E-Commerce Web Site Trust Assessment
Based on Text Analysis

Banatus Soiraya
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
1518 Pibulsongkram Road, Bangsue Thailand, 10800
banatuss@yahoo.com

Anirach Mingkhwan
Faculty of Industrial and Technology Management
King Mongkut's University of Technology North Bangkok
1518 Pibulsongkram Road, Bangsue Thailand, 10800
anirach@ieee.org

Choochart Haruechaiyasak
Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
112 Phahonyothin, Pathumthani, Thailand, 12120
choochart.haruechaiyasak@nectec.or.th

ABSTRACT
This paper reports the efficiency and effectiveness of our proposed text analysis module for a Web-site trust assessment system. One main factor that influences the level of trust for an e-commerce Web site is the textual content appearing on each Web page, especially the main page. The textual content refers to words, phrases, and sentences appearing on the page. A Web site that has a high trust level should contain meaningful keywords related to the e-commerce domain, such as return policy, payment option, and security. To analyze the textual content, our text analysis module adopts automatic classification techniques to learn from the example Web site data set. We perform experiments on two e-commerce domains (a jewelry store and a book shop) by using three well-known classification algorithms. A sample set of Web sites under each domain are collected and labeled as trust or untrust for performing our experiments. Two approaches for constructing the feature set are (1) using all extracted words from the textual contents (baseline), and (2) mapping extracted words into the meaningful groups of e-commerce terminology (EC-word). The best text analysis result of 83.5% accuracy was obtained when the support vector machines based on a sequential minimal optimization (SMO) algorithm was applied with the EC-word feature set.

Keywords: e-commerce, trust assessment, Web mining, text classification
1. INTRODUCTION
Currently, Web sites provide people with a convenient way to shop on-line. However, in some countries, including Thailand, the number of on-line shoppers is still very low compared with countries such as the United States. Based on some previous surveys, two major problems regarding Web sites are security and lack of trust. Although most e-commerce Web sites provide some form of secured payment, this provision does not guarantee that the Web site will gain better credibility.

The first or main page of a Web site is very important when users access the site for the first time. If it is not properly designed, a customer may leave the site after viewing the main page without browsing other parts of the site [Fogg et al., 2002]. W3 Trust Model (W3TM) is proposed for trust value calculation from meta-data provided on Web sites [Yang, 2004]; however, it does not include analysis of other important factors, such as contents and layout and, more important, the ability to automatically recommend trust-assessment improvement. At present, there are no effective tools for evaluating trust assessment on Web sites. An effective trust assessment tool must be able to correctly identify and determine the trustworthiness level of sites. In addition, the tool should provide recommended information, based on trust assessment, in order for Web masters to use it to improve the site for better trust.

We have developed a research framework called “E-commerce Web-Site Trust Assessment Framework” (WTA) [Soiraya, et al., 2007]. It has three major core components:
1. Web content mining for analyzing Web site contents, such as text and other meta-data
2. Web content mining for page layout detection of the Web sites
3. Recommendations that are generated for both on-line shoppers and the Web master of the site

The recommendation messages given to the Web master include, for example, “Your Web site is lacking a privacy policy; please fill in,” or “Your Web site contains too many advertisements; please reduce them.”

In this paper, we mainly discuss Web content mining. We collect a data set from the jewelry store and book shop domains. We evaluate three well-known classification techniques:
- Naïve Bayes (NB)
- k-Nearest Neighbor (kNN)
- Support Vector Machines, based on Sequential Minimal Optimization (SMO) training [Chakrabarti, 2003; Witten and Frank, 2000; and Platt, 1999] in the WTA aspect

We propose two approaches for constructing the feature set:
1. Using all extracted words from the textual contents (baseline)
2. Mapping extracted words into the meaningful groups of e-commerce terminology (EC-word)
From the previous research [Liu, 2006], we find that most of the best accuracy in text classification is the SVM classifier. In this paper, our experiment has shown similar results. The best accuracy in both approaches is the SMO classifier (SVM classifier base on sequential minimal optimization training).

Although we propose two approaches, we found several limitations and impracticalities in the first approach (baseline). For one thing, it was necessary to store all of the words from every Web site before our classification. Both informative data (meaningful information such as e-commerce word and various policies, as shown in section 4.1) and non-informative data (meaningless information such as product description, product characteristics, and Web site characteristics) are collected together, which can reduce the accuracy in the classification process. Another disadvantage is that the approach is both storage consuming and time consuming, which impacts the efficiency of our system.

The second approach (EC-word) has five advantages:

- **Elimination of non-informative data.** We must select the meaningful information feature selection (words and phrases) from our corpus (trust Web sites training set). Only e-commerce or trust e-commerce words will be chosen from trust categories, as shown in section 4.1. Non-informative (the other attribute of product or other characteristic of Web sites besides trust categorization sample set) data will be removed from our repository.

- **Increasing accuracy of classification.** After we eliminated non-informative data from our corpus, we have only the pure EC-word that will discriminate between trust Web sites and untrust Web sites. In our results, we notice that the best accuracy in the second approach is 83.5% by using the SMO classifier, compared with 79.5% in the first approach.

- **Feature selection reduction.** After the elimination step, we have only the e-commerce word in our corpus. In our results, we notice that the number of features in the second approach (EC-Word) is smaller than those of the first approach (baseline) by 14 times.

- **Rapid classification.** The small number of features in the second approach will be processed for classification by using a time shorter than the first approach. As shown in the section 4.3, for the book shop domain, we use 375 features from the second approach instead of 5,382 features. In other words, we can reduce the comparison time from 5,382 to 375 in our experiments.

- **Guideline for developing feature selection algorithm.** We could develop a feature selection algorithm according to the trust categorization in section 4.1. This algorithm could identify the best suit features (word, phrase). We could increase the accuracy of classification in future research [Ozono and Shintani, 2006].
The rest of this paper is organized as follows. In Section 2, we present the topics related to trust assessment. Section 3 discusses the conceptual idea of our proposed framework. Section 4 provides an example of e-commerce Web site categories. Section 5 discusses automatic classification algorithms. Section 6 presents our evaluation methods and experiment results, which integrate all necessary components and their functionality. Section 7 contains our conclusions and suggestions for future work.

2. RELATED WORK
Egger [2003] developed a model of trust in e-commerce called MOTEC, which describes characteristics of trust in e-commerce in terms of interaction, information, usability, and relationship. Consumer trust patterns were considered to be the component of trust assessment terms such as information policy, return policy, and warranty policy [Kaluscha and Grabner-Krauter, 2003]. At present, only one prototype tool for a trust assessment of Web sites is performed with W3 Trust Model (W3TM) by assessing the relevant metadata of Web documents [Yang, 2004]. This tool proved impractical when we applied it to real Web sites. This tool cannot detect various types of trust information such as content and layout on the Web site.

Web mining has become a popular method of discovering useful information and knowledge from the Web hyperlink structure, page content, and use data by using data mining techniques. In general, Web mining consists of three categories:

- Web content mining
- Web structure mining
- Web use mining

Some of the well-known classification techniques for Web content mining such as Naïve Bayes, k-Nearest Neighbor (k-NN), and Support Vector Machines (SVM) will be adopted in this paper [Chakrabarti, 2003; and Liu, 2006]. Most of the previous researches were Web page or Web site classification using the categories [Kriegel and Schubert, 2004; and Liu, 2006], but there were no previous works on Web site classification using the trust assessment feature in the area of e-commerce area.

3. OUR PROPOSED FRAMEWORK
In our previous paper, “E-commerce Web-Site Trust Assessment Framework,” we proposed a framework for helping users select trust Web sites and improve Web site design in terms of trust assessment [Soiraya, et al., 2007]. Our proposed framework consists of two phases:

1. Content-based analysis
2. Decision analysis.

The major components are illustrated in Figure 1.
(1) Content-Based Analysis. This phase analyzes a Web site in terms of both text and layout. The output is a content-based score that corresponds to the trust assessment level of the Web site. The first step is information gathering, in which information is collected from the Web site and the Web pages are stored automatically in the Web repository. The second step is feature extraction and pre-processing, which eliminate the html tag and unnecessary words in the plain text. The third step is the analysis of the text and layout, which includes a comparison between the data set and the test set, using the trust assessment conditions.

(2) Decision Analysis. This phase uses the total trust score from the previous phase. Two processes are then performed: (a) trust criteria, and (b) recommendations for improving the Web site.

(a) Trust criteria. We consider four trust levels:
- High trust Web sites: Web sites that provide high-quality layout and textual content
- Moderate trust Web sites: Web sites that provide medium-quality layout and textual content
- Low trust Web sites: Web sites that provide low-quality layout and textual content
- Untrust Web sites: Web sites that provide no-quality layout and textual content

Figure 1. E-Commerce Web-Site Trust Assessment Framework
(b) **Recommendations for improving the Web site**: For example, the following types of messages could be given to the Web site administrator.

- “Your Web site lacks a privacy policy. Please provide the information to improve the Web site’s credibility.”
- “Your Web site layout looks unprofessional and there are too many advertising banners. Please arrange your layout and reduce the advertising banners.”

The user benefits from this framework in terms not only of improving Web site trustworthiness, but also in obtaining recommendations for overall Web site improvement.

4. **E-COMMERCE WEB SITE DATA SET**

For our experiments, we collected and labeled example pages from Web sites as *trust* and *untrust*. In this paper, we perform text analysis on two different domains: a jewelry store and an online book shop. We classify trust patterns (textual content) that appear on each Web site; i.e., privacy policy, 30-day return policy, and secured transaction [Adkisson, 2002; Egger, 2003; and Yang, 2004].

4.1. **Jewelry Store Domain**

In the jewelry store domain, we classify Web sites into four categories:  
(1) *Trust* Web sites with high-quality content and layout
(2) *Candidate trust* Web sites with high-quality layout
(3) *Candidate trust* Web sites with high-quality content
(4) *Untrust* Web sites with low-quality content and layout

(1) *Trust* Web sites with high-quality content and layout

Figure 2 shows the main page of the web site www.jewelrydays.com, which has high-quality content and layout. Template design and layout design are state-of-the-art. Five blocks (A, B, C, D, and E) are in the appropriate position. Each block has trust e-commerce words. In Block A, the words *Blog, FAQ, About Us, Privacy Information,* and *Contact Us* explicitly represent trustworthiness to customers. A greater sense of security is imparted by references to secure transaction, guarantee, and satisfaction and return policy in blocks B, C, and D. Moreover, Block E provides customer service and testimonials, which increase customer confidence even more.
(2) Candidate trust Web sites with high-quality layout

We use the same concept from the previous Web sites (content analysis and layout analysis). The Web site www.gifts-you.com does not have trust content, but the layout is rather good (Figure 3). This site has a professional layout design, but lacks a trust pattern (textual content); i.e., privacy policy, 30-day return, and secure transaction. Though we classify in two blocks (A and B), we hardly notice those trust patterns. A few words in Block A (Contact Us) and Block B (Shopping Cart) do not provide a greater sense of confidentiality. The only signs on the top right side of the screen are several credit card pictures (Visa, MasterCard, Amex), but we exclude from our research any graphic format or jpg.
Candidate trust Web sites with high-quality content

From www.jansjewels.com, we notice that the template design and layout design are unprofessional (Figure 4). We classify it in two blocks (A and B). Block A provides more confidentiality and security information to the customer, such as Shopping Cart, Contact Information, Guarantee and Security Policy, and Shopping Cart Problems. Block B provides secure ordering and a guarantee to customers. Although credit card logos (Visa, MasterCard, and Paypal) appear on the Web site, we exclude any graphic format or jpg from our research.
Jan's Jewelry Supplies offers over 9,000 jewelry making supplies including vintage rhinestones, settings, findings, stampings, screw findings, cameos, charms, and sew-on rhinestone findings. Jewelry making supplies for the professional jewelry designer and the hobbyist.

Try to update the site at least every other week, so please bookmark this page.

E-Mail Notification will let you know when the site has been updated. To see what has been added between visits, check What's New.

Click here for Jewelry Making Kits as featured in BeadStyle magazine.

MULTIFUNCTION Jewelry

Have been designing jewelry for years and have come up with a new concept called "MULTIFUNCTION Jewelry." The key to the concept is a custom designed pendant that can be worn as a pendant, necklace, belt loop, jewelry purse decor, sweater guard, belt, bracelet and more. This design is cost and space effective. Please take a look and let me know what you think.

Tips on visiting the site:

All pieces on my site have a small icon image and a larger detailed image. Just click on the image or item number for a detailed picture.

A navigation menu is provided at the bottom of each page. The "Home" link brings you back to this page. The "Next" link will take you to the next page in a collection. The "Catalog" link moves you to a site map.

For detailed information on placing orders, please see How to Order & Guarantees.

For detailed information on placing orders, please see How to Order & Guarantees.

Jan Kostera
C H Marketing Inc.
3333 NW 50th Street
Oklahoma City, OK 73116

Phone: 405.542.3341
Fax: 405.542.0671

You can E-Mail me at jan@janjewels.com

Figure 4. Example of a Candidate Trust Web Site with High-Quality Content

International Journal of Business and Information
(4) **Untrust** Web sites with low-quality content and layout

From the Web site www.jewelbiz.com, we notice that there are only product descriptions and that the trust context is absent on the main page (Figure 5). A few words such as **Contact Us** and **About Jewelbiz** do not provide more confidentiality to customers. The Web site design is also unprofessional.

![Jewelbiz](image)

**Figure 5. Example of Untrust Web site with Low-Quality Content and Layout**
4.2. **Book Shop Domain**

In the book shop domain, we classify Web sites into three categories:

1. *Trust* Web sites with high-quality content and layout
2. *Candidate trust* Web sites with high-quality layout
3. *Untrust* Web sites with low-quality content and layout

Following are the Web site characteristics of the three categories.

1. **Trust** Web sites with high-quality content and layout
   
   For www.barnesandnoble.com (Figure 6), we used the same concept from the previous domain. We classified it in three blocks (A, B, and C). There are some words in Block A, such as *Shopping Cart, Order Status*, and *Search*. In Block B, there are some words such as *Help* and *Affiliate Policy*. Block C provides more security to customers, such as *Return Policy* and *Customer Service*.

2. **Candidate trust** Web sites with high-quality layout
   
   For www.dymocks.com.au (Figure 7), we classify the site in two blocks. Both Block A and Block B provide service to customers, but there are no security or trust words such as *Secure Payment, Privacy*, or *Return Policy* on this page. This Web site has a very nice graphic and professional layout design, but lacks of high-quality textual content.

3. **Untrust** Web sites with low-quality content and layout
   
   From www.mjtbooks.com (Figure 8), we do not see the trust context, nor a professional layout design. Furthermore, there is an absence of product detail on this web site. Most customers might not be satisfied with this Web site.
Figure 6. Example of Trust Web Site with High-Quality Content and Layout
Figure 7. Example of a Candidate Trust Web Site with High-Quality Layout
5. AUTOMATIC CLASSIFICATION ALGORITHMS
Automatic classification is used to classify the Web sites into trust and untrust categories. Various techniques are the foundation of data mining algorithms such as Naïve Bayes, kNN, and Support Vector Machine [Chakrabarti, 2003; Liu, 2006]. This section discusses the classification method for WTA.

5.1. Naïve Bayes Classifier
A Naïve Bayes classifier; [Liu, 2006; Yong, Wang et al., 2003] is constructed using the training data to estimate the probability of each category, given the document feature values of a new instance. We use Bayes theorem to estimate the probabilities.
We assume $E$ denotes the attribute set and $c_i$ denotes the class variable. We have the conditional probability that we called $P(c_i \mid E)$ in terms of posterior probability for $c_i$. And we called $P(c_i)$ in terms of prior probability. $P(E \mid c_i)$ is called the class-conditional probability, and $P(E)$ is called the evidence. Bayes theorem can express the posterior probability in terms of the prior probability. A naïve Bayes classifier estimates the class condition probability by assuming that the attributes are conditionally independent, given the class label $c_i$.

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$  (1)

$$P(E \mid c_i) = P(e_1 \land e_2 \land \cdots \land e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$  (2)

5.2. kNN Classifier

The kNN algorithm [Liu, 2006; Shintani et al., 2006] is quite simple. It finds the $k$ nearest neighbors of the test document from the training documents. The categories of these nearest neighbors are used to weight the category candidates. The similarity score of each neighbor document to the test document is used as the weight for the categories of the neighbor document. Because there is no model creation, this method is called lazy learning. Typical similarity is measured with a cosine function or Euclidean distance. We illustrate Euclidean distance in the following equation.

$$d(p,q) = \sqrt{\sum (p_j - q_j)^2}$$  (3)

We can measure the distance between two objects in term of coordinates of a pair of objects. We assume $d(p,q)$ denotes the distance and $p$, $q$ is the coordinate of the x or y axis.

5.3. SVM and SMO Classifier

Support Vector Machine (SVM) is based on the statistical model. In this paper, we focus only on the linear form [Liu, 2006; Aixin, Sun et al., 2002]. SVM has a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin.

Let the set of training examples $D$ be

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is an r-dimensional input vector, $y_i$ is its class label (output value). and $y_i \in \{-1, +1\}$. 1 denotes the positive class and -1 denotes the negative class.
To build a classifier, SVM finds a linear function of the form

\[ f(x) = \langle w \cdot x \rangle + b \]  

(4)

so that an input vector \( x_i \) is assigned to the positive class if \( f(x_i) \geq 0 \), and to the negative class, otherwise; i.e.,

\[ y_i = 1 \quad \text{if} \quad \langle w \cdot x \rangle + b \geq 0 \]
\[ y_i = -1 \quad \text{if} \quad \langle w \cdot x \rangle + b \leq 0 \]  

(5)

We will call \( \langle w \cdot x \rangle \) that dot product of \( w \) and \( x \) (or Euclidean inner product). In other words, we can describe in

\[ f(x_1, x_2, \ldots, x_r) = w_1 x_1 + w_2 x_2 + \ldots + w_r x_r + b, \]

In essence, SVM finds a hyperplane that separates positive and negative training examples.

\[ \langle w \cdot x \rangle + b = 0 \]  

(6)

This hyperplane is called the decision boundary or decision surface, since SVM maximizes the margin between positive and negative data points. SVM looks for the separating hyperplane with the largest margin, which is also called the maximal margin hyperplane, as the final decision boundary.

We consider a positive data point \((x^+, 1)\) and a negative \((x^-, -1)\), which are closest to the hyperplane \( \langle w \cdot x \rangle + b = 0 \). We define two parallel hyperplanes, \( H^+ \) and \( H^- \), which pass through \( x^+ \) and \( x^- \), respectively, in Figure 9.

\[ H^+ \langle w \cdot x \rangle + b = 1 \]  

(7)

\[ H^- \langle w \cdot x \rangle + b = -1 \]  

(8)
Now, we compute the distance between the two margin hyperplanes, $H^+$ and $H^-$. Their distance is the margin $(d_+ + d_-)$. We ignore the step for Euclidean distance computation. Finally, we have this equation:

$$\text{Margin} = (d_+ + d_-) = \frac{2}{\|w\|} \quad (9)$$

where $\|w\|$ is the Euclidean norm of $w$ ($\|w\| = (w_1^2 + w_2^2 + \ldots + w_r^2)^{1/2}$) \quad (10)

Because of SVM’s slow training and algorithm complexity, Sequential Minimal Optimization (SMO) is used to solve the problem, and we adopt the SMO technique in our research [Platt, 1999]. SMO is new algorithm for training Support Vector Machine. It is constructed using three components:

1. An analytic method to solve for the two Lagrange multipliers
2. A heuristic for choosing which multipliers to optimize
3. A method for computing

This paper does not discuss the technique in mathematical detail.

6. EVALUATION METHOD AND RESULTS

This section discusses evaluation methods (6.1), data preparation (6.2), and the results obtained (6.3).

6.1. Evaluation Methods

Our objective is to investigate the effectiveness of two approaches – Approach A (baseline) and Approach B (EC-word) -- using three types of classifiers -- Naive Bayes (NB), k-Nearest Neighbor (kNN), and Support
Vector Machine (SVM), based on Sequential Minimal Optimization (SMO) training.

Two domains of Web sites (jewelry store and book shop) are studied in two approaches, using the Weka [www.cs.waikato.ac.nz/ml/weka/] machine learning tool. We use three tools in our experiments: Weka [Ian and Witten, 2000], Java, and HTML Tag Removal. Weka is used as a classification tool for accuracy measurement of each approach, The Java program is used for feature selection algorithms such as word extraction, stop word removal, and stemming. HTML Tag Removal is used to convert HTML files to normal text.

6.1.1. Approach A (Baseline)

This is called the baseline approach since it uses all extracted words from the textual content. It has two major components: (1) feature selection for creating the feature or keyword file from the training file, and (2) format ARFF for creating file format input for Weka.

(1) Feature Selection selects features (keywords) from the training file or training set. Feature selection is a program for tokenizing the text to individual token words. This program checks against stop words (a, an, the) and removes them from the keyword file. Then, the stemming process transforms those words in either singular or plural forms into root words such as Privacy, Privacies (Privaci) to eliminate the ambiguity shown in Figure 10. Moreover, this step detects and keeps the word occurrences in the keyword file since they have a frequency equal to at least three times in our experiment. The output of this part is a keyword file that stores feature words.

(2) Format ARFF uses the keyword file from the previous step to check words and map them with the training file. The training file is also processed with tokenizing and stemming. The result of mapping is illustrated in weight values from the word frequency. The output of this part is an ARFF file that is used in Weka, as shown in Figure 10. The ARFF format is shown in Figure 11. We prepare ARFF for both the training data set and the testing data set.
Figure 10. Two Approaches (A and B) of the Evaluation Method
Figure 11. ARFF File Format

ARFF consists of five major areas:

(a) Relation name (such as @relation jewelry.train) is the first line in the ARFF file. It identifies the starting point in each file.

(b) Attribute declaration is an ordered sequence of attribute statement. The <datatype> can be any of several types, such as numeric, integer, and string. We use the numeric (0-374) for feature selection in Approach B.

(c) Class declaration defines how many classes (such as trust and untrust) are in our experiment. We can use the last attribute (375) to create the class.
(d) @ data declaration is followed by a single line denoting the end of the instance. It starts with an open parenthesis and ends with a close parenthesis, which includes the class label (trust, untrust). We adopt a sparse ARFF format in our research. Because of the column reduction, data with value 0 is not explicitly represented. The first position is the location of each attribute. The next position is the frequency of each attribute, along with the row, until the last attribute is shown. If this position is 0, it is not represented.

(e) Data declaration is the same as (d), Notice that the untrust class is labeled at the end of each line. The number of lines after @data depends on the training set.

6.1.2. Approach B (EC-Word)

This approach is called EC-Word because it maps extracted words into the meaningful groups of e-commerce terminology. This approach is similar to Approach A, except for the feature selection part, as shown earlier in Figure 10. There are two major parts of Approach B: (1) Manipulation EC-Word, for creating the feature or keyword file from trust assessment criteria, and (2) Format ARFF for creating file format input for Weka.

1) Manipulation EC-Word creates trust e-commerce words using the trust assessment criteria. We adopt the previous papers [Egger, 2003; Yang, 2004] and use our own empirical research from real Web sites. There are six classes for building trust assessment in EC-Word, as shown in Figure 12.

Figure 12. Trust Knowledge Base Structure

International Journal of Business and Information
(a) **Company Class**: This property provides useful company information and appropriate company characteristics, as illustrated in Table 1. The appropriate word is called EC-Word in the right column.

**Table 1. Company Class Property Information**

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Company Information</td>
<td>About Us</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corporate Information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Company Overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Company History</td>
</tr>
<tr>
<td>2</td>
<td>Investor Information</td>
<td>Investor Relation Press</td>
</tr>
<tr>
<td>3</td>
<td>Information Updated</td>
<td>Last updated</td>
</tr>
</tbody>
</table>

(b) **Product Class**: This property provides useful product or service information and their characteristic to customers, as illustrated in Table 2. The appropriate word is called EC-Word in the right column.

**Table 2. Product Class Property Information**

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Product &amp; Service Reputation</td>
<td>Award Winning</td>
</tr>
<tr>
<td>5</td>
<td>Warranty, Guarantee</td>
<td>Insurance</td>
</tr>
<tr>
<td>6</td>
<td>Professional Service</td>
<td>Professional Service</td>
</tr>
</tbody>
</table>

(c) **Security Class**: This property provides the confidentiality guarantee service and security to customers, as illustrated in Table 3. The appropriate word is called EC-Word in the right column.
Table 3. Security Class Property Information

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>Standard Compliant Policies</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Secure Payment Secure Shopping Secure Ordering Paypal Visa, Master Card, JCB, etc.</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Certification SSL Certificate</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Return Policy Return Policy</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Financial Compensation Money Back Guarantee Replacement</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Third-Party Guarantee Insurance, Verisign, Bizrate, Hacker</td>
<td></td>
</tr>
</tbody>
</table>

(d) Privacy Class: This property provides customer privacy and assures confidentiality for protection personal of customers, as shown in Table 4. The appropriate word is called EC-Word in the right column.

Table 4. Privacy Class Property Information

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.</td>
<td>Customer Information Personal Confidential</td>
<td>Privacy Policy</td>
</tr>
<tr>
<td>14.</td>
<td>Privacy Seal Web Trust Web Seal Trust E BBBonline</td>
<td></td>
</tr>
</tbody>
</table>

(e) Relation Management Class: This property provides the customer relation management and useful service to customers, as illustrated in Table 5. The appropriate word is called EC-Word in the right column.
Table 5. Relation Management Class Property Information

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.</td>
<td>Online Contact</td>
<td>Chat, email, login, logon</td>
</tr>
<tr>
<td>16.</td>
<td>FAQ</td>
<td>FAQ</td>
</tr>
<tr>
<td>17.</td>
<td>Customer Support</td>
<td>Affiliate Program, Support, Customer Care, Customer Service</td>
</tr>
<tr>
<td>18.</td>
<td>Track Order</td>
<td>Track Orders, Shipping Information, Ordering, Order Tracking, Order Status</td>
</tr>
<tr>
<td>19.</td>
<td>Customer Review or Customer Recommend</td>
<td>Feedback, Testimonials, Customer Satisfaction, Customer Commend, Suggestion Box</td>
</tr>
</tbody>
</table>

(f) **Usability Class:** This property provides the useful facilitation tool for helping customers, as illustrated in Table 6. The appropriate word is called EC-Word in the right column.

Table 6. Usability Class Property Information

<table>
<thead>
<tr>
<th>Index</th>
<th>Trust Description Name</th>
<th>EC-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.</td>
<td>Help</td>
<td>Help</td>
</tr>
<tr>
<td>21.</td>
<td>Web Directory</td>
<td>Site Map, Site Index, Directory of all stores</td>
</tr>
<tr>
<td>22.</td>
<td>Search</td>
<td>Search, Go</td>
</tr>
<tr>
<td>23.</td>
<td>Shopping Indicator</td>
<td>Shopping Basket, Shopping Bag</td>
</tr>
<tr>
<td>24.</td>
<td>Order Indicator</td>
<td>View Basket, View Bag, Wish List</td>
</tr>
<tr>
<td>25.</td>
<td>Account Information</td>
<td>Account Information, Sign in, My Account, Your Account</td>
</tr>
</tbody>
</table>

Volume 3, Number 1, June 2008
6.2. Data Preparation

We collected the first or main page of two domain Web sites – i.e., a jewelry store and a book shop – from the search engine tools (Google directory) and Dmoz (open directory project). We performed with 100 trust Web sites and 100 untrust Web sites in each domain, according to trust criteria. The criteria for the untrust Web sites are not good in visualization or graphical design, security context in each Web site, and the lack of trust categories in Approach B. We filtered out the undefined Web sites from our corpus (i.e., only graphic page, flash, unidentified with text). From our experiment, we noticed that there were some different business types between the jewelry store domain and the book shop domain.

In the first, the jewelry store domain, we noticed that there were three business types in trust Web sites and untrust Web sites, as shown in Table 6.

<table>
<thead>
<tr>
<th>No.</th>
<th>Trust Web Sites</th>
<th>%</th>
<th>No.</th>
<th>Untrust Web Sites</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Focus luxury jewelry and diamond</td>
<td>50</td>
<td>1</td>
<td>Focus luxury jewelry and diamond</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Variety jewelry</td>
<td>10</td>
<td>2</td>
<td>Variety jewelry</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Special jewelry body and lower price jewelry</td>
<td>40</td>
<td>3</td>
<td>Special jewelry body and lower price jewelry</td>
<td>82</td>
</tr>
</tbody>
</table>

In the second, the book store domain, there were three business types in trust Web sites and four business types in untrust Web sites, as shown in Table 7.

<table>
<thead>
<tr>
<th>No.</th>
<th>Trust Web Sites</th>
<th>%</th>
<th>No.</th>
<th>Untrust Web Sites</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variety book selling</td>
<td>46</td>
<td>1</td>
<td>Variety book selling</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>Focused book selling</td>
<td>51</td>
<td>2</td>
<td>Focused book selling</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>Used book selling</td>
<td>3</td>
<td>3</td>
<td>Used book selling</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>Independent &amp; small press</td>
<td>8</td>
</tr>
</tbody>
</table>

It was necessary to prepare normal text data before starting in each approach; so, we used HTML Tag Removal to clean the Web document to normal text.
6.3. Experimental Results

We used the Weka machine learning tool for our experiments. A standard measurement for general classification performance is classification accuracy. The data were prepared in ARFF format (Weka’s input). Each classification technique was based on 10-fold-cross validation; that is, we divided data for the training and testing set from the same source file into 10 groups. Each group was divided into 9:1 by training: testing ratio. We had two domains for this evaluation -- the jewelry store Web sites and the book shop Web sites. The results of both approaches in the first domain are shown in Figure 13.

![Figure 13. Comparison of Three Classifiers with Two Approaches (Jewelry Store Domain)](image)

We compared each classification technique and approach with accuracy. The highest performance was an SMO algorithm, which yielded 80% accuracy on both approaches. The worst performance was a kNN algorithm, which yielded 55% accuracy on Approach A. We noticed that, by using the kNN and SMO algorithms, the accuracy of Approach A was lower than, or equal to, Approach B. On the other hand, the Naïve Bayes algorithm yielded a higher accuracy for Approach A than Approach B. Therefore, the best accuracy was the SMO on both approaches, but Approach A has 3,731 features and Approach B has 254 features.

The result of both approaches in the second domain is shown in Figure 14. We compared each classification technique and approach with accuracy. The
highest accuracy was an SMO algorithm on both approaches, which yielded 79.5% for Approach A and 83.5% for Approach B (EC-Word approach), respectively. The worst performance was a kNN algorithm, which yielded 60.5% for Approach A and 70% for Approach B. We noticed that, by using the kNN and SMO algorithms, the accuracy of Approach A was lower than that of Approach B. On the other hand, the Naïve Bayes algorithm yielded a higher accuracy on Approach A than on Approach B. Therefore, the best accuracy was the SMO algorithm for both approaches, but Approach A has 5,382 features and Approach B has 375 features.

![Figure 14. Comparison of Three Classifiers with Two Approaches (Book Shop Domain)](image)

In our comparative analysis of the two domains, we noticed that the best accuracy for both was the SMO algorithm. We observed that the number of features for Approach A of the book shop domain was higher than that of the jewelry domain equal to (5,382-3,731) 1,651, providing more information on the book shop than on the jewelry shop. We also noted that the number of features in Approach B for the book shop in the second domain was higher than those for the jewelry domain equal to (375-254) 121. This finding illustrated the discriminate e-commerce trust word (EC-word) was more than jewelry Web sites. The features ratio between Approach A and Approach B in the jewelry shop domain was 14.67 (3,731/254), and the features ratio between approach A and Approach B in the book shop domain was 14.35

*International Journal of Business and Information*
(5,382/375). Hence, we will adopt the SMO algorithm into our system in future research. For the Naïve Bayes algorithm, we noted that there were different results from the other algorithms, as mentioned previously.

7. CONCLUSIONS
In this paper, we adopted automatic classification algorithms to learn from the example Web site data sets. From our experiments using two approaches, we gained accuracy from the Support Vector Machine (SVM), based on Sequential Minimal Optimization (SMO) training techniques of at least 80% for both the jewelry store domain and the book shop domain. The best accuracy, using SMO, was 83.5% for Approach B (EC-word), with 375 features for the book shop domain. The SMO was very powerful in our experiments. In addition, we found that Approach B can work efficiently, saving both data storage space and rapid processing for WTA in the future.

In future research, we plan to develop and improve the accuracy of WTA with the new feature selection algorithm. At present, our proposed system can work efficiently with high quality in a textual content. Currently, we are developing the additional layout and link analysis for increasing the accuracy of the system.

REFERENCES


ABOUT THE AUTHORS

Banatus Soiraya received his master of science in information technology from Assumption University, Thailand. Currently, he is a Ph.D. candidate in the Department of Information Technology at King Mongkut's University of Technology, North Bangkok. He has researched and published papers on web mining. In addition, he is also interested in data networking and Internet security.

Anirach Mingkhwan received his Ph.D. (computer network) in computing and mathematical sciences from Liverpool John Moores University. He has researched and published in various fields, such as wireless network, sensor network, and information science. Currently, he is assistant professor of Industrial and Technology Management at King Mongkut's University of Technology, North Bangkok (Thailand).

Choocchart Haruechaiyasak received his Ph.D. in electrical and computer engineering from the University of Miami. He has researched and published in various fields such as search technology, data/text/Web mining, and information filtering and recommender system. Currently, he is chief of the Intelligent Information Infrastructure Section under the Human Language Technology Laboratory (HLT) at National Electronics and Computer Technology Center (NECTEC), Thailand.