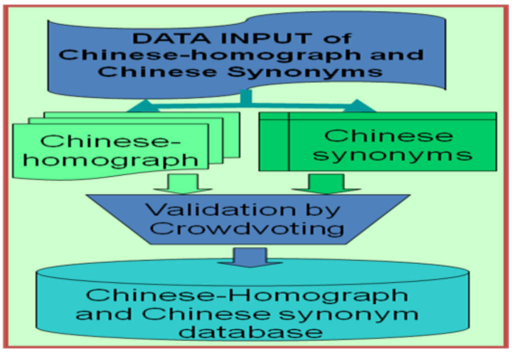


Figure 2. The flow of the update of the Chinese-homograph and Chinese synonym database

3. Architecture of Detection of IDN Phishing based on Chinese-Homograph and Chinese Synonym Database

In this section, the architecture of the detection of IDN phishing based on the Chinese-homograph and Chinese synonym database is described. In Section 3.1, we introduce the construction of the proposed Chinese-homograph and Chinese synonym database. In Section 3.2, the architecture of the detection of the IDN phishing is described.

3.1 The construction of the Chinese-homograph and Chinese synonym database

The Chinese-homograph is a group of different Chinese characters within the similar shape but different meanings. They may have the same or different pronunciations. The synonym is defined as a group of the different word, or phrase within the same meaning as another, in some or all uses. In Chinese phrase, the synonym sometimes appears in the abbreviation or that can be written by the reverse. The abbreviation is a shorter way to write a phrase. In Chinese, some phrases written by the reverse words have the same meaning with the original one, especially it often appears in the phrase within two Chinese words.

The website is created to collect the Chinese-homograph and Chinese synonyms by the collective intelligence method, described in Section 2.1. Besides, the data validation has also been implemented by the crowdvoting method, described in Section 2.2, to increase the trustworthiness of our database. The detail attributes of the database within the categories, words, the score of the crowdvoting, and the validation flag are shown in the following.

Field	Data Type	Description
Type_id	Integer(4)	Chinese-homograph or Chinese Synonyms
Serial_id	Integer(4)	Serial Number of the Category
Original_ Words	Character(32)	Original words
Homograph h_words	<Character(32), Character(32),…>	Chinese-homograph
Syn_words	<Character(32), Character(32),…>	Chinese Synonyms
First_time	Timestamp with time zone	Time stamp of Data input
Score	Integer(4)	Score of the crowdvoting
Validation_ flag	Integer(4)	Validation status
Phishing_ flag	Integer(4)	Ever appeared in the phishing website

3.2 Architecture of the detection of IDN Phishing

In this study, the detection of phishing website related to the Chinese IDN is divided into three stages, such as suspect detection, website checking, and confirmation, as shown in Figure 3.

The suspect detection stage is to find the possible suspect of the phishing Chinese IDN based on the proposed database. The stage for the website checking is to analyze the content or the activity of the suspicious phishing website. The last stage is to make the confirmation of the phishing website.

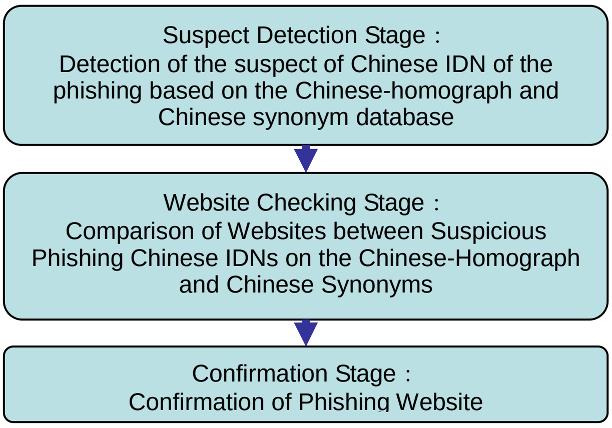
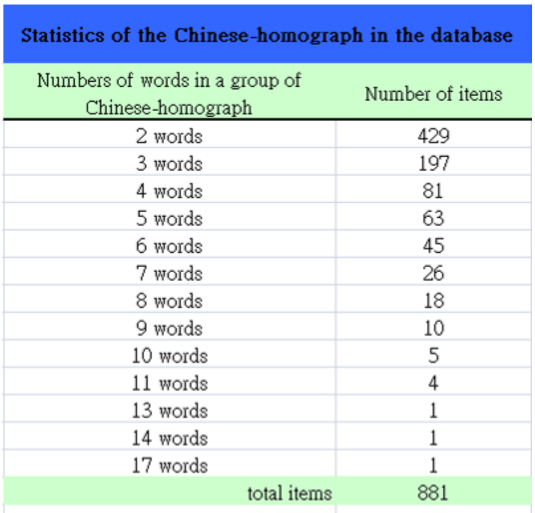

Suspect Detection Stage :
Detection of the suspect of Chinese IDN of the phishing based on the Chinese-homograph and Chinese synonym database
Website Checking Stage :
Comparison of Websites between Suspicious Phishing Chinese IDNs on the Chinese-Homograph and Chinese Synonyms
Confirmation Stage :
Confirmation of Phishing Website

Figure 3. Architecture of IDN Phishing Detection based on Chinese-Homograph and Chinese Synonym Database

4. Experiment

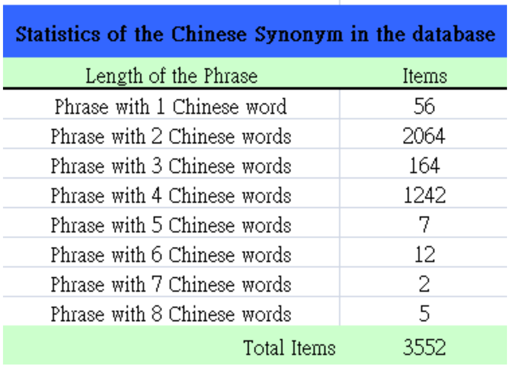
The prototype of the Chinese-homograph and Chinese synonym database has been constructed based on our approach. There are 881 items of the Chinese-homograph within similar Chinese words in the database, as shown in Figure 4. Besides, 3552 items of the Chinese synonyms are stored in the database, as shown in Figure 5. The interface of the query of the Chinese-homograph is shown in Figure 6. The interface of the query of the Chinese synonyms is shown in Figure 7.



Statistics of the Chinese-homograph in the database

Numbers of words in a group of Chinese-homograph	Number of items
2 words	429
3 words	197
4 words	81
5 words	63
6 words	45
7 words	26
8 words	18
9 words	10
10 words	5
11 words	4
13 words	1
14 words	1
17 words	1
total items	881

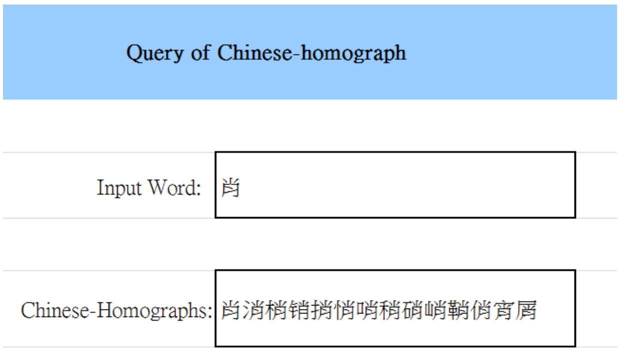
Figure 4. The statistics of the Chinese-homograph in the database.



Statistics of the Chinese Synonym in the database

Length of the Phrase	Items
Phrase with 1 Chinese word	56
Phrase with 2 Chinese words	2064
Phrase with 3 Chinese words	164
Phrase with 4 Chinese words	1242
Phrase with 5 Chinese words	7
Phrase with 6 Chinese words	12
Phrase with 7 Chinese words	2
Phrase with 8 Chinese words	5
Total Items	3552

Figure 5. The statistics of the Chinese Synonym in the database.



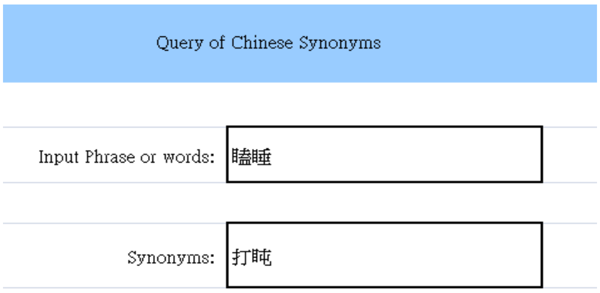
Query of Chinese-homograph

Input Word: 肖

Chinese-Homographs: 肖消梢销捎悄哨稍硝峭鞘俏宵屑

Figure 6. Query of database on the Chinese-homograph.

According to the constructed Chinese-homograph and Chinese synonym database, the IDN phsihing from the homograph attack can be easily detected. In Figure 8, the process of the detection of the Phishing website related to the Chinese IDN within the homograph and synonym has been illustrated. The desired Chinese IDN will be compared with the existing Chinese IDN based on the constructed Chinese-homograph and Chinese synonym database. The suspicious phishing website will be checked by the content or the activity of the website.



Query of Chinese Synonyms

Input Phrase or words: 瞌睡

Synonyms: 打盹

Figure 7. Query of the database on the Chinese synonym.

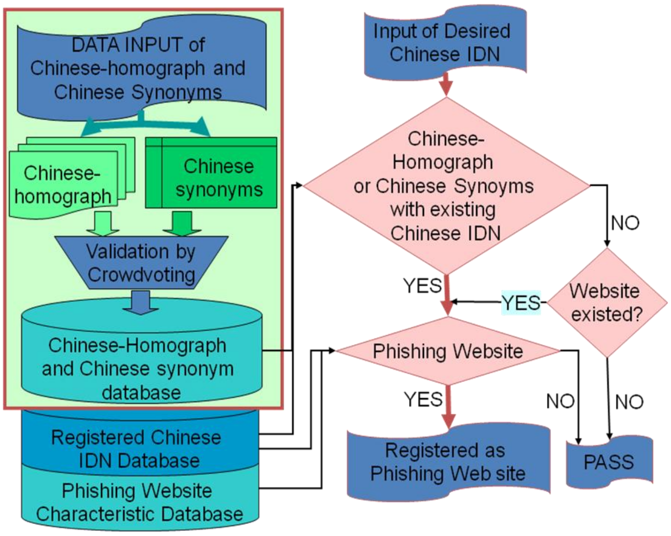

DATA INPUT of Chinese-homograph and Chinese Synonyms
Chinese-homograph
Chinese synonyms
Validation by Crowdvoting
Chinese-Homograph and Chinese synonym database
Registered Chinese IDN Database
Phishing Website Characteristic Database
Input of Desired Chinese IDN
Chinese-Homograph or Chinese Synoyms with existing Chinese IDN
NO
YES
YES
Website existed?
Phishing Website
YES
NO
NO
Registered as Phishing Web site
PASS

Figure 8. Flowchart of the detection of the Phishing Website related to the Chinese IDN of the Chinese-homograph and Chinese synonym.

5. Conclusion

In this paper, we successfully proposed a collective intelligence approach which aims to construct the Chinese-homograph and Chinese synonym database by Internet crowd collectively. Besides, the data validation has also been implemented by the crowdvoting method to increase the trustworthiness of our database.

Accordingly, we developed the architecture of IDN phishing detection based on the proposed database. Therefore, our approach for the detection of Chinese IDN phishing consists of three stages, such as finding the suspecious phishing IDN, checking the suspecious phishing website, and the confirmation of the phishing website.

In the experiment, we successfully construct the Chinese-homograph and Chinese synonym database within 881 items of the Chinese-homograph, and 3552 items of the Chinese synonyms. Besides, the flowchart of the detection of the Phishing website related to the Chinese IDN of the Chinese-homograph and Chinese synonym is also proposed. The research result can also be used in the Internet browser or email client to achieve homograph identification or blocking in the future.

Acknowledge

This paper was partially sponsored by the National Science council under Grant No. NSC 101-2511-S-468-007-MY3 and the China Internet Network Information Center (CNNIC) under Grant No. DNSLAB-2012-S-U.

References

[1] Ming Qi and Chang-Yi Zou, “A study of anti-phishing strategies based on TRIZ”, Proceedings of 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009, PP. 536-538.

[2] Phishing Attack Trends Report -3Q2012, Anti-phishing Working Group(APWG), February 1, 2013, http://www.antiphishing.org/

[3] 2012 annual report of Anti-Phishing Alliance of China(APAC), 2012, http://en.apac.cn/news/201301/P020130122639769507177.pdf

[4] Internationalized Domain Names (IDNs), ICANN, http://www.icann.org/en/resources/idn

[5] IDNs in Phishing, Symantec, June 2009, http://www.symantec.com/connect/blogs/idns-phishing

[6] André Boder, “Collective intelligence: a keystone in knowledge management”, Journal of Knowledge Management, 1997.

[7] Martijn C. Schut, “The Scientific Handbook for Simulation of Collective Intelligence”, Version: 2, February 2007.

[8] Jason Milletary, “Technical trends in Phishing attacks”, http://www.us-cert.gov/sites/default/files/publications/phishing_trends0511.pdf, US-CERT.

[9] Costello, “Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA),” March, 2003, http://www.ietf.org/rfc/rfc3492.txt.

[10] Evgeniy Gabrilovich and Alex Gontmakher, “The Homograph Attack,” Communications of the ACM, 45(2):128, February 2002, http://www.cs.technion.ac.il/~gabr/papers/homograph_full.pdf

[11] Johanson, Eric, “The State of Homograph Attacks”, rev1.1, The Shmoo Group, 2005.

[12] Evgeniy Gabrilovich and Alex Gontmakher, “The Homograph Attack”, Communications of the ACM., February 2002.

[13] IDN homograph attack, Wikipedia, http://en.wikipedia.org/wiki/IDN_homograph_attack

[14] Collective intelligence, Wikipedia, http://en.wikipedia.org/wiki/Collective_intelligence

[15] Crowdsourcing, Wikipedia, http://en.wikipedia.org/wiki/Crowdsourcing

[16] Brabham, Daren, "Crowdsourcing as a Model for Problem Solving: An Introduction and Cases", Convergence: The International Journal of Research into New Media Technologies, vol. 14 (1), pp. 75–90, 2008.