A system to support the identification and extraction of information related to privacy policies in electronic commerce websites

Meziane, F, Kasiran, MK and Prince, P

A SYSTEM TO SUPPORT THE IDENTIFICATION AND EXTRACTION OF INFORMATION RELATED TO PRIVACY POLICIES IN ELECTRONIC COMMERCE WEBSITES.

Farid Meziane, Mohd Khairudin Kasiran and Phil Prince
School of Computing, Science and Engineering,
University of Salford,
Salford M5 4WT, UK.

ABSTRACT

Privacy has been identified by many studies as the main problem why customers are not completing their online transactions. They fear to provide sensitive information such as personal and financial details. It is not surprising that most websites nowadays include privacy statements in their websites to encourage customers to complete their transactions. However, only few customers can find this information easily on merchants' websites rendering the use of this information obsolete. The aim of this paper is two folds. First it identifies the link or document dealing with private policies on the website for the customer to consult. However, as the current research shows, these documents are long, tedious to read and contain information that may not be relevant to customers. The second aim of this paper is to automatically extract the most relevant information to the customer in the form of short statement. These statements include information on sharing and selling of personal details, the use of secure technologies, customer satisfaction with the goods purchased, the use of cookies and unsolicited communications. We use the Semantic distance similarity model as the basis to develop the information extraction component.

Keywords: Trust; Electronic Commerce; Privacy Policies; Information Extraction; Semantic Distance.

[1] INTRODUCTION AND MOTIVATION

In [3], Meziane and Kasiran developed a trust model for Business to Customer (B2C) Electronic Commerce (EC). The model identified four major factors that are known to increase customers' trust when they are available on a merchant's website. These factors are: existence, affiliation, policy, and fulfilment. There are a number of elements that contribute to each of the overall factors. For example, the evaluation of the Existence factor would be achieved by assessing the amount of contact details, such as address, phone number and fax number displayed on the site. From the policy perspective, there are different types of policies included in EC websites. Evaluating these policies is a very challenging task as they present many difficulties that can be summarised as follows:

- They are written in natural language and therefore can be ambiguous and any interpretation requires the use of advanced linguistic tools.
- They are difficult to localise on EC websites.
- They are usually included as large textual files and customers hardly read them.

The identification of the components of these policies will contribute to the policy factor. As sites are developed independently and there are no standards to follow when developing these standards, their contents vary from one site to another. However, in general, most sites would include details on:

- ***Privacy:*** This would include how a site intends to use personal information and how they will ensure the safety of customer details.
- ***Warranty***: It would be expected that the sites display their warranty information for products purchased. In a sense this is no different from regular shopping as warranty agreements are a standard feature of purchasing arrangements.
- ***Returns:*** Similarly to warranty information, the sites returns details should be displayed, stating under what conditions returns will be accepted.
- ***Customer Satisfaction***: This is something that would show to potential customers that the site has a reputation for providing good customer service.

The aim of this research is to develop a system that first localises merchants' policies on EC websites and then extracts information related to the details of these policies. The analysis will be achieved by manually studying policies contained within E-commerce sites. The outcome from this will include the statements that the system will need to be searching for and the rules to enable the program to extract these statements. Whether these websites have specific policies for each element or include them in one terms and conditions policy is not relevant. This project will therefore not be looking for the existence of policies only but it will attempt at extracting the information that should be contained in the policies as the existence of a policy itself does not mean that the policy contains the relevant information that will be useful to customers and so should not be the element that identifies good policy information.

The remaining of this paper is organised as follows. In section 2 we describe the privacy policies considered in the development of this system; in section 3 we give an overview of the system developed and our experiment. In section 4 we evaluate our system in terms of precision of the extracted information and discuss its future development.

[2] PRIVACY POLICIES IN ELECTRONIC COMMERCE

2.1 Introduction

Online privacy policy is understood as the set of statements explaining how consumers' privacy is dealt with and protected by the web merchant. Public surveys indicate that privacy is the major concern for people using the Internet [2]. A study by the University of California has shown that 94.4% of Americans are concerned about the privacy of their personal information when buying online [6]. Privacy related complaints that are made to the US Federal Trade Commission include complaints about unsolicited email, identity theft, harassing phone calls, and selling of data to third parties [4]. Important requirements for EC security are the need to protect sensitive information that is stored on computers before and after an EC transaction, to verify the identity of the other party in the transaction, to ensure that no one can intercept the information being exchanged during the transaction, and in general to prevent disruption of services and applications [5].

In EC, policies such as privacy, customer satisfaction and guarantee can help consumers evaluate the trustworthiness of a merchant. These policies can influence the level of risk involved in the transaction. Merchant policy such as money back guarantee can lower consumers' risk by giving more control to the user towards the output of the transaction since they can return the product without total loss if they are not satisfied with their quality. The variables retained for the policy components are: customer satisfaction policy (P1), privacy statement (P2) and warranty policy (P3).

Policies found on EC sites are often very long and are therefore frequently ignored by customers. This is a dangerous thing for a customer to do, before they part with any money they should know that the company they are purchasing from is going to be liable if certain things occur. Also it is of great importance that a customer realises what is going to happen to their personal information once it has been past to an online vendor. The security levels that the site uses are also something that a customer should be aware of before typing any details such as credit card numbers. All this information should be made available to a customer through the company's policy wordings. However not all sites include such information and this is where it becomes confusing for customers to know whether they are legitimate or not. Automatically extracting important information for customers is therefore an interesting idea. With such a system customers will have the ability to see for themselves (with minimum amount of effort on their part) whether a site contains information that would show that it is trustworthy.

To achieve an extraction system it is vital that a good understanding of the information to be extracted is achieved. In this instance the important information to customers that is held in policies, is what the system needs to extract. To identify this information a study of the policies on randomly chosen EC sites was required. The only way to accomplish this was to literally visit web sites and read the policies to see what details they contained. After manually searching and reading policies five statements were identified as the most common statements included in privacy policies. The statements and reasons for including them are discussed in section 2.2. For the program to be successful it should be able to find the statement on sites if it exists. If this were achieved then the program would be able to show whether a site can be seen to be trustworthy in terms of the policy information it contains.

2.2 Expected Statements

The statements identified from policies that need to be searched for and extracted by the program will be described. The justifications for choosing these statements will also be discussed, this will include why they are important to customers of EC sites.

2.2.1 Description of Statement 1

The first statement proposed is, "We will not sell or share your personal details with unrelated third parties". This is a particularity important aspect for people purchasing from E-Commerce sites. Personal details have to be entered and it is essential that sites state how they are going to use the potentially sensitive data. With credit card information and personal contact details being required the handling of this data and in particular who the company will let see the data is of paramount importance. If there is no statement that specifically says that they will not sell or share personal information then it is a possibility that they will. Anything other than passing information on to other companies to help in the retail process is something that is not good for customers and companies that practice this should be avoided. This type of statement is generally found in Privacy Policies or in the general Terms and Conditions of sites.

2.2.2 Description of Statement 2

The second statement proposed is, "We use Secure Socket Layer (SSL), or other encryption technology". The levels of security put into place by the site are of extreme importance.

Credit card details should only be imputed on secure connections and it is expected that sites explain in their policy wording that they use some sort of encryption technology to protect their customers. This sort of information is generally not something that inexperienced computer users will understand therefore it is unlikely that they would look for it themselves before purchasing from a site. This type of statement is generally found in Privacy Policies or in specialised Security Policies.

2.2.3 Description of Statement 3

The third statement proposed is, “If you are not completely satisfied with a product or it is faulty, you may return it”. It is important that customers realise how company’s refund policies work. This is especially important for large purchases. Legitimate sites will work much like regular high street stores where if there is a problem with a purchased item they will offer a refund or replacement. In the faceless E-Commerce environment it is important that a version of this statement is clearly defined in a policy document. This type of statement is generally found in Return/Refund Policies, Customer Satisfaction Policies or in the general Terms and Conditions of the site.

2.2.4 Description of Statement 4

The fourth statement proposed is “Our site uses cookies to automatically collect information about you”. Cookies are a way of gathering information from customers without their prior knowledge or consent. Therefore it is expected that sites that use cookies tell their customers that they are doing so. This is again something that inexperienced users will not know very much about so will not be looking for when reading policy documents. Most legitimate sites also inform customers that they can switch off cookies if they wish to do so. This type of statement is can generally be found in Privacy or Security Policies.

2.2.5 Description of Statement 5

The fifth statement proposed is, “If you wish to stop receiving information from us, then contact us”. Many sites use the personal information given to them (such as email address and postal address) to send advertisements and other mail. It is expected that sites that do this would give the customer the option to opt-out from receiving these mailings. This is something that many customers neglect to think about when giving out personal details. This leads to a huge amount of spam/junk mail being received after giving out details on Internet sites. Legitimate sites would always give customers the chance to no longer receive advertising information. Similar to statement one this type of statement is generally found in Privacy Policies or in the general Terms and Conditions of sites.

2.3 How to Identify a Statement

It is important to accepted that sites are unlikely to contain these statements in the way they have been expressed. Synonyms, grammatically different structures and different writing styles will be used. To extract statements a set of rules will be defined that a statement must adhere to if it is to be extracted. As what will be searched for is a sentence there may be certain times when something is extracted that does not mean the same as the statement it was intended to. This is an inherent problem when searching text. However in the vast majority of cases it is expected that a sentence will have the same meaning as the expected statement and in this case it will be accepted as a viable alternative to the statements. Subsequently it will be up to the users own judgement to recognise if what has been extracted is comparable to the statement. This is a problem that cannot be overcome due to the vast array of different writing styles and techniques that developers will be using when writing policy documents on sites. When manually searching for statements, personal judgement will need to be used to decide whether a statement has been found.

The rules therefore will need to be as tight as possible, but without becoming so tight that they never find a sentence that is a valid statement. There is a need for an element of flexibility in the rule or it will only find sentences that are written in exactly the same way. However, finding an exact statement replicated from one site to another is unlikely to occur.

The manual searching for statements will be done by taking 25 random EC sites and searching each policy they contain for evidence of the five statements. Section 3.4 covers this process in detail and discusses the output that was accrued. This will then enable the rules to be written and these will eventually be used in the program that will be developed.

2.4 Searching Policies for Statements

As every statement was found the web address, the link leading to this statement is also noted. This was important for the development of the system as web pages often contain many links and the system needs to identify which links are likely to lead to a privacy policy. The rest of the links are then ignored.

The number of manually found statements on the random sites gives some justification for the five statements chosen. An overall average of over 81 percent signifies that approximately 4 out of the 5 statements are found per random web site. This high figure shows that the statements chosen to search for occur often enough on sites to provide a good platform for the project. Table 1 shows the number and percentages for each of the statements that have been found in the random selected sites.

Table 1: Occurrences of Statements in 25 Random Sites

Statement	Number of Occurrences	Percentage
1	20	80%
2	21	84%
3	20	80%
4	20	80%
5	21	84%
Total	**102**	**81.60%**

2.5 Identifying the Links for Policy Pages

After studying the statements found on the random sites it was possible to determine the links that the policy information is found on. As previously stated, as web pages often contain many links it would not be possible to search each one for policy information. Therefore a way of validating links is required to enable the program to eliminate links that would not show policy information. By using the links that the policy information was found on it was possible to identify a set of keywords. The keywords that were identified are shown in Table 2. During the manual search it was seen that many policies could be accessed directly from the homepage of a site. However some sites have a homepage that serves as an advert to the main site. In this case there would be no direct link to policy information; therefore searching the next layer down the hierarchy of a site would be required. It is important to note that these keywords are what have been obtained during the initial search of 25 sites. As this is an extraction system in order to obtain a greater precision the number of keywords would be expected to grow as more sites are searched.

Table 2: Occurrences of Statements in 25 Random Sites

Link Layer	**Keywords**
Policy Links	Privacy, Terms, Warranty, Help, Policy, Info, User, Security, Secure, Content, Returns, Customer, Satisfaction, Order, Safe, Article, RMA, About.
Intermediate Links	Index, Home, Main, Default, Redirect, Contact, Session, Catalogue.

3 THE EXPERIMENT

Two steps are involved in the development of the information extraction system. These are the construction of a lexicon that contains worlds denoting the same meaning. For example for the first statement "We will not sell or share your personal details with unrelated third parties" a sample of the lexicon constructed includes the following information:

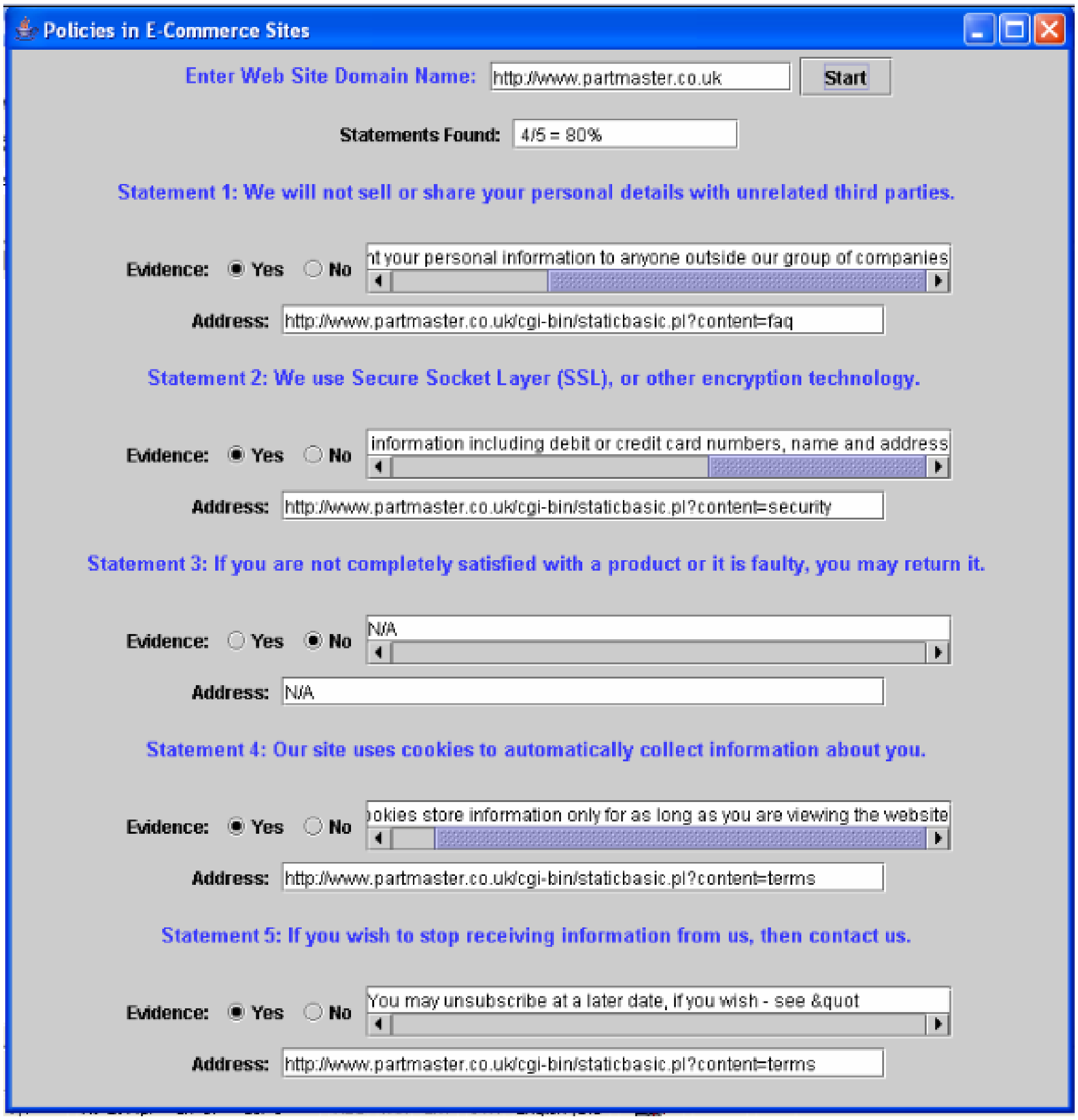
{{not sell, never sell, not share, never share, not disclose, not pass, never disclose, never pass}, {information, data, details, personal details}, {third party, third parties}}

The expressions in each set denote the same semantics in this specific context.

The second step is then to look at the inclusion of these expressions in sentences and how close to each other they are. If an instance of each expression is present and they are within a significant distance from each other, then the probability is that the sentence denotes one of the statements. An example of rules identified for the first statement is:

IF the words {**not, share, information}** are found in a sentence and the SD(not, share) = 1 and SD(share, information) <= 10, THEN **statement found**

Here SD stands for the semantic distance, which in this case is a simple subtraction between the positions of two words within a sentence. A total of 85 rules have been hand crafted in the first instance based on the privacy policies extracted from the 25 websites studied. The system is implemented in Java and the interface is shown in Figure 1.



Enter Web Site Domain Name: http://www.partmaster.co.uk

Statements Found: 4/5 = 80%

Statement 1: We will not sell or share your personal details with unrelated third parties.

Evidence: ● Yes ○ No nt your personal information to anyone outside our group of companies

Address: http://www.partmaster.co.uk/cgi-bin/staticbasic.pl?content=faq

Statement 2: We use Secure Socket Layer (SSL), or other encryption technology.

Evidence: ● Yes ○ No information including debit or credit card numbers, name and address

Address: http://www.partmaster.co.uk/cgi-bin/staticbasic.pl?content=security

Statement 3: If you are not completely satisfied with a product or it is faulty, you may return it.

Evidence: ○ Yes ● No N/A

Address: N/A

Statement 4: Our site uses cookies to automatically collect information about you.

Evidence: ● Yes ○ No ookies store information only for as long as you are viewing the website

Address: http://www.partmaster.co.uk/cgi-bin/staticbasic.pl?content=terms

Statement 5: If you wish to stop receiving information from us, then contact us.

Evidence: ● Yes ○ No You may unsubscribe at a later date, if you wish - see &quot

Address: http://www.partmaster.co.uk/cgi-bin/staticbasic.pl?content=terms

Figure 1: The system interface

4 SYSTEM EVALUATION AND DISCUSSIONS

A sample of 25 EC websites is used to evaluate the system. The results are summarised in Table 3.

Table 2: Occurrences of Statements in 25 Random Sites

	Actual	Extracted	Precision
Statement 1	23	16	70%
Statement 2	23	20	87%
Statement 3	22	12	55%
Statement 4	23	21	91%
Statement 5	19	11	58%

As a starting point for an extraction system the precision calculated is very favourable and this is shown when comparing the figures to a similar system [Meziane and Kasiran, JORS]. In general the accuracy and robustness of machine extraction systems can be improved greatly due to the shallow understanding of the input [1]. Therefore an average 73% success rate for a first generation program as developed here can be seen as a success.

The precision of the system can be improved in different way. The first is with regards to the methodology used for the information extraction. Approaches such as fuzzy logic and a-gram models may improve the overall precision of the system. On the other hand a better understanding of the way policies are written by a study of a larger sample and a search in deeper layers of the websites can also improve the precision of the system.

5 REFERENCES

[1] Caride, C. (1997), Empirical methods in information extraction - Natural Language Processing,http://www.findarticles.com/p/articles/mi_m2483/is_n4_v18/ai_20475200#continue, [5 September 2005].

[2] Cavoukian A and M Crompton (2000), Web Seals: A Review of Online Privacy Programs, http://www.ipc.on.ca/english/pubpres/papers/seals.pdf, 2000.

[3] Meziane, F and Kasiran, MK. (2008), 'Evaluating trust in electronic commerce: a study based on the information provided on merchants' websites', *Journal of the Operational Research Society*, 59 (4) , pp. 464-472.

[4] Mithal M (2000), Illustrating B2C Complaints in the Online Environment, the Joint Conference of the OECD, HCOPIL, ICC: Building Trust in the Online Environment: Business to Consumer Dispute Resolution, The Hague, http://www1.oecd.org/dsti/sti/it/secur/act/online_trust/presentations.htm.

[5] Patton M A and Jøsang A (2004), Technologies for Trust in Electronic Commerce, *Electronic Commerce Research*, (4):9–21.

[6] University of California-Los Angeles Centre for Communications policy (2001), the UCLA Internet report 2001: Surveying the digital future.