財團法人台灣網路資訊中心因公出國人員報告書

102 年 11 月 13 日

報告人 姓　名	曾憲雄 呂愛琴 顧靜恆	服務單位及職稱	董事長 副執行長 網址組經理
出國期間	102 年 10 月 15-19 日	出國地點	中國大陸北京
出國事由	參加 The Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIM-HSP 2013) 研討會		
報告書內容應包含： 一、出國目的 二、考察、訪問過程 三、考察、訪問心得 四、建議意見 五、其他相關事項或資料 （內容超出一頁時，可由下頁寫起）			
授　權 聲明欄	本出國報告書同意貴中心有權重製發行供相關研發目的之公開利用。 授權人：　　　　　　（簽章）		

附一、請以「A4」大小紙張，橫式編排。出國人員有數人者，依會議類別或考察項目，彙整提出報告。
註二、請於授權聲明欄簽章，授權本中心重製發行公開利用。

一、出國目的

第九屆智慧資訊隱藏與多媒體訊號處理國際研討會議（The ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing，IIH-MSP 2013）。今年的會議是由 IEEE 和國立高雄應用科技大學（KUAS）作技術贊助，北京物資學院作經費贊助，並由中國大陸北京工業大學主辦。會議內容包括國際最新智慧資訊隱藏與多媒體訊號處理，以及網路相關應用等方面的研究論文發表與專題演講。

多媒體技術與智慧能力的不斷提高，對於創造一個全球性的資訊基礎環境的啟用過程是急需的，以將世界各地的異質電腦網路和各種形式的資訊技術互相連結。此次 IIH-MSP 2013 研討會議於 2013 年 10 月 16 日至 18 日假中國大陸北京舉行，參加此研討會除與各國專家學者進行經驗與學術交流，並為投稿的論文進行發表報告，論文題目為 Building a self-organizing phishing model based upon dynamic EMCUD。

二、考察、訪問過程

此次研討會安排了許多論文發表的場次以及大會所規劃特別主題的專題演講，大會會場照片請見圖一：

2013 Ninth International Conference On
Intelligent Information Hiding and Multimedia Signal Processing
October 16-18, 2013 • Beijing, China
IEEE
NSFC
HBiS
Texas Instruments

圖一、IIH-MSP 2013 大會會場

10 月 16 日至 10 月 18 日論文發表議程中，大會特地於 16-17 日上午場次邀請國際專家學者做專題演講，其餘投稿論文發表的場次也都以英文進行簡報，會議議程內容如下：

Conference Schedule

Conference Site: Grand Gongda Jianguo Hotel (工大建国饭店)

October 15, 2013	
09:00 ｜ 17:00	**Registration and Welcome Reception Venue: Lobby (1F) 一楼大厅**
October 16, 2013	
08:30 ｜ 12:30	**Registration and Welcome Reception Venue: Lobby (1F) 一楼大厅**

October 16, 2013			
08:30 ｜ 09:00	**Opening Venue: Grand Ballroom(3F) 三楼宴会厅**		
09:00 ｜ 10:00	**Invited Keynote Speech I – Prof. Wen Gao, a Member of Chinese Academy of Engineering, and a Fellow of IEEE. Perking University, China.** ***Title: Challenges to Video Surveillance System***		
10:00 ｜ 10:30	Coffee Break		
10:30 ｜ 11:30	**Invited Keynote Speech II – Prof. Ioannis Pitas, IEEE Fellow. Aristotle University of Thessaloniki, Greece.** ***Title: Human activity recognition for video content analysis and description.***		
11:30 ｜ 12:30	**Invited Keynote Speech V – Prof. Wan-Chi Siu, IEEE Fellow. Hong Kong Polytechnic University, Hong Kong.** ***Title: Challenges on Video Surveillance under Quality Expectations and Big Data Environment.***		
13:30 ｜ 15:45	Room A—— Hibiscus Room (3F) 芙蓉厅	Room B—— Rose Room (3F) 玫瑰厅	Room C—— Peony Room (3F) 牡丹厅
	Session IS10	Session IS02	Session IS13
15:45 ｜ 16:00	Coffee Break		
16:00 ｜ 18:45	Room A—— Hibiscus Room (3F) 芙蓉厅	Room B—— Rose Room (3F) 玫瑰厅	Room C—— Peony Room (3F) 牡丹厅
	Session IS03,IS05	Session IS07	Session IS08

<table>
<tr><th colspan="4">October 17, 2013</th></tr>
<tr><td colspan="4">08:30-12:00 Venue: Grand Ballroom (3F)</td></tr>
<tr><td>08:30
|
09:30</td><td colspan="3">Invited Keynote Speech IV – Prof. Patrizio Campisi, Dept. of Applied Electronics at the Universita degli Studi "Roma TRE", Italy.
Title: PassBrain: can brain waves be used to recognize people.</td></tr>
<tr><td>09:30
|
10:30</td><td colspan="3">Invited Keynote Speech III – Prof. Qionghai Dai, Tsinghua University, China.
Title: Ultra-fast Lens-less Computational Imaging.</td></tr>
<tr><td>10:30
|
11:00</td><td colspan="3">Coffee Break</td></tr>
<tr><td>11:00
|
12:00</td><td colspan="3">Invited Keynote Speech VI – Prof. Sin-Horng Chen, National Chiao Tung University, Taiwan.
Title: Mandarin Speech Prosody Modeling and its Applications to ASR and TTS.</td></tr>
<tr><td rowspan="2">13:00
|
15:25</td><td>Room A——
Hibiscus Room (3F)
芙蓉厅</td><td>Room B——
Rose Room (3F)
玫瑰厅</td><td>Room D——
Orchid&amp; Jasmine Room (3F)
兰花厅和茉莉厅</td></tr>
<tr><td>Session IS01</td><td>Session IS11</td><td>Session IS21</td></tr>
<tr><td>15:25
|
15:40</td><td colspan="3">Coffee Break and Post Session</td></tr>
<tr><td rowspan="2">15:30
|
17:50</td><td>Room A——
Hibiscus Room (3F)
芙蓉厅</td><td>Room B——
Rose Room (3F)
玫瑰厅</td><td>Room D——
Orchid&amp; Jasmine Room (3F)
兰花厅和茉莉厅</td></tr>
<tr><td>Session IS04</td><td>Session IS12</td><td>Session IS22</td></tr>
<tr><td>18:00
|
20:00</td><td colspan="3">Banquet</td></tr>
</table>

October 18, 2013			
08:30 | 10:25	Room A—— Hibiscus Room (3F) 芙蓉厅	Room B—— Rose Room (3F) 玫瑰厅	Room D—— Orchid& Jasmine Room (3F) 兰花厅和茉莉厅
	Session IS16	Session IS17	Session IS20
10:25 | 10:40	Coffee Break		
10:40 | 12:30	Room A—— Hibiscus Room (3F) 芙蓉厅	Room B—— Rose Room (3F) 玫瑰厅	Room D—— Orchid& Jasmine Room (3F) 兰花厅和茉莉厅
	Session IS18, IS19	Session IS14	Session IS15

三、考察、訪問心得

此次大會邀請了許多 IEEE Fellow 進行專題演講，11月16日上午邀請了中國大陸的Prof. Wen Gao（見圖二）、希臘的Prof. Ioannis Pitas（見圖三）以及香港的Prof. Wan-Chi Siu 進行專題演講，講題分別為 Challenges to Video Surveillance System、Human activity recognition for video content analysis and description 以及 Challenges on Video Surveillance under Quality Expectations and Big Data Environment。

11月17日上午邀請了義大利的 Prof. Patrizio Campisi、中國大陸的 Prof. Qionghai Dai（見圖四）以及台灣的 Prof. Sin-Horng Chen（陳信宏教授）進行專題演講，講題分別為 PassBrain: can brain waves be used to recognize people、Ultra-fast Lens-less Computational Imaging 以及 Prosody Modeling of Mandarin and its Applications（見圖五）。

圖二、Prof. Wen Gao 專題演講，講題為 Challenges to Video Surveillance System

Human activity recognition for vi...
content analysis and des...
Information
Aristotle Unive...
IMPART

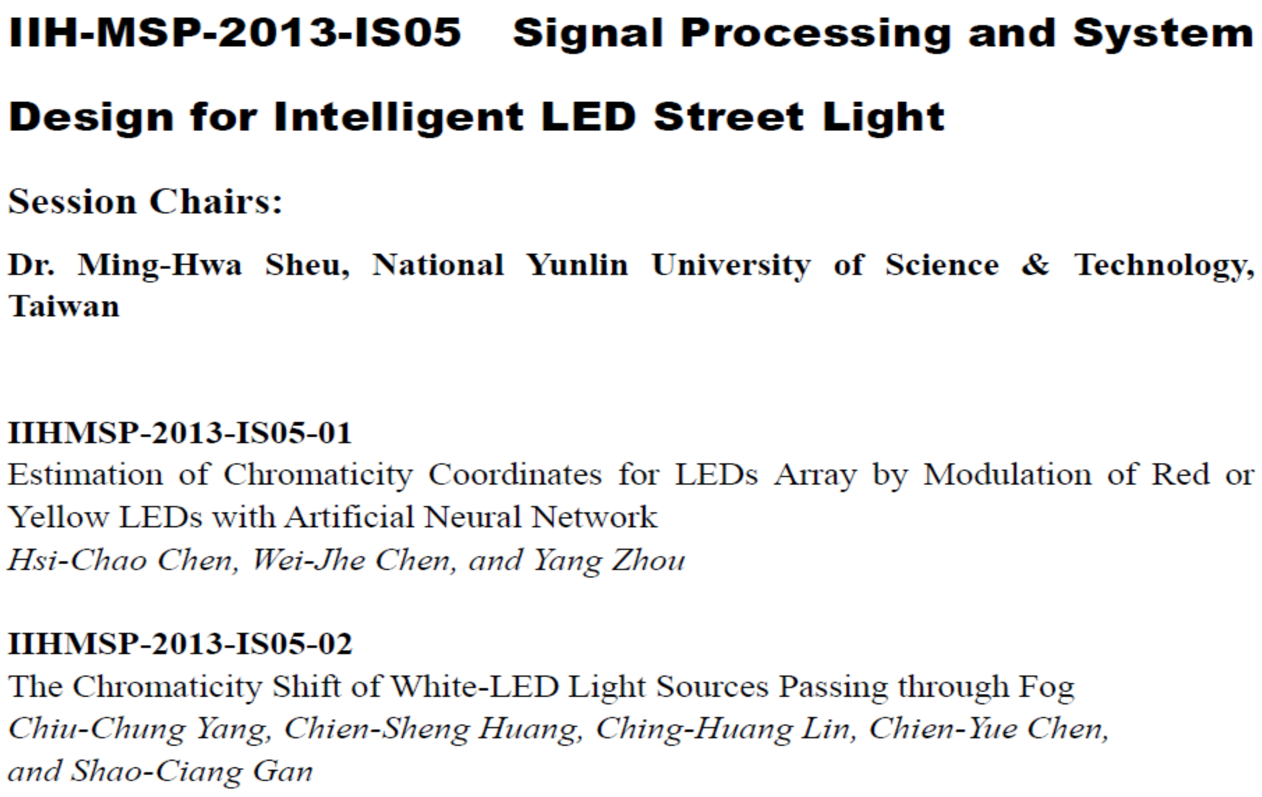
圖三、Prof. Ioannis Pitas 專題演講，講題為 Human activity recognition for video content analysis and description

圖四、Prof. Qionghai Dai 專題演講，講題為 Ultra-fast Lens-less Computational Imaging

Prosody Modeling of Mandarin
Speech and its Applications
Sin-Horng Chen
National Chiao Tung University
NCTU
2013/10/17

圖五、Prof. Sin-Horng Chen 專題演講，講題為 Prosody Modeling of Mandarin and its Applications

本次會議中投稿了一篇論文，並且被審核通過，安排在 10 月 16 日下午 Session IS05 的場次中發表。這篇論文，是由曾憲雄董事長、顧靜恆經理、呂愛琴副執行長、王智傑、耿光剛研究員共同撰寫的 Building a self-organizing phishing model based upon dynamic EMCUD 論文。該場次共有 7 篇論文發表，詳細議程如下：



IIH-MSP-2013-IS05 Signal Processing and System Design for Intelligent LED Street Light

Session Chairs:

Dr. Ming-Hwa Sheu, National Yunlin University of Science & Technology, Taiwan

IIHMSP-2013-IS05-01
Estimation of Chromaticity Coordinates for LEDs Array by Modulation of Red or Yellow LEDs with Artificial Neural Network
Hsi-Chao Chen, Wei-Jhe Chen, and Yang Zhou

IIHMSP-2013-IS05-02
The Chromaticity Shift of White-LED Light Sources Passing through Fog
Chiu-Chung Yang, Chien-Sheng Huang, Ching-Huang Lin, Chien-Yue Chen, and Shao-Ciang Gan

IIHMSP-2013-IS05-03
The Investigation on Physiological Influences and the Working Efficiencies of Several Lighting in Market
Chien-Yue Chen, Ming-Da Ke, Jeng-Han Wu, and Pei-Jung Wu

IIHMSP-2013-IS05-04
Multi-channel LED Driver with PWM Dimming and Temperature Self-Protection
Shih-Chang Hsia, Jyun-Jia Ciou, and Sheng-Yueh Lai

IIHMSP-2013-IS05-05
Fast Image Blending and Deghosting for Panoramic Video
Chen-Hong Yuan, Jeng-Shyang Pan, Ming-Hwa Sheu, and Tzu-Hsiung Chen

IIHMSP-2013-IS05-06
A Computation Efficiency AND-CFAR for FMCW Radar Receiver
Ho-En Liao, Guan-Yu Lin, Ming-Hwa Sheu, Siang-Min Siao, and Sin-Siang Wan

IIHMSP-2013-IS22-05
Building a Self-Organizing Phishing Model Based upon Dynamic EMCUD
Shian-Shyong Tseng, Ching-Heng Ku, Ai-Chin Lu, Yuh-Jye Wang, and Guang-Gang Geng

每一篇論文報告時間為 20 分鐘，並且接受大家的提問。此次會議曾憲雄董事長、呂愛琴副執行長及顧靜恆經理（見圖六）皆出席參加，並由顧靜恆經理代表進行論文簡報（見圖七），論文全文請詳見附件一。

2013 Ninth International Conference On
Intelligent Information Hiding and Multimedia Signal Processing
October 16-18, 2013 • Beijing, China

圖六、曾憲雄董事長(中)、呂愛琴副執行長(右)及顧靜恆經理(左) 於大會會場合影

Introduction of ideas of self-organizing phishing model
What are key factors to successfully build the self-organizing phishing model in this study?
We use the EMCUD method to build up the phishing attack knowledge according to the identification of phishing attributes
Since the phishing attack knowledge needs to be dynamically evolved over time, we use the VODKA method to evolve existing phishing knowledge to add modify phishing attributes and adapt the original embedded rules for phishing representative ones.

圖七、顧靜恆經理進行論文報告

本次會議有許多國際專家學者都出席參加，藉此機會互相觀摩學習，在會議中並與國立交通大學電機工程學系陳信宏特聘教授（見圖八）及國立中正大學資訊工程學系張真誠榮譽教授（見圖九）進行學術交流討論。陳信宏特聘教授除了於大會進行專題演講之外，亦發表了論文，題目為 A New Model-Based Prosody Coder for Mandarin Speech。張真誠榮譽教授發表之論文題目為 A Two-Staged Multi-level Reversible Data Hiding Exploiting Lagrange Interpolation。

圖八、曾憲雄董事長與國立交通大學電機工程學系陳信宏特聘教授合影

圖九、曾憲雄董事長與國立中正大學資訊工程學系張真誠榮譽教授合影

四、建議事項

(一) 網際網路的應用越來越廣泛，多媒體訊號處理以及智慧資訊隱藏處理等技術，於多媒體應用及智慧資料的分析上都是重要的發展趨勢，各國教授及研究人員的論文成果值得多加觀摩交流與學習。

(二) 此次研討會除有許多來自各國的教授參與，借由此研討會可以取得互相交流的機會，對於建立更多相互合作的機會有很大的幫助。

附件一：

The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 2013

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.

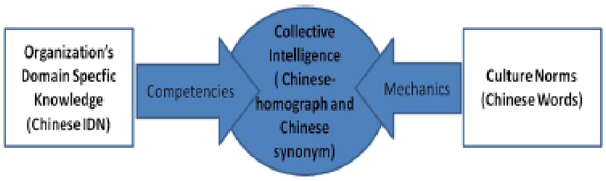

Organization's Domain Specfic Knowledge (Chinese IDN)
Competencies
Collective Intelligence (Chinese-homograph and Chinese synonym)
Mechanics
Culture Norms (Chinese Words)

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.

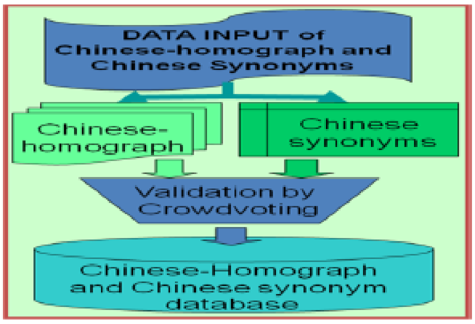


Figure 2. The flow of the update of the Chinese-homograph and Chinese synonym database

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
Homograp 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.

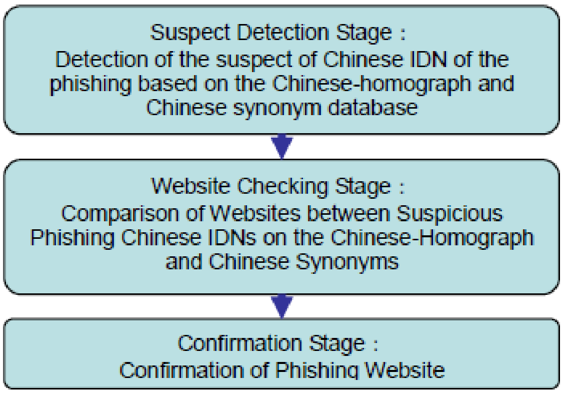
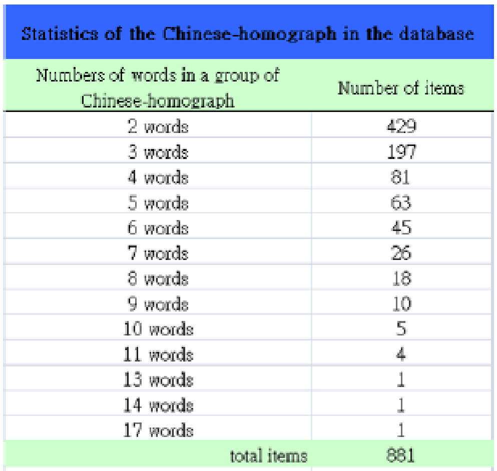

Suspect Detection Stage :
Detection of the suspect of Chinese IDN of the phishing based on the Chinese-homograph and Chinese synonym database
Website Checking Stage :
Comparison of Websites between Suspicious Phishing Chinese IDNs on the Chinese-Homograph and Chinese Synonyms
Confirmation Stage :
Confirmation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

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

4. Experiment

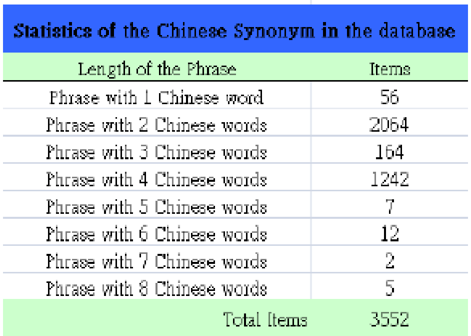
The prototype of the Chinese-homograph and Chinese synonym database has been constructed based on our approach. There are 881 items of the Chinese-homograph within similar Chinese words in the database, as shown in Figure 4. Besides, 3552 items of the Chinese synonyms are stored in the database, as shown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

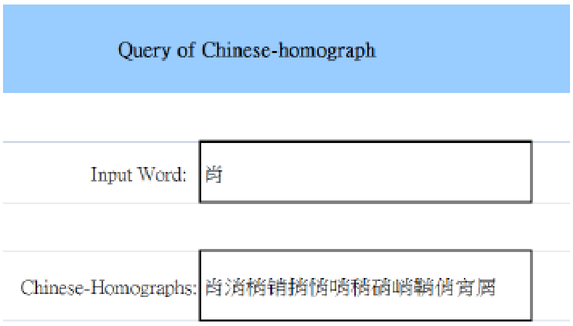
Figure 4. The statistics of the Chinese-homograph in the database.



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.



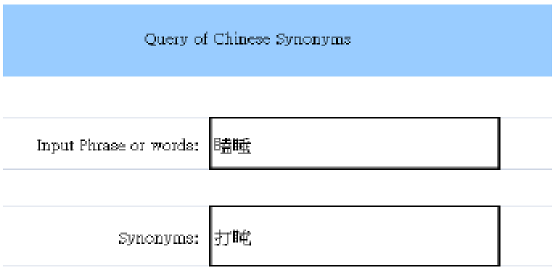
Query of Chinese-homograph

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.

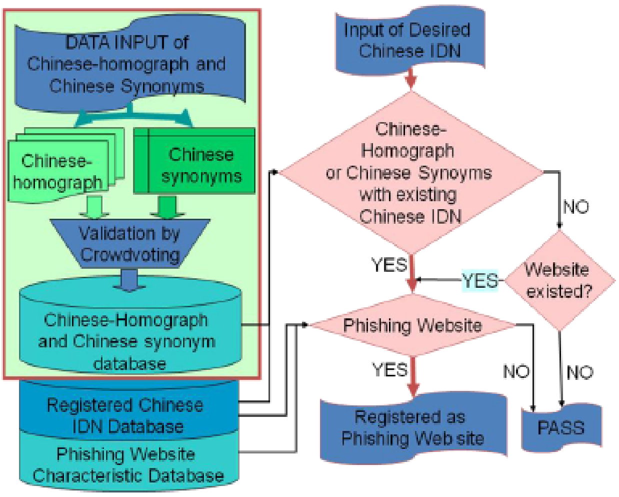

DATA INPUT of Chinese-homograph and Chinese Synonyms
Chinese-homograph
Chinese synonyms
Validation by Crowdvoting
Chinese-Homograph and Chinese synonym database
Registered Chinese IDN Database
Phishing Website Characteristic Database
Input of Desired Chinese IDN
Chinese-Homograph or Chinese Synoyms with 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.