

eCommerce Web-Site Trust Assessment Framework Based on Web Mining Approach

Banatus Soiraya
Faculty of Information Technology
King Mongkut's Institute of Technology North Bangkok 10800
banatuss@yahoo.com

Anirach Mingkhwan
Faculty of Industrial and Technology Management
King Mongkut's Institute of Technology North Bangkok 10800
anirach@ieee.org

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

Abstract

Despite the proliferation of the Internet and the growth of Web sites during this decade, eCommerce and other on-line business activities in Thailand remain relatively low compared to other activities such as email, chat and general Web page viewing. Two major reasons are Web site security and the lack of Web site trust. At present there is no effective tool which could perform trust assessment on Web sites and at the same time recommended information for improvement of those Web sites automatically. This paper introduces a novel trust assessment framework which performs evaluation on eCommerce Web sites based on Web Content mining techniques. We adopt two aspects of the Web Content mining concept for the framework: (1) Web content mining for analyzing Web-site contents such as text and other meta-data and (2) Web content mining for page layout detection of the Web sites. This framework will help on-line shoppers select and screen for reliable

and trusted eCommerce Web sites. At the same time, the web masters of these Web sites can benefit from using the recommendation automatically generated from the framework to improve their web sites for higher reliability and trust.

1. Introduction

Web sites provide people with a convenient way to disseminate information. Although the Internet has expanded tremendously during the past decade, the high reluctance for using eCommerce and on-line business activity still remains. Based on the previous surveys, two major problems are Web site security and the lack of Web site trust. Although most eCommerce Web sites provide some forms of secured payment method, it doesn't guarantee that the Web sites will gain better credibility. Based on many previous research and studies, there are many other significant factors which could influence the trust level of the Web sites such as Web-site contents and Web page

layout. W3 Trust Model (W3TM) proposed trust value calculation from meta data provided on the Web sites [9], however, it does not include the analysis on other important factors such as contents and layout analysis and more importantly 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 identify and determine the trustworthiness level of the Web sites correctly. In addition, it should provide recommended information based on trust assessment in order for Web masters to use as a guideline to improve the Web site for better trust.

We proposed an effective trust assessment framework for evaluating eCommerce Web sites based on Web mining techniques [8]. Two aspects of Web Content mining techniques are included in the framework: (1) Web content mining for analyzing Web-site contents such as text and other meta-data and (2) Web Content mining for page layout detection of the Web sites

2. Related Works

The Merriam-Webster English dictionary defines “trust” as assured reliance on the character, ability, strength, or truth of someone or something.

F. N. Egger et al [2] developed a model of trust in eCommerce called MoTEC (model of trust in eCommerce), which could classify characteristic of trust in eCommerce in terms of company, product and service, security, privacy, usability and relationship management. Moreover the significant factors and tools for a trust assessment in Web sites which can be illustrated via the following three methods: (1) Visual Design Factor [5, 10], (2) Context Design Factor [3] and (3) W3 Trust

Model (W3TM) [9] which assesses Web documents by using relevant meta data.

Jaideep Srivastava [8] defined Web mining *as* the application of data mining techniques to extract knowledge from Web data. In general, Web mining tasks can be classified into three categories: (1) Web content mining, (2) Web structure mining and (3) Web usage mining. Some of the well-known classification techniques for Web mining such as Naïve Bayes, kNN and SVM will be used in this paper [4, 6]. The most typical application for classification technique is Web page or Web site classification into predefined categories [4, 6, 7]. However, there are no Web site classifications by using attributes related to trust assessment features.

We need to analyse Web page by using visual information. This information represents Web page in terms of visual cue and semantic level. We can segment a page to different visual areas. There are several methods for representing visual structure of Web sites [1]. The most popular approaches are DOM-based segmentation, location-based segmentation, and Vision-based Page Segmentation (VIPS).

3. eCommerce Web-Site Trust Assessment Framework

Our proposed framework consists of two phases: (1) Content-based Analysis and (2) Decision Analysis. The major components will be described in details as follows (Figure 1).

3.1 Content-Based Analysis

This phase analyses a Web site by using both text analysis and layout analysis. The output is a content-based score which corresponds to the trust assessment level of the Web site. The initial step is the information gathering which retrieves and collects information from Web sites and then stores those Web pages into Web repository automatically.

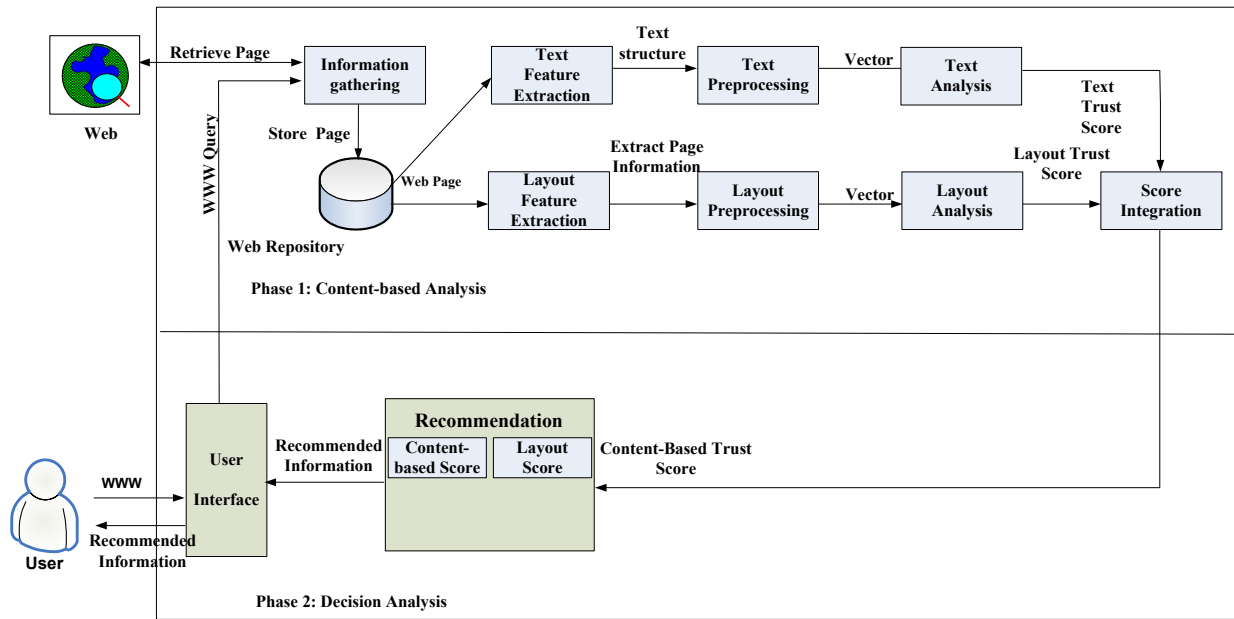


Figure 1. ECommerce Web-Site Trust Assessment Framework

3.1.1 Text Analysis:

The overall process for text analysis involves the following three steps.

a) Text Feature Extraction: This step parses and extracts Web page from Web repository for information under HTML document format Information which are extracted and used under our framework are, for example, title, meta-data and content.

b) Text Preprocessing: We separated this step into three processes:

(1) term extraction: this will tokenize text into a list of words. It then removes noisy information, i.e., stop words, HTML tags such as fonts, table and heading. The extracted terms are then formatted into a vector via the Vector-Space Model [6]. The weight which is assigned to each term is based on the TF-IDF approach.

(2) Identification of term creation and comparison. We adopt the MOTEC [2] and W3TM [9] for identifying the term features and rearranging them to identify context for trust. (3) Transformation. The tokenized terms from step (1) are checked against the

term features under step (2) and formatted into a vector with the TF-IDF weight assignment for the content-based analysis. These three processes are illustrated in Figure 2.

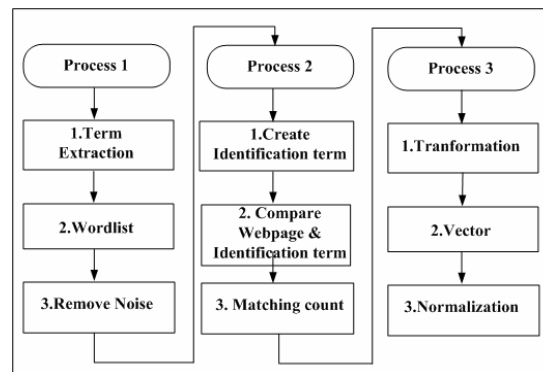


Figure 2. Text Preprocessing Process

c) Text Analysis: This step uses text classification techniques to classify Web pages. The vectors which are preprocessed from text preprocessing step is input into this step. The result is text trust score. We will perform evaluation in order to select the high-performance classification algorithms from many well-known techniques such as

Naïve Bayes, kNN and SVM [4,6]. This step is shown in Figure 3.

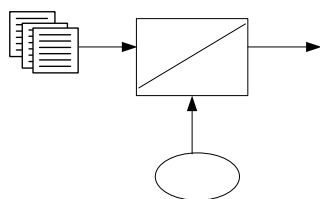


Figure 3. Text Analysis

3.1.2 Layout Analysis

This component will analyze web sites by layout-based analysis. And then it will give layout score. The overall process for Layout analysis involves three steps, Layout Feature Extraction, Layout Preprocessing and Layout Analysis.

a) Layout Feature Extraction: After Web Page is retrieved from Web repository, it will go through layout detection and classification based on our predefined characteristics. These characteristics include (1) the distrusted Web sites were more image-base than the trusted sites [5]. (2) layout positioning .The formal layout such as Newspaper theme will give more credibility than informal layout [5]. Another factor to consider is advertisements which should be placed in appropriate position. Advertisements should not be put and mix up with the useful content. (3) The noise information extraction. We should be able to detect advertisements in both normal and abnormal patterns such as flash animation, GIF animation.

b) Layout Preprocessing: This step is similar in text preprocessing which starts from First Process (Remove Noise) to Third Process (Vector and Normalization). These three processes are illustrated in Figure 4.

c) Layout Analysis: This step uses classification technique to classify Web pages. The vector obtained from layout preprocessing will be sent to this step. The

result is the layout trust score. We will perform evaluation based on some well-known classification algorithms such as Naïve Bayes, kNN and SVM [4,6].

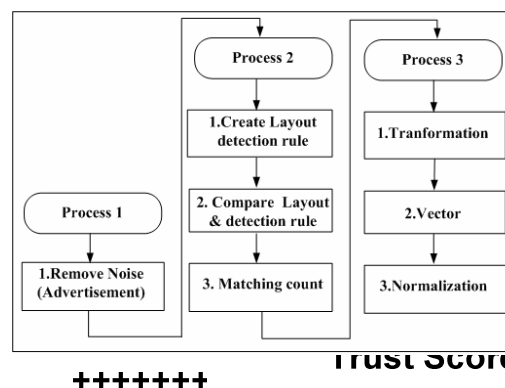


Figure 4. Layout preprocessing process

This step is shown in Figure 5. We adopt two techniques, i.e., page segmentation and (Related with Identification term) for this step.

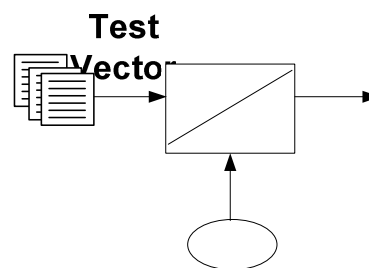


Figure 5. Layout Analysis

-Score Integration: After the end of two processes, text trust scores and layout trust scores will be merged together. Empirical experiments will be performed to evaluate the appropriate weights from both scores for maximum trust assessment accuracy. Then the equation is defined as:

$$\text{Score Integration} = \text{Weight Text} * \text{Text Trust Score} + \text{Weight Layout} * \text{Layout Trust Score}$$

Under condition (Weight Text = 1- Weight Layout, $0 \leq \text{Weight Layout} \leq 1$)

3.2 Decision Analysis

This phase uses the total trust score generated by previous phase. Two processes are then performed: (a) Trust Criteria and (b) Recommendation for improve this Web site. Two processes are illustrated as follows.

(a) Trust criteria for categorizing Web sites, we consider four trust levels:

- High Trust Web sites
- Moderate Trust: Web sites
- Low Trust Web sites
- UnTrust Web sites

(b) Recommendation for Improving the Web site: For example, the following messages could be given to the Web site administrator.

-Your Web site lacks of privacy policy, please provide the information for improving the Web site credibility.

-Your Web site layout looks unprofessional and there are too many advertising banners, please arrange your layout and reduce the advertising banners.

4. Conclusion and Future Works

This paper presents a Web mining framework for trust assessment on eCommerce Web sites. We adopt Web mining techniques in the framework for analysing Web-site contents such as text and page layout. Our framework consists of two phases: (1) content-based analysis and (2) decision analysis. This framework gives to users two benefits: (1) criteria and trust percentage on a specified web site and (2) recommendation for improving this web site according to content, layout guidelines. The future works include the study of algorithms for classifying eCommerce web sites, text-based and layout classification algorithms, empirical experiment should be done by maximize trust accuracy.

5. References

- [1] R. Burget (2006). "Visual Structure of Web Documents."
http://www.fit.vutbr.cz/research/pubs/TR/2006/sem_uifs/s060313slidy2.pdf
- [2] F. N. Egger (2003). "Designing the Trust Experience for Business to Consumer Electronic Commerce," Ph. D. Thesis.
- [3] S. G.-K. Ewald A. Kaluscha (2003). "Towards A Pattern Language for consumer trust in electronic commerce," EuroPLop 2003 ,the 8th European Conference on Pattern Languages of Programs, pp. C9-1-C9-16.
- [4] M. S. Hans-Peter Kriegel (2004). "Classification of Websites as Sets of Feature Vectors," Proceeding of the IASTED International conference DATABASES and Application, pp. 127-132.
- [5] J. P. Kristiina Karvonen (2001). "SIGNS of Trust: A Semiotic study of Trust formation in the web," First International Conference on Universal Access in Human-Computer Interaction(Vol. 1, pp. 1076–1080). Mahwah, NJ: Erlbaum.
- [6] A. A. Parham Moradi, Mohammad Ibrahim Shiri (2006). "Novel Method for Improving Web Text Classifiers Performance Through Machine Learning," IEEE, Vol. 0-7803-9521, pp. 534-539.
- [7] J. M. Pierre (2000). "Practical Issues for Automated Categorization of Web Sites," ECDL 2000 Workshop on the Semantic Web, Lisbon, Portugal ,21 September.
- [8] J. Srivastava, "Web Mining: Accomplishments & Future Directions." <http://www.cs.umn.deu/faculty/srivasta.html>
- [9] Y. Yang (2004). "W3 Trust Model (W3TM) "A Trust-Profiling Framework to Asses Trust and Transitivity of Trust of Web-Base Service in A

Heterogeneous Web Environment,” Phd Thesis.

- [10] H. H. E. Ye Diana Wang (2005). “An overview of online trust: Concepts, elements, and implications,” *Computer in Human Behavior* :ELSEVIER, vol. 21, pp. 105-125.