Trademark Protection for Chinese Domain Names

Junling Zhang[1,2], Zhiwei Yan[1], Guanggang Geng[1], Xiaodong Lee[1], Jong-Hyouk Lee[3]

[1] China Internet Network Information Center, China
[2] Computer Network Information Center, Chinese Academy of Science, China
[3] Department of Computer Software Engineering, Sangmyung University, South Korea
{zhangjunling, yanzhiwei, gengguanggang, xl}@cnnic.cn, jonghyouk@smu.ac.kr

Abstract

Since the New Generic Top-level Domain (New gTLD) Program of Internet Corporation for Assigned Names and Numbers (ICANN) have carried out for some time, Chinese New gTLD and IDN have attracted a lot of attention from related stakeholders. At the same time, the protection for brand's domain name and intellectual property rights are undergoing a severe situation because of the huge character space of IDN and greatly expanded Domain Name System (DNS). In order to protect interests of trademark holders, ICANN developed the Trademark Clearinghouse (TMCH) schema. But the TMCH is based on exact string-matching, its protection scope for trademarks is limited. Based on these facts, we put forward a new solution for Chinese trademark protection in this paper. First construct three libraries of similar Chinese characters by using computer vision and machine learning techniques. Then generate similar Chinese domain names of a given brand domain name using three similar-character libraries and calculate similarity of similar domain names. Next query the WHOIS server for registration information to see whether domain names have been registered. These can support our services like Chinese IDN similarity evaluation, domain name recommendation and trademark brand counterfeiting detection, thus provide a more comprehensive protection for Chinese trademarks.

Keywords: Internationalized domain names, Chinese domain name, Similar domains, Trademark protection

1 Development of IDN

Internationalized Domain Names (IDN), also known as Multilingual Domain Names, are represented by local language characters and could contain characters with diacritical marks as required by many European languages, or characters from non-Latin scripts (for example, Arabic or Chinese). The introduction of IDNs in the Internet's address space is one of the most significant innovations in the Internet since its inception. For Internet users from non-Latin countries, it is of very important value and cultural significance to access the Internet resources through non-English characters.

1.1 Advantages of IDN

In the past 20 years, the Internet has experienced unprecedented rapid development, more and more people are using the Internet to get information, news and entertainment. Internet users also expand from the initial high level educations to the general public, which means that more and more Internet users want to use languages they are familiar with to access the Internet. So Internet Corporation for Assigned Names and Numbers (ICANN) comes up with the IDN to meet this demand. The advantages of IDN include: (1) allowing the Internet users to access the Internet using languages they are familiar with, (2) eliminating confusion caused by transliteration or free translation of domain names, (3) showing respect to the local languages and cultures, and (4) easy to use and remember by the local users.

1.2 Current Scales

After 2009, there are more and more languages and scripts joining into IDNs. From 2010 to 2014, ICANN created 38 IDN country/region code top-level domains (ccTLD) in the root zone through IDN ccTLD Fast Track Process [1], such as ".中国" (which means China), ".香港" (which means HongKong) and ".台湾" (which means Taiwan). Besides, ICANN created 35 IDN generic top-level domains in the root zone from 2013 to 2014 through New Generic Top-level Domain (New gTLD) Program (including IDN gTLDs) [1], such as ".中文网" (which means Chinese network), ".网站" (which means website), ".MOCKBA" and so on. There will be more and more top-level IDNs created in the root zone in the future, and more and more second-level or third-level IDNs will be registered.

Chinese IDNs are part of the Internationalized Domain Names. According to statistics, the top three countries (or regions) accounted for the largest percentages of Internet users are China (22%),

*Corresponding Author: Guanggang Geng; E-mail: gengguanggang@cnnic.cn

DOI: 10.3966/160792642018031902001

America (10%) and India (8%) [1]. So Chinese IDNs are of great significance and importance to Chinese Internet users. Chinese IDN is a new generation domain name that contains Chinese characters. Chinese IDNs can choose from Chinese characters, English letters (A-Z, a-z, case equivalent), digits (0-9) and hyphen (-) to form a domain name string, and must contain at least one Chinese character. What's more, Chinese IDNs technically meet the International multilingual domain name standard [2] released by IETF. The same as English domain names, Chinese IDNs are the entry point to access the Internet. It's one of the most important fundamental services of Internet, which could support applications like WWW, EMAIL, FTP and so on. There are a lot of top-level domains that are support Chinese IDN registration at the moment, for example ".COM", ".NET", ".CN", ".中国" (which means China), ".公司 (which means company), ".网络" (which means network).

1.3 Problems Caused by IDN

Currently, the support for IDN registration in top-level domains is improving, and the scale of New gTLD and IDN ccTLD are expanding rapidly as well. Compared with the domain name system that can only use 63 ASCII characters ("A-Z", "a-z", "0-9" and "-"), the new domain name system that contains IDNs is more complicated and diversity. Meanwhile, the number of similar characters in IDNs has increased tremendously, which not only makes Internet users to be more confused about some domain names, but also heavily increases the difficulty of companies to protect their intellectual property rights of trademark and brand domain names. As an example, the word "microsoft" in English and the word "microsoft" in Cyrillic look the same, but they are actually two different domains. There are mainly two cases of similar characters in IDN [3]. The first is similar characters in the same language, such as digit "0" and letter "O" in English, character "日" and character "曰" in Chinese scripts. The second is similar characters between different languages, such as letter "a" in Latin and Cyrillic language.

On the other hand, with the continuous popularity of iPad, iPhone and other smart terminal equipments, more and more new input technologies are developed by leaps and bounds. As one of the most important human-machine interfaces, speech recognition technology has developed rapidly in recent years, speech input has also become an important input method. Several companies of China like iFlyTek, Sogou, Baidu and Tencent all own self-developed technologies in speech recognition. When using speech input function, equipments may list a number of candidate domain names with the same pronunciation, for instance, when we speak "京东" (which means the B2C website Jingdong) in Chinese, Chinese words "晶东", "京东" or "京冬" may be listed. These homonym domain names may give chances to brand spoof, such as phishing attacks.

In terms of Chinese IDNs, besides characters in similar shape and Chinese homophones, there are a lot of variants in Chinese scripts. All of these together result in more and more serious problems in Chinese IDNs, such as cybersquatting and counterfeiting. According to a report [4] released by Anti Phishing Working Group (APWG), 103 IDNs were used for phishing in the second half of 2014, among which Chinese IDNs are "工商银行首页.com" (which means Industrial and Commercial Bank of China), "淘宝服务中心.cc" (which means the C2C website Taobao), "淘宝申请中心.cc" (which means the C2C website Taobao) and "淘宝退款官方网站.xyz" (which means the C2C website Taobao). With the increasing popularity of IDN, there would be more and more phishing attacks using IDNs. The target trademarks that were attacked by phishing include many Chinese famous brands, such as Taobao, Alibaba, Sina, CCTV and so on. These attacks not only threaten network security of Internet users, but also damage the reputation and images of the relevant brands.

1.4 Analysis of Current Domain Names Protection Method

In order to enable trademark holders to protect their rights during the Domain Name System (DNS) expansion, ICANN, working with Intellectual Property experts and various community stakeholders, developed the Trademark Clearinghouse (TMCH) mechanism, which aimed at protecting trademark holders' rights before disputes and providing information for trademark holders. This mechanism reduces cybersquatting risks to some extent when ICANN carrying out New gTLD Program. The Trademark Clearinghouse mechanism functions by authenticating information from rights holders and providing this information to registries and registrars [5]. Benefits of registering a trademark with the Clearinghouse include access to Sunrise registration with new gTLD registries and notification from the Clearinghouse when a domain matching your trademark has been registered [6].

TMCH mechanism uses the method of exact string matching when authenticating trademark information from rights holders, without taking similar domain names into account. Thus the TMCH is not fully effective to solve the problems of cybersquatting and phishing. For example, for the Chinese trademark "康师傅", TMCH can only provide rights holders with priority access of domain names associated with "康师傅". But for similar domain names of "康师傅" like "康帅傅", it cannot provide the same protection and notification. In terms of this aspect, the improper registration and use of domain names may greatly damage the legitimate interests of trademark holders, and what's more, give chance to Internet phishing attacks.

2 Motivations

2.1 Prevent Chinese IDN Cybersquatting

With the trend that more and more New gTLD and IDN ccTLD are created in the root zone and more and more top-level domains support IDN registration, the cybersquatting problem of famous trademark brands is getting worse. Because the registration of Domain names follows the principle of "first come first served", and has no limitation, speculators may foreseen the potential value of certain domain names and register them in advance. What's worse, some speculators may rush to register valuable domain names through cybersquatting companies. As a consequence, companies, especially with famous trademarks, are faced with increasingly tough situation of trademark protection. Thus the primary tasks of trademark protection is to prevent cybersquatting and improper use of trademark domain names, and get informed of cybersquatting and improper use timely, then take immediate action to minimize losses.

2.2 Prevent Chinese IDN Counterfeiting

The environment of Chinese IDNs is becoming mature in China. With the popularity and widely use of Chinese IDNs, there would be more and more domain counterfeiting problems. Chinese IDN counterfeiting examples may as follows (Table 1).

Table 1. Possible counterfeiting names of famous trademarks

Chinese IDN	Counterfeiting names
淘宝 (Taobao)	(a) 淘宝; (b) 掏宝; (c)裪宝
工商银行 (Industrial and Commercial Bank of China)	(a) 工茵银行; (b)工商佷行; (c)工商银伃
腾讯 (Tencent)	(a)腠讯; (b)勝讯; (c)塍讯
京东 (Jingdong Mall)	(a)京东; (b)京乐; (c)宗东

Taking the C2C website "淘宝" as an example, if someone maliciously registered a similar domain name "淘宝.com" and provided the same online shopping service as Taobao, and then the address was disseminated through spam, pseudo base-stations or social networks, it would deceive a large number of customers. These counterfeiting actions not only infringe the interests of users, but also damage reputation and images of trademark brands. Therefore, it is necessary and meaningful for us to come up with a solution to protect famous trademarks from counterfeiting.

3 Problem Solution

3.1 Solution Overview

In order to protect the legitimate rights of Chinese trademark holders, we could query WHOIS servers of top-level domains for registration information of the related Chinese IDN based on exact string match. The top-level domains that support Chinese IDN mainly include ".中国" (which means China), ".公司" (which means company), ".网络" (which means network), ".cn", ".com", ".net", ".cc", ".biz", ".co", ".business", ".sale". We can conclude from the registration information that whether trademarks have been maliciously registered in advance. We could also find out similar domain names that are quite similar to original ones by some means, and then query for registration information about these similar ones to decide whether trademarks have been counterfeited. We can also take advantage of these similar domain names to prevent Chinese IDN counterfeiting and phishing. The whole solution for trademark protection is showed in the following figure (Figure 1).

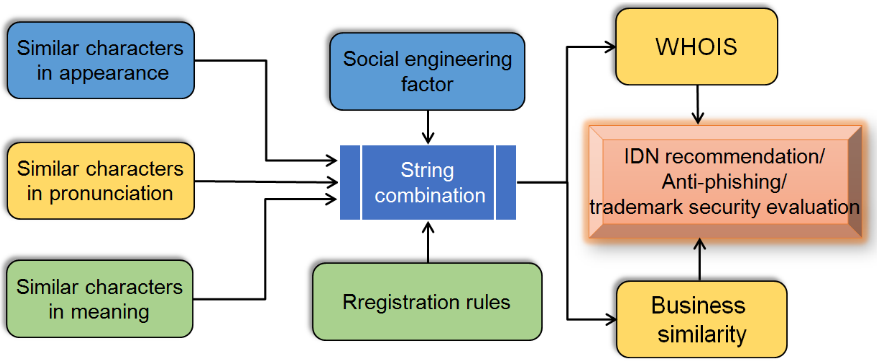

Similar characters in appearance
Similar characters in pronunciation
Similar characters in meaning
Social engineering factor
String combination
Rregistration rules
WHOIS
IDN recommendation/ Anti-phishing/ trademark security evaluation
Business similarity

Figure 1. Flow diagram of our trademark protection service

As Figure 1 shows, firstly, we construct three libraries of similar Chinese characters, which record similar characters in appearance, pronunciation and meaning separately, by using computer vision and machine learning techniques. Then we utilize the three libraries to compound similar domain names of the target Chinese IDN. We also sort the similar domain names based on similarity by using the theory of social engineering [7] (the theory we utilize here is human visual weakness when reading words). Next, we get registration information about the target domain name as well as similar domain names through WHOIS interface. For the similar domain names that have been registered we can detect the services they provide. Based on the information above, we could recommend domain names to trademark holders for the purpose of preventing cybersquatting and counterfeiting, evaluate the security level of the target Chinese IDN, and support the work of anti-phishing.

3.2 Construction of Three Similar-character Libraries

We take the comprehensiveness and general use of this solution into account when constructing three similar character libraries. There are more than 90,000 Chinese Hanzi, but not all of them are used for domain registration. Based on Maximal Starting Repertoire Version 2 [8] (MSR-2), which is released by ICANN for the development of Label Generation Rules for the Root Zone, there are 19852 Hanzi [9] can used for IDN registration. So we construct three similar-character libraries that record similar scripts of the Chinese Hanzi that are allowed to be used for Chinese IDN registration, as the foundation of the solution. The three libraries organize similar characters in appearance, pronunciation and meaning separately.

3.2.1 Construction of Library Storing Similar Appearance Characters

We combine computer graphics with machine learning methods to find out near homographs of Chinese Hanzi used for domain name registration. The method we use to determine similarity between Hanzi is described below. There are three stages in this method.

Stage 1: Calculate similarity based on visual information. For any two Chinese ideographs, we map them to corresponding representative optical-character images firstly. Then we measure similarity of the two ideographs from image aspect by using statistical methods. The similarity determination procedure is as follows.

(1) Preprocess representative optical-character images: Adjust the size, position and shape of the optical-character images with the second moment matrix normalization [10] approach and then map the optical-character image to the size of 64*64 using nonlinear function.

(2) Extract features of optical-character images: Extract 512 original features from 8 directions by using the Normalization-cooperated feature extraction method [11-12]. Then we decrease the features to 160 by means of Principal Component Analysis algorithm [13-15].

(3) Compute similarity based on features: We compute Euclidean distance between two representative optical-character images of the two ideographs. Then we normalize the Euclidean distance to the range of [0, 1] utilizing Softmax Regression [16-18]. We take the normalized distance as similarity value of the two characters herein.

Stage 2: Calculate similarity based on structure information. For any two Chinese ideographs, we map them to corresponding representative glyph structures. Then we measure similarity of the two ideographs from structure aspect by means of structure matching.

The glyph structure library is constructed by adopting character recognition method. We use semi-supervised machine learning method to detect the structures of all related characters and cut the characters into components. Finally, we get a glyph structure library of all related characters, among which every character has a glyph structure.

The structure information used in matching process includes vertical or horizontal information (up-down or left-right), components type and position of the components. The calculation of structure similarity value depends on many empirical values. In terms of vertical or horizontal structure, the similarity value of the same structure is higher. In terms of components type, characters are firstly represented by components sequence, then we compare the two sequences to get best match result by using dynamic programming algorithm. And the similarity value is given based on the number of matched components and the matched position.

Stage 3: Combine result of stage 1 and stage 2. Method based on visual information focuses on measuring the shape similarity between Chinese characters, while method based on structure information focuses on measuring the internal structure similarity between Chinese characters. When combining the two methods together, we get a more effective way to measure the similarity between characters.

The difficulty in combining two methods is that the metric space is different when calculating similarity values, although the similarity values of the two methods both fall into [0, 1]. So the similarity values of the two approaches are not comparable.

The combination strategy we used here is described below. For each character, we firstly get the top n semblable characters of each approach, then divide the interval [0, 1] into a certain number of small intervals. We assign different weights to image similarity values that fall into different intervals and transform the weights into metric space which is comparable with structure similarity. We use the maximum voting to merge the results of two approaches. Finally, we sort the similarity values of these similar characters to get a better result. The character with the largest similarity value is the most similar character of the specified character. The following table shows some examples of similar characters found by the above method. Each similar character is followed by a similarity value in the parenthesis.

Table 2. Examples of similar characters

Chin-ese	Similar Chinese characters			
chara-cters	char1	char2	char3	char4
淘	掏(0.8033)	裪(0.7867)	啕(0.6400)	渆(0.6000)
宝	宔(0.8183)	宅(0.6667)	宅(0.6600)	壶(0.4373)
京	京(0.9900)	宗(0.8033)	東(0.7133)	克(0.6333)
东	乐(0.8000)	尔(0.5729)	牙(0.5559)	--

From the table we can conclude that the method we use is an effective way to find similar characters of Chinese Hanzi. By using this method, we construct the library organize similar appearance characters.

3.2.2 Construction of Library Storing Same Pronunciation Characters

We can easily get all homophones of a Chinese character according to Chinese dictionary. Take Hanzi “京” as an example, “京” may have a lot of homophones, such as “京”, “惊”, “晶”, “经”, “精”, and so on, though some of which are look totally different from “京”. Hence, we don’t add all homophones into the library. We adopt glyph similarity determination methods to compute the similarity between character and its homophones. Only homophones whose similarity values are greater than threshold are added to the library. Wherein the threshold require manual selection and testing.

3.2.3 Construction of Library Storing Variant Characters

Chinese (character) variants are characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts [19]. In the Chinese language, there are two types of variant [19]. One is created by Simplified Chinese and Traditional Chinese, such “宝” in Simplified Chinese and “寶” in Traditional Chinese, they are variants to each other. Another is the generic variant. Generic variants are slightly different visually, but are treated the same and have universal interchangeability, such as “户/戸” and “黄/黃”. Chinese Variants are common in Han script and are an important source of similar characters. We adopt the variants table [20] for domain registration, released by Chinese Domain Name Consortium (CDNC), to construct variants library. The table covers all Chinese characters and its variant(s) used in the Chinese IDN registration.

3.3 The Compound of Similar Domain Strings

For a domain name, we can take advantage of three libraries to compound similar domain names. Firstly, for each character included in a certain Chinese IDN, we select its similar characters (including similar appearance characters, same pronunciation characters and variants) from three libraries. Next, according to Chinese IDN registration rules, we compound similar domain names by permutating and combining similar characters. Then we sort the similar domain strings based on overall similarity. In the sorting process, we also take advantage of a social engineering principle.

3.3.1 Select Similar Characters and Compute Similarity Against Original Ones

For each Hanzi in a certain domain name, we select its similar appearance characters, same pronunciation characters and variants from three libraries. The three types of similar characters may be crossed, we can show them in the below figure (Figure 2).

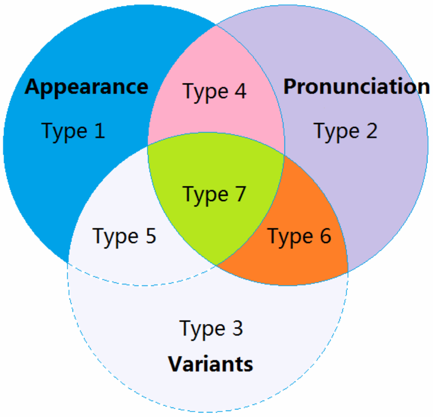


Figure 2. The relation of three types of similar characters

We can see from the figure that similar characters of a Hanzi may be included in the following types of subsets: (1) only similar in appearance, (2) only with same pronunciation, (3)only variant, (4) both similar in appearance and same in pronunciation, (5) similar in appearance and variant, (6) with same pronunciation and variant, (7) similar in appearance, same in pronunciation and variant. Because variants must have same pronunciation, while characters with same pronunciation are not necessarily variants, so variants set is a subset of characters with same pronunciation. Therefore the numbers of elements of the third and fifth type are both 0, that is to say these two types of subset don’t exist. There exist 5 types of subsets at last.

For similar characters included in different types of subsets, we calculate similarity values against original ones by using the following algorithm (Figure 3).

```
/*Algorithm Description: Calculate similarity values of similar characters against original one, return a Map set
 of similar characters and corresponding similarity value.
   */
Map<CharType, double>  cal_Similarity(Map< CharType, double>  shapeSimSet,
                                 Set<CharType>  spellSameSet,
                                 Set<CharType>  variantSet)
  {
    Map<CharType, double>  simCharUnion;
    double  similarity = 0.0;
    foreach(Map.Entry<CharType, double>  shapeSimChar  in  shapeSimSet)
    {
      if(!spellSameSet.contains(shapeSimChar.getKey())&&
       !variantSet.contains(shapeSimChar.getKey()))
    {
        similarity = shapeSimChar.getValue();
        }
        else if(spellSimSet.contains(shapeSimChar.getKey())&&
              !variantSet.contains(shapeSimChar.getKey()))
    {
        similarity = min(max(shapeSimChar.getValue(), PARAM_Y_1) * K, 1.0);
        }
        else if(spellSameSet.contains(shapeSimChar.getKey())&&
              variantSet.contains(shapeSimChar.getKey()))
      {
        similarity = 1.0;
        }
        if(!simCharUnion.containsKey(shapeSimChar.getKey()))
        {
          simCharUnion.put(shapeSimChar.getKey(), similarity);
        }
    }
    foreach(CharType  spellSameChar  in  spellSameSet)
    {
      if(!shapeSimSet.containsKey(spellSameChar)&&
         !variantSet.contains(spellSameChar))
    {
          similarity = PARAM_Y_1;
        }
        else if(!shapeSimSet.containsKey(spellSameChar)&&
              variantSet.contains(spellSameChar))
    {
          similarity = min(max(PARAM_Y_1, PARAM_Y_2) * K, 1.0);
        }
        if(!simCharUnion.containsKey(shapeSimChar.getKey()))
        {
          simCharUnion.put(shapeSimChar.getKey(), similarity);
        }
    }
  }
```

Figure 3. Algorithm for calculating similarity value

The above algorithm has three parameters, which are (1) shapeSimMap, a Map set of similar appearance characters, saving similar characters and corresponding similarity values, (2) spellSimSet, a Set of characters with same pronunciation, and (3) variantSet, a Set of variants. The algorithm would return a Map set of similar characters and corresponding similarity value at last. We calculate the similarity value of a similar character based on which subset it belongs to. If the character falls into the 1st subset type, we take its appearance similarity as comprehensive similarity directly. If the character belongs to the 2nd subset type, we take PARAM_Y_1 as its similarity value. If the character falls into the 4th subset type, we get its comprehensive similarity using the given formula, among which PARAM_Y_1, PARAM_Y_2 and PARAM_K are predefined constant. If the character belongs to the 6th subset type, we get its comprehensive similarity using the given formula as well. If the character falls into the 7th subset type, it shows that the character is extremely similar against original one, so we assign 1.0 directly to comprehensive similarity. For similar characters and their corresponding similarity values, we save them into the Map Set simCharUnion.

There are three predefined constants in the algorithm, i.e. PARAM_Y_1, PARAM_Y_2 and K. In analogy to shape similarity value, PARAM_Y_1 and PARAM_Y_2 can be seen as similarity values in pronunciation and meaning separately. The range of PARAM_Y_1 is [0.8, 0.9] and we can select a suitable value according to application scenarios, for instance, we can give PARAM_Y_1 a bit larger value in speech-input scenario. As for PARAM_Y_2, it can take value from [0.9, 1.0). K is a number greater than 1 and it can take value from (1, 1+σ), wherein σ is a very small decimal greater than 0 to make sure similarity value lifted a bit.

3.3.2 Construct Similar Domain Name Strings and Calculate String Similarity

When constructing similar Chinese IDN, we firstly get all similar characters of every script in Chinese IDN, then compound similar Chinese IDNs by permutating and combining similar characters. For these similar Chinese IDNs, we compute their similarity values against original one by adopting Bayesian decision method [13].

The following describes the calculation of the overall similarity of similar Chinese IDNs against original one. Supposing A and B are domains both contains N scripts and similarity values between N counterparts are S1, S2, S3, …, Sn, we calculate similarity value between A and B using the following formula [3]:

$$Similarity(A,B)=\frac{\prod_1^n S_n}{\prod_1^n S_n+\prod_1^n (T-S_n)}$$

In the formula, T is a Minimal number greater than 1 and its function is to avoid becoming 0 when Sn=1(n=1, 2, 3, …, n), i.e. the counterpart exactly the same.

We can measure the similarity between two Chinese IDNs with value between [0, 1] by using the formula above. For example “工商银行” and “工茵银行”, the similarity values of “工”, “银” and “行” in the corresponding position are 1.0, while the similarity value between “商” and “茵” is 0.8733. Given T=1.01, Similarity (工商银行, 工茵银行) =

$$\frac{1\times 0.8733\times 1\times 1}{1\times 0.8733\times 1\times 1+(1.01\text{-}1)\times(1.01\text{-}0.8733)\times(1.01\text{-}1)\times(1.01-1)}$$

Because of human visual weakness, we tend to take a similar Chinese IDN, which is only different in one script from original one, as target IDN. Based on this social engineering principle, we increased similarity values of IDNs that is only different in one script from original one. For example, the Similarity (工商银行, 工茵银行) is 0.999999 by using formula above. There is only one script different between “工商银行” and “工茵银行”, so we increase Similarity (工商银行, 工茵银行) to 1.999999 by adding 1 to Similarity (工商银行, 工茵银行). We also sorted similar IDNs based on similarity values.

3.4 Query for WHOIS Information

For constructed similar Chinese IDNs and original Chinese IDN, we obtain their registration information by taking advantage of WHOIS service. We build a web application in order to show related information to users. We used ThinkPHP at the front-end and Java at the back-end in order to improve the efficiency of information inquiry. The workflow of information inquiry is shown as follow (Figure 4).

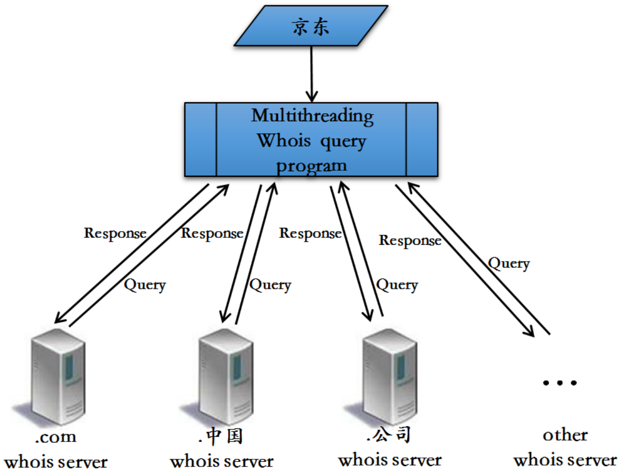


Figure 4. Workflow of information inquiry

For a second-level domain name, the program could send queries to related Whois servers for registration information of this domain and get responses from Whois servers. We take Chinese IDN “京东” as an example in the figure. If the domain name has been registered, related servers would return registration information about this domain. Registration information returned includes registrant name, registrant organization, registrant phone, registrant email, registrar, domain status and so on. Trademark holders could precaution cybersquatting and counterfeiting based on registration information. For example, trademark holders could register brandable domains timely and pay close attention to domains that have been preregistered and business they are offering. What’s more, trademark holders could check that whether the similar domains are phishing websites.

In order to improve the query efficiency, the Whois query program uses a combination of serial and parallel manner. Because the whois servers of top-level domains is not all the same, we can use multithreading to query each top-level domain’s whois server separately. While for domains that have same server, we use serial program to query the server.

4 Effectiveness Evaluation

4.1 Registration Status of Brand Chinese IDNs

We can get registration status of a certain domain by using Whois query program. Take brand “京东” as an example, the following table shows its registration status under some key TLDs.

From Table 3 we can see that the domain “京东 has been registered under related TLDs except “.co”. But the holders of this second-level domain under “.中国”, “.cn”, “.business” and “.sale” are not “京东”. Therefore, “京东” should keep an eye on those four domains which are not owned by “京东” to avoid behaviors that may damage the company's reputation through web phishing. As for brandable domain “京东商城” (which means Jingdong mall), its registration status is shown in Table 4.

Table 3. Registration status of “京东” under some key TLDs

Domain	Status	Holder	Web Service	Mailbox service
京东.中国	registered	罗贤	no	no
京东.公司	registered	京东	--	--
京东.网络	registered	京东	--	--
京东.cn	registered	罗贤	no	no
京东.com	registered	京东	--	--
京东.net	registered	京东	--	--
京东.cc	registered	京东	--	--
京东.biz	registered	京东	--	--
京东.co	unregister-ed	--	--	--
京东.business	registered	101doma-in, Inc.	yes	yes
京东.sale	registered	101doma-in, Inc.	yes	yes

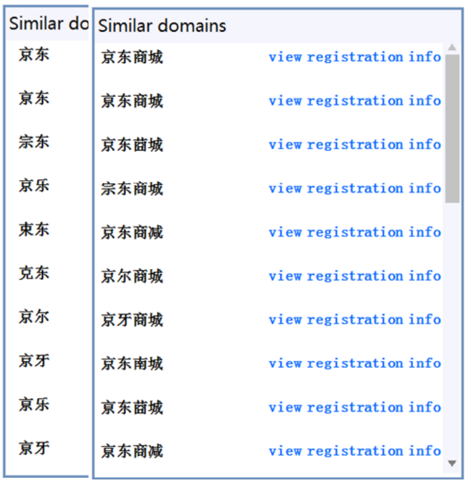
Table 4. Registration status of “京东商城” under some key TLDs

Domain	Status	Holder	Web Service	Mailbox service
京东商城.中国	registered	方学董	yes	no
京东商城.公司	registered	京东	--	--
京东商城.网络	registered	京东	--	--
京东商城.cn	registered	方学董	no	no
京东商城.com	registered	匿名	no	no
京东商城.net	registered	Wu Yang	no	no
京东商城.cc	registered	Fei Jian xin	no	yes
京东商城.biz	registered	Wang Xun	yes	no
京东商城.co	unregistered	--	--	--
京东商城.business	unregistered	--	--	--
京东商城.sale	unregistered	--	--	--

The company of “京东” is faced with more severe problems in protecting Chinese IDN “京东商城”. As shown in Table 4, “京东” only owns this Chinese IDN under “.公司” and “.网络”. Hence “京东” should care about those domains that are not owned by itself.

4.2 Registration Status of Similar Chinese IDNs

We can take advantage of three similar-character libraries to get similar Chinese IDNs of “京东” and “京东商城”, which are shown in Figure 5.



Similar do
京东
京东
宗东
京乐
東东
克东
京尔
京牙
京乐
京牙

Similar domains	
京东商城	view registration info
京东商城	view registration info
京东蔺城	view registration info
宗东商城	view registration info
京东商减	view registration info
京尔商城	view registration info
京牙商城	view registration info
京东南城	view registration info
京东蔺城	view registration info
京东商减	view registration info

Figure 5. Similar Chinese IDNs of “京东” & “京东商城”

From Figure 5 we can see that, the similar domains constructed based on three similar-character libraries are very similar to original ones indeed. If these domains were maliciously registered and used in illegal ways by speculators, it is very likely that the brand image of “京东” would be damaged and the property security of users would be threatened also. We queried for registration status of “京东” and “京乐” (two similar domains of “京东”) under some key TLDs, which are shown in Table 5 and Table 6 below.

Table 5. Registration status of “京东” under some key TLDs

Domain	Status	Holder	Web Service	Mailbox service
京东.中国	registered	罗贤	no	no
京东.公司	registered	京东	--	--
京东.网络	registered	京东	--	--
京东.cn	registered	罗贤	no	no
京东.com	registered	Zhang Zaifa	no	no
京东.net	unregistered	--	--	--
京东.cc	unregistered	--	--	--
京东.biz	unregistered	--	--	--
京东.co	unregistered	--	--	--
京东.business	unregistered	--	--	--
京东.sale	unregistered	--	--	--

Table 6. Registration status of “京乐” under some key TLDs

Domain	Status	Holder	Web Service	Mailbox service
京乐.中国	unregistered	--	--	--
京乐.公司	unregistered	--	--	--
京乐.网络	unregistered	--	--	--
京乐.cn	registered	肖敏	no	no
京乐.com	registered	anonymous	no	no
京乐.net	registered	anonymous	no	no
京乐.cc	unregistered	--	--	--
京乐.biz	unregistered	--	--	--
京乐.co	unregistered	--	--	--
京乐.business	registered	101domain, Inc.	yes	yes
京乐.sale	registered	101domain, Inc.	yes	yes

From the Table 5 and Table 6, we can see that except for “京东.公司” and “京东.网络”, the domain “京东” under other TLDs were either preregistered by others or unregistered. Hence “京东” should register related domains timely and pay attention to related domains that have been preregistered.

From the above results, we can conclude that our solution for trademark protection offers registration status not only about brand related domains but also similar domains of trademarks. Compared with TMCH mechanism developed by ICANN, our solution provides more comprehensive protection for Chinese IDNs as well as defends Internet users against web phishing.

5 Conclusion

This paper introduced IDN and described the problems and challenges caused by IDN firstly, then analyzed the deficiency of TMCH. Aimed at providing more comprehensive protection for Chinese IDNs, we studied on this and put forward our solution. We constructed three similar-character libraries by using computer vision and machine learning methods. Then we generated similar Chinese IDNs of brand domain names by taking advantage of three libraries above. Next we got registration information of related domain names through extended Whois query program. The registration information could support services like Chinese IDN similarity evaluation, domain name recommendation and trademark brand counterfeiting detection. In the future, we could construct similar domain names by adding interference characters, for example “淘-宝.com”, “工商银行首页.com”, “淘宝服务中心.cc” and so on.

Acknowledgement

This paper is supported by grants from the National Natural Science Foundation of China (Nos. 61375039 and 61272433), and supported by National Key R&D Program of China (No. 2016YFC0801100).

References

[1] ICANN, *Access Domain Names in Your Language*, 2015. https://www.icann.org/sites/default/files/assets/idn-access-domain-names-03sep15-en.pdf

[2] P. Faltstrom, RFC 3490 International Domain Names in Applications (IDNA), 2003.

[3] B. Hong, G. Geng, L. Wang, W. Mao, Wei, A Method to Detect Chinese Domain Name Homograph Attack, *Application Research of Computers*, Vol. 30, No. 11, pp. 3426-3429, November, 2013.

[4] APWG, *Global Phishing Survey: Domain Name Use and Trends in 2H2014*, https://apwg.org/apwg-news-center/

[5] ICANN, *Trademark Clearinghouse (TMCH)*, http://newgtlds.icann.org/en/about/trademark-clearinghouse

[6] ICANN, *Trademark Clearinghouse Rights Protection Mechanism Requirements*, http://newgtlds.icann.org/en/about/trademark-clearinghouse/rpm-requirements-30sep13-en.pdf

[7] F. Mouton, M. M. Malan, K. K. Kimppa, H. S. Venter, Necessity for Ethics in Social Engineering Research, *Computers & Security*, Vol. 55, pp. 114-127, November, 2015.

[8] ICANN, *Maximal Starting Repertoire Version 2 (MSR-2) for the Development of Label Generation Rules for the Root Zone*, https://www.icann.org/news/announcement-2-2015-04-27-en

[9] ICANN, *Maximal Starting Repertoire - MSR-2-Overview and Rationale*, https://www.icann.org/en/system/files/files/ msr-2-overview-14apr15-en.pdf

[10] T. Lindeberg, Image Matching Using Generalized Scale-Space Interest Points, *Journal of Mathematical Imaging and Vision*, Vol. 52, No. 1, pp. 355-367, October, 2014.

[11] C. Liu, N. Kazuki, S. Hiroshi, F. Hiromichi, Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques, *Pattern Recognition*, Vol. 37, No. 2, pp. 265-279, February, 2004.

[12] C.-L. Liu, Normalization-Cooperated Gradient Feature Extraction for Handwritten Character Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 8, pp. 1465-1469, September, 2007.

[13] X. Zhang, *Pattern Recognition*, Tsinghua University Press, 2010.

[14] M. Yamamoto, K. Hayashi, Clustering of Multivariate Binary Data with Dimension Reduction via L1-regularized Likelihood Maximization, *Pattern Recognition*, Vol. 48, No. 12, pp. 3959-3968, May, 2015.

[15] L. Kozma, A. Ilin, T. Raiko, Binary Principal Component Analysis in the Netflix Collaborative Filtering Task, *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, 2009, pp. 1143-1147.

[16] Computer Science Departmentof Stanford University, *Softmax Regression*, http://deeplearning.stanford.edu/wiki/index.php/Softmax_Regression

[17] Y. Zheng, R. S. Zemel, Y.-J. Zhang, H. Larochelle, A Neural Autoregressive Approach to Attention-based Recognition, *International Journal of Computer Vision*, Vol. 113, No. 1, pp. 67-79, May, 2015.

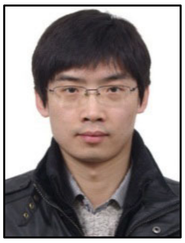
[18] M. Geist, Soft-max Boosting, *Machine Learning*, Vol. 100, No. 2, pp. 305-332, September, 2015.

[19] X. Lee, J. Yee, C. Dillon, *Report on Chinese Variants in Internationalized Top-Level Domains*, http://archive.icann.org/en/topics/new-gtlds/chinese-vip-issues-report-03 oct11-en.pdf

[20] Chinese Domain Name Consortium, *Merged Character Table* (utf8 encoded), http://www.cdnc.org/gb/research/file/CDNC_utf8.txt

Biographies

Junling Zhang is currently a master's candidate of Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. Her research interests include information retrieval on the Web, Web data analysis, and domain names.



Zhiwei Yan received his Ph.D. degree from Beijing Jiaotong University. He joined China Internet Network Information Center in 2011 and is currently an Associate Professor of Chinese Academy of Sciences. Since April 2013, he has been an Invited Researcher of Waseda University. His research interests include mobility management, network security, and next generation Internet.

Guanggang Geng received the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences. He joined the Computer Network Information Center, Chinese Academy of Sciences. He is currently an Professor in the China Internet Network Information Center. His current research interests include machine learning, adversarial information retrieval on the Web, and Web search.

Xiaodong Lee received his Ph.D. of Computer Architecture in the Institute of Computing Technology of Chinese Academy of Sciences (CAS) in 2004. He organized and accomplished several international and domestic technology standards in the fields of domain name and email, as well as the research and development of the first series of software and hardware system of domain name service in China.

Jong-Hyouk Lee (M'07-SM'12) carried the M.S. and Ph.D. work in Computer Engineering at Sungkyunkwan University, Korea (M.S., 2007; Ph.D., 2010). In 2009, he joined the project team IMARA at INRIA, where he undertook the protocol design and implementation for IPv6 vehicular (ITS) communication and security.