

Regular Expression Patterns for Searching Trust Attributes in e-Commerce Website

MUHAMMAD RUSHDI RUSLI¹, AB. RAZAK CHE-HUSSIN²,
HALINA MOHAMED DAHLAN³

Department of Information Systems
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia 81310 Skudai, Johor, Malaysia.
¹mrushdi84@gmail.com, ²abrazak@utm.my, ³halina@utm.my

Abstract – E-Commerce websites must provide trust to attract consumers. There are five most important trust attributes that should be placed in an e-Commerce website. However, consumers may not know about the attributes and it takes time for users to search them especially for a beginner computer user. Besides, there are no tools available to search these five attributes to establish confidence between e-Commerce parties. Tools that can search the attributes should be developed to assist customers assess the trustworthiness of an e-Commerce website. These trust attributes in e-Commerce websites have been identified and located usually in ‘Homepage’ and ‘Contact Us’ pages. Since the trust attributes are usually placed in unstructured text, information extraction is used to extract the data. Data from e-Commerce websites in United Kingdom (UK), United States (US) and Malaysia have been collected to create patterns using regular expression. These patterns are applied in prototype that has been built according to the proposed search algorithm. Finally, the prototype and the patterns are tested against the e-Commerce websites from UK, US and Malaysia. Based on the results, the patterns show that they can be used in the proposed tool to search the trust attributes in order to assist the consumer to place trust in an e-Commerce website.

Key-Words – Information Extraction, Regular Expression, e-Commerce, Trust

1 Introduction

Recent statistics from the Census Bureau of the Department of Commerce [1] indicate that total e-Commerce sales for 2008 were estimated around USD 133.6 billion, which is a 4.6 percent increase over the sales for 2007. However, an e-Commerce transaction is different from traditional commerce whereby customers cannot see the seller’s face or physically examine the products that are marketed through e-Commerce websites. To gain the customer's trust is an important task for the seller in order to secure an e-Commerce transaction. As a result, many people have focused upon trust issues in e-Commerce including researchers. Some of them have conducted studies on factors that influence consumer trust of e-Commerce websites. These factors are defined by the researchers with various definitions like trust influencers [2], antecedents of trust [3] and trust attributes [4] and each of them would have their own point of view.

Hussin et al. [4] have defined the trust factor with definition of trust attributes which are more firmer than from other studies. The trust attributes will be mentioned in the next section. The focus of this study is how the trust attributes will be represented in regular expression patterns.

This paper is organized as follows. In section 2, we describe the five trust attributes that exist in e-Commerce websites. Followed with the location of the trust attributes in e-Commerce. Section 3 describes the trust attributes patterns and how to write patterns using regular expression. The patterns of data are gathered from e-Commerce websites within the UK, US and Malaysia. Once the patterns are created, search algorithm is proposed and applied in the system to search the five trust attributes. Furthermore, the system is tested on the e-Commerce websites from UK, US and Malaysia and comes out with the comparison between a manual search and system search for the attributes. Based on the results, the patterns can be used to search the trust attributes which will somehow help the customer to gain trust towards e-Commerce website. The paper ends with the conclusion and suggestion for future work.

2 Existence of Trust Attributes in E-Commerce Website

According to the online questionnaires administered by Che-Hussin et al. [4], there are five trust attributes that should be placed on the first page of

e-Commerce website to gain the trust of consumers. Table 1 shows the five trust attributes.

Table 1: Top five trust attributes adapted from Che-Hussin et al. [4]

Rank	Trust attributes
1.	Company telephone number (CTN)
2.	Company email address (CEA)
3.	Privacy Policy (PP)
4.	Company address (CA)
5.	Third party for secure transaction (e.g. VeriSign) (TPST)

Based on the attributes in Table 1, company telephone number and company email address are the most important attributes that will influence trust of customers towards an e-Commerce website. These trust attributes are followed by the privacy policy which states the policy of the vendor regarding their products or services, the company address and third party for secure transaction. The last trust attribute is to guarantee that their personal information is safe and secure.

There are only five trust attributes that have been considered in this study since these attributes are the most important information for the consumer [4]. As we know, there many pages in an e-Commerce website and the trust attributes only exist in several of them. The search time for the system to find the trust attributes can be reduced by searching in the appropriate page only. The next section will explain how the locations are identified and what the common pages of the merchants who placed the trust attributes.

2.1 Location of Five Trust Attributes

Since there is no standard way to present the trust attributes in the e-Commerce websites, there are possibilities that these trust attributes are located in the different places. A survey was conducted to find the common place of trust attributes in e-Commerce website. The observation has been done on 40 UK, 40 US and 40 Malaysian e-Commerce websites and Table 2 shows the most location of trust attributes are placed and keyword for PP attribute in e-Commerce websites.

Table 2: Trust Attributes Locations.

Trust Attributes	Location	PP Keyword Used
CTN	Contact Us	Privacy Policy
CEA	Contact Us	
PP	Homepage / Contact Us	Privacy & Security
CA	Contact Us	Privacy Statement
TPST	Homepage / Contact Us	Privacy Notice

According to the table, most e-Commerce websites from UK, US and Malaysia place their trust attributes in Homepage and Contact Us. Besides, e-Commerce websites always put their privacy policy under “Privacy Policy”, “Privacy & Security”, “Privacy Notice” or “Privacy”.

3 Converting Trust Attributes to Regular Expression Patterns

Even though there are five trust attributes, only four patterns are created for CTN, CEA, CA and TPST. This study only searched links for PP attributes to identify whether an e-Commerce website contained this attribute. The next section will explain how the patterns are created for each trust attribute. Before patterns are created using regular expression, the pattern of the desired data should be recognized first. In order to recognize the pattern, a survey was conducted on several e-Commerce website. The survey was conducted on 40 UK, 40 US and 40 Malaysian e-Commerce websites.

Since most of the e-Commerce websites are in unstructured format, these trust attributes are also written and placed in this format. One of the solutions to search these attributes in unstructured text is by using regular expression.

3.1 Company Telephone Number (CTN)

In Table 3, number 1 to 4 regular expressions are used to extract company telephone numbers from United Kingdom, United States and Malaysia.

Table 3: Regular expression patterns to CTN

No.	Country	Regular Expression
1.	UK	Pattern (1a): $\backslash(?\d{5})?[-\s.](?\d{6})?$ Example: i. (01422) 330008 Pattern (1b): $\backslash+?\d{2}[-\s.](?\d{1})?[-\s.]?\d{0,3}[-\s.]?\d{3,4}[-\s.]?\d{3,6}$ Example: i. +44 (0) 28 9068 1015
2.	US	$\backslash(?\d{3,4})?[-\s.]?\d{3}[-\s.]?\d{4}$ Example: (800) 564-5740
3.	UK & US	$\backslash\{1\}[-\s.]?+?(?\d{2,3})?[-\s.]?\d{3,4}[-\s.]?\d{3,4}[-\s.]?\d{0,4}$ Example: i. 1-888-237-8289, 650.560.6500
4.	Malaysia	Pattern (4a): $\backslash+?\d{3,4}[-\s.]?\d{7}$ Example: i. +6016-6242492 Pattern (4b): $\backslash+?(?\d{2,4})?[-\s.]?\d{4}[-\s.]?\d{3,4}$ Example: i. +603-8922 1513

3.1.1 CTN patterns for UK

Pattern (1a) in Table 3 is used to find UK company telephone numbers. In regular expression, backslash character '\ ' is used to match a special character. For example, in order to match the telephone numbers which starts with five digits and followed by whitespace and 6 digits. Some of them write the first five digits between left and right parenthesis. Since character '(' and ')' are categorized as special characters, the character '\' should be placed first. The character '?' is to show that the character before it can exist or not because to match both strings that have character '(' and ')' or not. The character '\d{5}' is used to match digit characters as much as five characters. The pattern '\(?:\d{5}\)?' can be used to match the first string of telephone numbers such as (01422) or 08700. The character '[-s.]' is used to match character '-', '.' or whitespace. And the pattern '\(?:\d{6}\)?' is the same as the previous which is used to match digit characters as much as six characters.

Pattern (1b) in Table 3 is used to search for UK company telephone numbers. In order to match character '+44' or '44', pattern '\+?d{2}' is used. Followed with '[-s.]' pattern to match character '-', '.' or whitespace. As in example 2, there are also data that is without whitespace in '(0)20' which is different with '(0) 28', the pattern '[-s.]' should include the '?' character. The pattern 'd{0,3}' is to match digit number which is not more than 3 digit and can be 0 digit. This is used for the string in example 3 where the arrangement of the digit is different from the first two examples. According to the example 1 and 2, the arrangement of the digit is 2 digits, 4 digits and followed by 4 digits. However, in the last example, the arrangement of the number is 4 digits followed by 6 digits.

3.1.2 CTN patterns for US

The pattern in the US category in Table 3 is used to find the company telephone numbers from US. Pattern '\(?:\d{3,4}\)?' is to match the first three digits '(800)' and '[-s.]\d{3}' is to match the whitespace and three digits in the middle of the string. Pattern '[-s.]\d{4}' is to match the hyphen character (-) and the last four digits.

3.1.3 CTN patterns for both US and UK

The pattern in the US and UK categories in Table 3 is created to match telephone numbers from UK and US since the pattern of the arrangement of number is similar. The pattern is easier to create when the pattern of the string is similar with the other string which is desired to match. Pattern '\d{1}?[-s.]\+?\(?:\d{2,3}\)?' is to match the string '1-', '650.',

or '+44'. The rest of the pattern is to match the number which follows with three or four digits such as '888-237-8289', '560.6500', '844 844 0809' or '1829 771 886'. According to the example, the vendors place their telephone numbers in various formats. Some of them use character hyphen '-', periods '.' and whitespace (' ').

3.1.4 CTN patterns for Malaysia

There are only two types of telephone numbers in Malaysia which are mobile phone and fixed line. Pattern (4a) for Malaysia category in Table 3 is used to match mobile phone numbers where several e-Commerce websites in Malaysia also includes number '60' as the country code. The number will follow with two digits, most of them are service provider codes, and followed with seven digits.

Meanwhile, pattern (4b) is used to match fixed line telephone numbers in Malaysia where the numbers are placed in various ways such as inclusive of country code and the plus '+' character, and some of them put the country code in parenthesis '(')'.

3.2 Company Email Address (CEA)

This trust attribute follows a standard pattern for the e-Commerce companies from UK, US and Malaysia. Table 4 shows the regular expression to extract email address.

Table 4: Regular expression patterns to extract company email address (applies for all)

Regular Expression
/^[A-z0-9\-_]+\(\.[A-z0-9\-_]+\)*@(((A-z0-9)\+ \-?[A-z0-9]+\)+\.)+[A-z]{2,6}\$/

This pattern is used to match the email pattern. The first pattern '/^[A-z0-9\-_]+\(\.[A-z0-9\-_]+\)*@' is to find local name which may only contains letters, digits, hyphen '-', underscore '_' and periods '.'. The local name may not begin and/or end with a period and not contain two or more subsequent periods '..'. Pattern '((A-z0-9)\+|\-?[A-z0-9]+\)+\.)' is use to match sub domain and domain of the email where the domain name and sub-domain name may not begin and/or end with a hyphen, must be between 2 to 63 characters long, and may only contain letters, digits and hyphen. Besides, top-level domains may only contain letters and must be between 2 to 6 characters long and the sub-domain names, the domain name and the top-level domain names are separated by single periods '.'.

3.3 Privacy Policy (PP)

According to Table 2, e-Commerce websites from UK, US and Malaysia put their privacy policies under “privacy” section. The method used by authors to identify whether e-Commerce website from UK, US and Malaysia include a privacy policy was to find the title of each link in the homepage that contains the words "privacy". Since this method was only to find these words for the title link, deriving regular expressions pattern for this attribute is not necessary.

3.4 Company Address (CA)

Regular expression in Table 5 can only be used to extract company address from US and UK.

Table 5: Regular expression patterns to extract company address for US and UK

Country	Regular Expression
US	/[A-Z]{2},?[\s][]{0,}\d{5}/
UK	/[A-Z]{1,2}[0-9]{1,2}[\s][0-9]{1}[A-Z]{1,2}/

3.4.1. CA patterns for US

Based on the observation of US e-Commerce websites, most of them write their address with state code followed by with the zip code. The state code usually is in capital letters with two letters and the zip code in five digits. Pattern '[A-Z]{2}' is to match the first two letters of the state code. Pattern ',?[\s][
]{0,}' for US category in Table 5 is to match the character that separates the state code and zip code which is usually a whitespace ' ', html code for new line '
' and comma ','. Pattern '\d{5}' is used to match the zip code.

3.4.2. CA patterns for UK

UK address format is more specific than US address format where the codes are divided into postcode area, postcode district, postcode sector and postcode unit. The first letter or pair of letters represents the postcode area. The following number, from 0 to 99, determines the postcode district within that area. It follows with the whitespace ' '. After that the first character after the space is a digit from 0 to 9 which determines the postcode sector. The final two letters is for the postcode unit. Table 5 in UK category shows the regular expression to extract address from UK website.

3.4.3. CA patterns for Malaysia

Since Malaysian addresses do not have any state code, to extract the address is a difficult task. To overcome this problem, this study stored information that includes city and the state from the

country into the database. The system will search state name then search the city of the state in 60 characters before the state’s name is found.

Besides that, the system will also search for the postcode if there is no city for the state found in 60 characters before the state name is found. Postcode also is used to determine whether the state’s name is found in address or not.

3.5 Third Party for Secure Transaction (TPST)

Regular expression in Table 6 shows regular expression patterns to extract eight third party secure transaction information from an e-Commerce websites. This study only focuses on these eight third party since they are commonly used by the e-Commerce company. According to the table, three of them constitute more than one regular expression since the pattern of the data is different from the others.

In order to identify whether an e-Commerce website does have third party secure transaction, each pattern of the link for third party secure transaction should be identified first. To obtain the pattern, an observation was conducted on 40 e-Commerce websites from United States, 40 websites from United Kingdom and 40 websites from Malaysia that have this service. Some of the websites are shown in Table 12.

Table 6: Regular expression patterns to extract third party secure transaction information.

Name	Pattern
GEOTRUST	/\Vsmarticon.geotrust.com\Vs\i\js/i
VERISIGN	a. /https:\V\seal.verisign.com+[a-zA-z0-9=?\;_&.]*/i b. /https:\V\servicecenter.verisign.com\vcgi-bin\Xquery.exe\?/i
SCAN ALERT	a. /https:\V\www.scanalert.com+[a-zA-z0-9=?\;_&.]*/i b. /https:\V\www.mcafeesecure.com\VRatingVerify\?ref="+\$.Surlsecure."/i
GLOBALSIGN	/https:\V\secure.globalsign.net\en\find\sealct.cf m\?id=\d+/i
TRUSTWAVE	/https:\V\sealserver.trustkeeper.net\compliance\cert.php?\code=w+/i
WebSafeShield	/http:\V\seals.websafeshield.com\w+\websafeshield\js/i
Entrust	/https:\V\seal.entrust.net\seal\js?\domain="\$.Surlsecure."/i
Thawte	a. /https:\V\siteal.thawte.com\vcgi\server\thawte_seal_generator.exe/i b. /https:\V\siteal.thawte.com\vcgi\server\certdetails.exe?\code=w+/i c. /https:\V\www.thawte.com\vcgi\server\certdetails.exe?\referer=/i

According to the observation, some of the links for the secure transaction are not fixed to only one pattern but have more than one such as VERISIGN, sCAN ALERT, and Thawte. Therefore, the regular expressions patterns for these parties constitute more than one regular expression pattern. These patterns are created based on their link code that is used by the third party of secure transaction and utilized by their clients.

4.0 Search Algorithm

The steps to search five trust attributes in an e-Commerce Website that has been proposed in the prototype are as follow:

1. Convert the inserted URL into source code.
2. If success
 - 2.1. Check link’s title contains word “privacy”
 - 2.2. Check whether the homepage contains CTN, CEA, CA and TPST.
 - 2.3. Found CTN, CEA, CA and TPST?
 - 2.4.1. True
 - 2.4.1.1 Go to end
 - 2.4.2. False
 - 2.4.2.1. Search link’s title contains word “contact” in the homepage
 - 2.4.2.2. Repeat step 2.3
3. Else
4. Go to END

According to Fig. 1, there are several steps to search five trust attributes in an e-Commerce Website. Before the processes begin, the user has to insert the URL of the desired e-Commerce Website in order to assess the trustworthiness of the Website. The system will convert the inserted URL into source code. If the process is a success, the system will search all the links and the name of the link in the page that contains the words “privacy”. If the name of the link contains either of these two words, it means that the Website contain privacy policy trust attributes. For the four trust attributes which are company telephone number, company email address, company address and third party secure transaction, the system will search the attributes in the homepage that has been inserted with the URL page first. If it cannot find the attributes in the page, the system will search the link’s title that contains the word “contact” in that page. According to the Table 2, most e-Commerce Website placed their information such as email, address, telephone number and third party secure transaction in ‘Homepage’ and ‘Contact Us’ page.

5.0 Comparison between Manual Search and System Search

After the patterns have been created using regular expression, these patterns are finally applied in a web based system to search the five trust attributes. The comparison between manual search and system search is shown in Table 8. Table 7 shows the indicator of the results.

According to the table, the system is capable of finding the five trust attributes of almost all websites that have been tested. Some of trust attributes failed to be retrieved because of six factors, such as no mention of state code in the website, spelling error of the state name, address that is written different from with the pattern that has been created, using image for the contact us link, put email information in image format and using another third party for secure transaction that is not included in this study.

The table also shows that almost all the trust attributes are success to be searched from UK, US and Malaysia e-Commerce websites. Only 8 from 150 trust attributes (five attributes for each website) from them are failed to be searched. Its mean, in this case, only 5.33 percents trust attributes are failed to be searched by the proposed tool.

Table 7: Indicators of the results

Symbol	Description
√	Exist
×	Do not exist
Δ	Attribute is found
∞	Attribute is not found
TA	Trust Attribute

6.0 Conclusion and Future Works

It can be concluded that, trust is an important factor in order to establish e-Commerce transactions. Previous researchers have identified the trust attributes that should be placed in an e-Commerce website which are company telephone number, company email address, privacy policy, company address and third party for secure transaction. These trust attributes are the most important things to gain customer confidence towards e-Commerce websites. The study also shows that most of the websites placed the attributes in Homepage and Contact Us pages. Information extraction technique is used since almost all the information in the websites is in unstructured text. The technique used patterns of desired data to extract them. Sample data were taken from each 40 e-Commerce website from UK, US and Malaysia to create patterns of the trust attributes.

This study used regular expressions to write the patterns of the trust attributes. Once the patterns are created, a system that can search the attributes can be developed according to the search algorithm that has been proposed. This study is important since the task of searching the trust attributes is somehow hard especially for the beginner computer user and tools to help them in searching the trust attributes should be developed.

The experiment that has been conducted shows that proposed patterns can be apply to the proposed tool to search the trust attributes in order to assist the consumer to place trust in an e-Commerce website. The future work that will be carried out shall include testing the effectiveness and the accuracy of the technique that are using regular expression and finally testing the practicability of the system that has been developed.

Table 8: Comparison result between manual search and system search.

URL	Manual Search					System Search					Attribute that failed to be searched & the reason
	Trust Attributes					Trust Attributes					
	CEA	CA	CTN	PP	TPST	CEA	CA	CTN	PP	TPST	
US											
http://www.americanmedical-id.com/	×	√	√	√	√	×	∞	Δ	Δ	Δ	TA - CA Reason - No state code in address
http://www.scholarships.com	×	√	×	√	√	×	∞	×	Δ	Δ	TA - CA Reason - No state code in address
http://www.pageonce.com/	√	√	√	√	√	Δ	Δ	Δ	Δ	Δ	All TA are found
http://www.diapers.com/	√	√	√	√	√	Δ	Δ	Δ	Δ	Δ	All TA are found
http://www.babyearth.com/	√	×	√	√	√	Δ	Δ	Δ	Δ	Δ	All TA are found
http://www.unbeatable.com/	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.nurserydepot.com	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.iseeme.com/	√	√	√	√	√	Δ	Δ	Δ	Δ	Δ	All TA are found
http://www.alight.com	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.maconnection.com	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
UK											
http://www.buy-jeans.net/	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://www.customwaxseals.co.uk	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://www.distinctlybritish.com	√	√	√	×	√	Δ	Δ	Δ	×	Δ	All TA are found
http://www.elc.co.uk/	×	√	√	√	×	×	Δ	Δ	Δ	×	All TA are found
http://www.epicheroes.com	√	√	√	√	√	Δ	Δ	Δ	Δ	∞	TA - TPST Reason - Use Comodo third party secure transaction
http://www.framarhealth.com	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://www.majestic.co.uk/	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.jewellerynow.co.uk/	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.simplysalmon.co.uk	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://www.shoe-shop.com/	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
Malaysia											
http://www.onedropperfumes.com	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://mumbaby.com/	√	√	√	×	×	Δ	Δ	∞	×	×	TA - CTN Reason - Use image instead of text for contact us page
http://www.alicewonders.com/	√	√	√	√	×	Δ	∞	Δ	Δ	×	TA - CA Reason - No state's name
http://www.computermalaysia.com	√	√	×	×	×	Δ	∞	×	×	×	TA - CA Reason - Use image instead of text for contact us page
http://www.cardia.com.my/	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://fashionstore.com.my/	√	√	√	√	×	Δ	∞	Δ	Δ	×	TA - CA Reason - Spelling error for state's name
http://www.goeskincare.com/	√	√	√	×	×	Δ	Δ	Δ	×	×	All TA are found
http://www.rcplanet.com.my/	√	√	√	√	×	Δ	Δ	Δ	Δ	×	All TA are found
http://www.hobbysportz.com	×	√	√	√	×	×	Δ	Δ	Δ	×	All TA are found
http://www.beautyimpress.com	√	√	√	×	×	∞	Δ	Δ	×	×	TA - CEA Reason - Email in image format

References:

[1] The Census Bureau of the Department of Commerce. "Quarterly Retail E-Commerce Sales". <http://www.census.gov/mrts/www/data/html/08Q4.html>, 2009.

[2] K. Jones, L. N. K. Leonard and C. K. Riemenschneider, *Trust influencers on the web*, Journal of Organizational Computing and Electronic Commerce **19** (2009), no. 3, 196-213.

[3] H. Gill, K. Boies, J. E. Finegan and J. McNally, *Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust*, Journal of Business and Psychology, **19** (2005), no. 3, 287-302.

[4] A. R. Che-Hussin, L. Macaulay and K. Keeling, *The importance ranking of trust attributes in e-commerce website*, 11th Pacific-Asia Conference on Information Systems (2007).