Leveraging Website Genre and Structure Information for Fake Website Detection

Ahmed Abbasi
University of Wisconsin-Milwaukee
abbasi@uwm.edu

Abstract

In this study we assessed the efficacy of using website genre composition and design structure information for fake website detection. A genre tree kernel was proposed that creates a rooted tree from the website file directory structure, and labels the tree's file nodes with genre information. The genre tree kernel was compared against several benchmark kernel and non-kernel methods that utilized a rich feature set comprised of thousands of website content-based attributes. Experimental results revealed that the genre tree kernel outperformed all comparison methods on a test bed encompassing 900 legitimate, concocted and spoof sites. The results suggest that fake website detection systems could benefit from the use of genre and design structure information.

Keywords: Fake website detection, website genres, kernel-based methods

1. Introduction

Fake websites generate billions of dollars in fraudulent revenue by exploiting human vulnerabilities in online settings (Zhang et al., 2007). The most common types of fake websites used to target end users are concocted and spoof sites. Spoofs are imitations of existing commercial websites used for online identity theft (Dinev, 2006). Concocted sites are deceptive websites attempting to appear as unique, legitimate online entities, with an objective of failure-to-ship fraud (Abbasi et al., 2010). Recent studies have demonstrated the inability of existing detection tools and techniques to adequately combat concocted and spoof websites (Zhang et al., 2007; Abbasi et al., 2010). In this study, a new kernel-based approach for fake website detection is proposed. The approach centers around a novel genre tree kernel that leverages (1) website genre composition; (2) website design structure differences between legitimate and fake sites; for faster and more accurate fake website detection capabilities than comparison techniques. The results, which demonstrate the efficacy of using website genre and design structure information, have important implications for the design of future fake website detection systems.

2. Website Genre Composition and Design Structure

Document genres are a combination of purpose and form (Orlikowski and Yates, 1994; Roussinov et al., 2001). Website genres include homepages, product pages, search pages, frequently asked question (FAQ) sections, testimonials, newsletters, status and tracking pages, educational materials, publications, etc. (Roussinov et al., 2001). Each of these genres has a distinct and socially recognizable purpose (Orlikowski and Yates, 1994). For instance, testimonials are intended to increase credibility while newsletters and educational materials convey important information and knowledge. Given their objective, fake website developers only wish to present the appearance of legitimacy. Consequently, fake sites often differ from legitimate ones in terms of their website genre composition. For instance, fake sites often fail to incorporate substantial FAQ sections or membership and login pages (Abbasi and Chen, 2009). Conversely, fake websites include an abundance of customer testimonials in order to gain users' trust (Grazioli and Jarvenpaa, 2000). Nevertheless, website genre information has not been utilized in prior detection techniques.

Fraudsters' use of automated website development tools also results in design structure similarities between fake sites (Abbasi et al., 2010). For instance, spoof sites often have more levels/depth than legitimate websites, as indicated by the number of slashes "/" in these sites' web pages' URLs (Dinev, 2006). In contrast, concocted websites tend to be relatively flatter, with web pages concentrated in a few levels (Abbasi and Chen, 2009). Moreover, prior analysis has revealed that web pages at different levels also differ in terms of their quality, content, and genres (Ester et al., 2002); with lower/bottom level pages providing greater discriminatory potential for fake website detection (Abbasi and Chen, 2009). Information about a website's page levels can be derived from URL tokens and the file directory structure. The latter can also shed light on the location of key design-related files (e.g., images, banners, logos, scripts), which are often useful identifiers (Dinev, 2006).

3. The Genre Tree Kernel for Fake Website Detection

For complex structure information and problem-specific characteristics that cannot be described by standard feature vectors, kernel-based methods provide an effective alternative. Custom kernels have been used in recent document categorization and fake website detection work (Li et al., 2009; Abbasi et al., 2010). We propose a genre tree kernel that combines website genre and design structure information.

3.1 Genre Tree Construction

Trees are constructed by traversing the websites' file directories (i.e., folders), beginning with the root directory. All files and folders contained in the root directory are considered its child nodes, and are added to the tree with a label that corresponds to their file/folder name. Any child node folders (i.e., subfolders of the parent node) are also added to the traversal queue. The traversal and addition process is repeated until the contents of all subfolders have been added to the tree. Formally, the construction process results in a labeled rooted tree T with nodes $\{t_0 \ldots t_n\}$, where t_0 is the site's root directory and each node t_i has a label $v(t_i)$. We use $p(t_i) \in T$ to represent the parent node of t_i, while $c(t_i) \subset T \backslash \{t_0\}$ represents the set of children of the node t_i with cardinality $|c(t_i)|$ for all i>0. Once the tree has been constructed the nodes are relabeled. Indexable file nodes are labeled with genre information. We utilized twelve website genres, most of which have been described in prior genre analysis studies (Roussinov et al., 2001, Rosso, 2008). These are listed in Table 1.

Table 1: List of Website Genres Utilized by the Genre Tree Kernel

Genre	Label	Description
About	A	Information about the company, including history and background.
Contact	C	Comment posting, emailing/speaking with representatives, and live chat.
FAQ	Q	Frequently asked questions.
Homepage	H	The website's starting page.
Login	L	Login, logout, password retrieval, new member registration, etc.
Newsletters	N	Articles, newsletters, and other informational resources.
Order	O	Order and shopping cart information, including order tracking and shipping.
Policy	P	Policies, terms, guarantees, and privacy notes.
Price	D	Fees, rates, prices, and quotes.
Product	R	Description of products and services.
Search	S	Search and navigation pages, including site maps and directories.
Testimonial	T	Customer testimonials.

Prior website genre classification studies attained good results when using a page's URL tokens (Lim et al., 2005). For a given indexable file node $t_i \in T \backslash \{t_0\}$, the genre classification is performed by analyzing $v(t_i)$ and $v(p(t_i))$; the node's filename and the node's parent folder's name. The two strings $v(p(t_i))$ and $v(t_i)$ are concatenated, tokenized, and stemmed. The set of stemmed tokens is compared against semi-automatically learned sets of key words associated with folder/file names belonging to the 12 aforementioned genres. The key word sets were automatically learned from a training set comprised of over 1,000 legitimate and fake websites using the information gain heuristic, and then manually refined for improved accuracy. The indexable files' genres are assigned using a simple token matching scheme, where they are categorized as belonging to the genre with the most matches. The formulation is presented in Figure 1. Image and folder files are relabeled with ‘I’ and ‘F’, respectively. All remaining (i.e., unidentified) files are relabeled with an ‘X’.

Let $A(t_i)=\{a_1...a_v\}$ denote the stemmed tokens taken $v(p(t_i))$ and $v(t_i)$,
Let $B(j)=\{b_1...b_m\}$ denote the key word set associated with genre j,
where $C(j)$ is the label associated with j, as described in Table 1.
The number of matches between t_i and j is computed as follows:

$$s(t_i,j)=\sum_{k=1}^{v}\sum_{p=1}^{m}x(a_k,b_p)$$

$$\text{where } x(a_k,b_p)=\begin{cases}1, \text{ if } a_k=b_p\\ 0, \text{ if } a_k\neq b_p\end{cases}$$

And t_i is assigned to the genre with the greatest number of matches:

$$v(t_i)=\begin{cases}C\left(\underset{j}{\arg\max}\, s(t_i,j)\right), \text{ if } \sum_{j=1}^{12}s(t_i,j)>0\\ \text{X , otherwise}\end{cases}$$

Figure 1: Genre Tree Indexable Node Labeling Mechanism

Websites often vary considerably in terms of their size. In order to improve the accuracy of comparisons, as well as computation times, website pages are often pruned (Ester et al., 2002). One common pruning strategy is to limit the maximum number of pages associated with a particular label. Prior work on topic-based website categorization pruned pages containing duplicate topical information (Ester et al., 2002). Since certain website genres are more prevalent in terms of their occurrence frequency, we use a genre pruning parameter *g*. For a given node $t_{i,}$ *g* indicates the maximum number of child nodes in $c(t_i)$ that can be labeled as belonging to a particular genre *j*. Note that *g* only limits the number of non-folder child nodes.

3.2 Genre Tree Traversal and Comparison

Random walks provide a useful mechanism for traversing graph and tree structures. They have been used in prior work on graph kernels (Li et al., 2009). The genre trees are traversed using a series of random walks. Beginning with the root directory node t_0, the random walk q_x has a $(|c(t_0)|+1)^{-1}$ probability of selecting any $t_i \in c(t_0)$ or terminating. In other words, if t_0 has three child nodes, they each have a ¼ probability of being selected, while the random walk termination probability is also ¼. If the walk is not terminated, $q_x = (t_0,t_i)$ and from t_i, the random walk has a $(|c(t_i)|+1)^{-1}$ probability of selecting any $t_k \in c(t_i)$ or terminating. Hence, if $c(t_i) = \emptyset$, the probability of termination is 1. The random walk continues traversing the tree in a top-down manner until it is terminated. The process is repeated until *w* random walk paths have been generated. The genre trees from any two websites are compared based on these *w* paths Figure 2 shows the formulation of the genre tree comparison used to generate the kernel matrix *K*.

Let $\{q_1 \ldots q_w\}$ and $\{r_1 \ldots r_w\}$ represent the set of random walks along genre trees T and T'

$$K(T,T')=\sum_{k=1}^{w}\sum_{p=1}^{w}\frac{L(q_k,r_p)M(q_k)M(r_p)}{w}$$

where:

$$L(q_k,r_p)=\begin{cases}1, \text{ if } \left(v(q_{k1})...v(q_{kh})\right)=\left(v(r_{p1})...v(r_{ph})\right)\\ 0, \text{ otherwise}\end{cases}$$

$$M(q_k)=\begin{cases}1, \text{ if the path } q_k \text{ has not yet been matched to any of the paths of } T'\\ 0, \text{ otherwise}\end{cases}$$

h is the length of x, and the length of y.

Figure 2: Formulation of Genre Tree Comparison

4. Research Test Bed and Design

The training data set was comprised of over 1,000 legitimate, concocted, and spoof websites. A separate test bed of 900 websites (200 legitimate, 350 concocted, and 350 spoof) was used for evaluation. The spoof website URLs were taken from online repositories such as Phishtank.com. The concocted website URLs were taken from online databases such as Artists-Against 4-1-9. The 200 legitimate websites included ones that are commonly spoofed, as well as those belonging to genres relevant to the concocted website test bed. All websites were collected using automated spidering programs that fetched the website pages, images, and link information, while preserving the websites' file directory structures.

We compared the genre tree kernel against kernels used in previous fake website detection studies. The comparison kernels included the linear composite kernel proposed by Abbasi et al. (2010), as well as the standard linear, radial basis function (RBF), and polynomial kernels, all of which have worked well in prior studies on concocted, spoof, and web spam sites (Drost and Scheffer, 2005; Abbasi et al., 2010). The kernels were run using the Support Vector Machines (SVM) classifier. We also evaluated several non-kernel-based classification methods used in prior work, including logistic regression, Bayesian network, J48 decision tree, neural network, and naïve Bayes. All comparison methods were run using a feature set comprised of over 5,000 attributes derived from the websites' body text, source code, URL tokens, images, and linkage-based information (Drost and Scheffer, 2005; Abbasi and Chen, 2010). These features were learned from the training data set, using the information gain heuristic. The genre tree kernel's w and g parameters were tuned using cross validation on the training data, and were set to w=30 and g=10. All comparison methods underwent extensive parameter tuning in order to ensure the best possible results for these classifiers.

5. Experimental Results

The evaluation metrics employed included those used in prior research; overall accuracy and class-level precision, recall, and F-measure (Drost and Scheffer, 2005; Abbasi et al., 2010). Table 2 shows the experimental results. The genre tree kernel outperformed all comparison methods (kernel and non-kernel based) in terms of overall accuracy and class-level f-measure, precision, and recall for real, concocted, and spoof websites. The performance gain in terms of overall accuracy was approximately 5% over the best kernel method (linear composite kernel). Moreover, the genre tree kernel outperformed the best non-kernel method, logit regression, by over 8%. An important factor contributing to the genre tree kernel's enhanced performance was its ability to better detect concocted websites; it outperformed comparison techniques by at least 7% in terms of concocted recall. It also improved legitimate and spoof detection rates by at least 3% (based on the class-level recall values). Consistent with prior work, the concocted website detection rates were lower as this is considered a more challenging task as compared to spoof detection (Abbasi et al., 2010).

We constructed Receiver operating characteristic (ROC) plots, showing the tradeoffs between true and false positives/negatives (here true positives refer to correctly classified fake websites). In order to assess the impact of different parameter settings on the genre tree kernel's performance, we ran various combinations of values for w and g. Once again, the single best parameter setting was used for all 10 comparison methods. Figure 3 shows the ROC plots. Plots closer to the top left corner signify better results, since they denote high ratios of true to false positives (or negatives). The genre tree kernel had the best results for all parameter settings analyzed. The genre tree kernel's false positive and false negative rates ranged from 0%-2% and 2%-5%, respectively, with the highest error rates attained when g = 1. Several settings yielded 100% detection rates on the legitimate and spoof sites. Overall, the ROC analysis results suggest that the genre tree kernel's performance is fairly robust across different parameter settings.

Table 2: Experimental Results for Genre Tree Kernel and Comparison Methods

Learning Technique	Overall Accuracy (n=900)	Real Websites (n=200)			Concocted Detection (n=350)			Spoof Detection (n=350)		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Genre Tree	**97.78**	95.21	91.28	99.50	97.22	99.70	94.86	99.71	99.71	99.71
Linear Composite	92.56	85.21	76.29	96.50	91.82	97.74	86.57	97.12	97.97	96.29
Linear	90.78	82.00	72.42	94.50	90.61	96.45	85.43	95.36	96.76	94.00
Polynomial 2nd	90.44	81.38	71.75	94.00	90.30	96.13	85.14	95.07	96.47	93.71
RBF	90.22	81.20	70.90	95.00	90.41	96.74	84.86	94.89	97.02	92.86
Polynomial 3rd	89.89	80.60	70.26	94.50	88.58	96.31	82.00	95.96	96.80	95.14
Logit Regression	89.00	78.53	69.36	90.50	90.02	94.08	86.29	92.58	94.36	90.86
J48 Decision Tree	88.77	75.66	73.01	78.50	88.82	87.95	89.71	90.98	88.41	93.71
Bayesian Network	88.56	77.27	69.18	87.50	88.72	92.28	85.43	92.55	92.82	92.29
Naïve Bayes	77.67	63.12	49.86	86.00	86.49	91.14	82.29	77.47	89.51	68.29
Neural Network	66.22	54.21	38.79	90.00	70.63	90.99	57.71	73.28	91.45	61.13

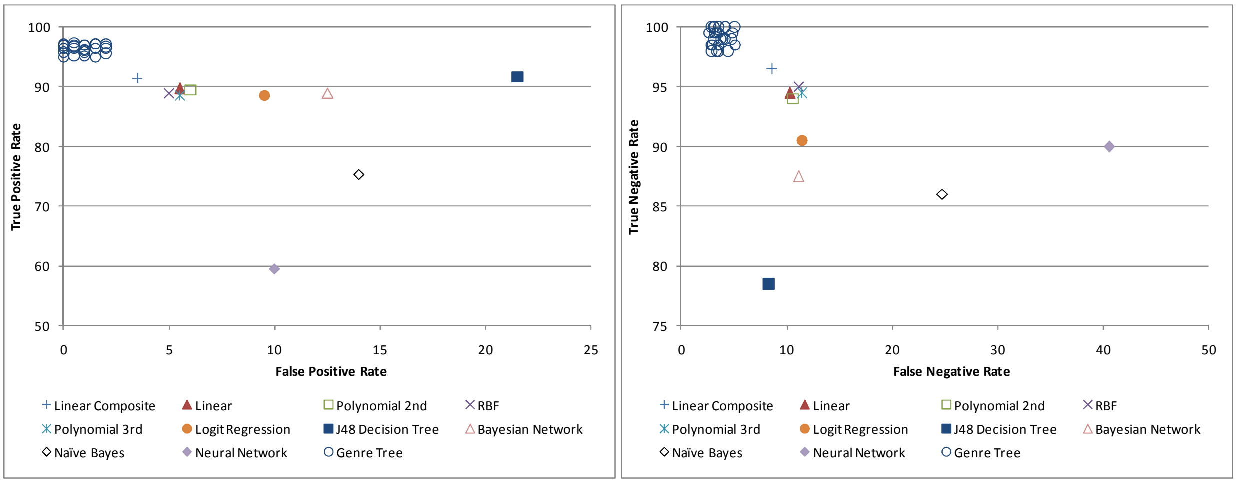

Figure 3: ROC Plots for Various Genre Tree Parameter Settings and Comparison Methods

6. Conclusions and Future Directions

In this study, we investigated the effectiveness of using website genre and design structure characteristics for enhanced fake website detection. The proposed genre tree kernel outperformed several content-based classifiers in terms of overall accuracy and class-level detection rates. In addition to accuracy, computation times are another important consideration. By employing an efficient URL token-based genre labeling mechanism, the genre tree kernel was able to detect fake websites in a computationally faster manner than the comparison content-based methods which require the extraction of thousands of text, image, and linkage attributes. On average, the genre tree kernel took 1.5 seconds per website while the comparison methods each took at least 3 seconds. In our future work, we plan to conduct a more detailed analysis of the genre tree kernel's computation times relative to comparison approaches. We also intend to compare the genre tree kernel against additional kernels and existing detection systems. Additionally, we plan to further investigate the impact of different parameter settings.

7. Acknowledgements

This research has been supported in part by the following U.S. National Science Foundation grant: CNS-1049497 “A User-Centric Approach to the Design of Intelligent Fake Website Detection Systems,” October 2010 – September 2012.

References

Abbasi, A. and Chen, H. “A Comparison of Fraud Cues and Classification Methods for Fake Escrow Website Detection,” *Information Technology and Management*, 10, 2009, pp. 83-101.

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker Jr., J. F. “Detecting Fake Websites: The Contribution of Statistical Learning Theory,” *MIS Quarterly*, 34(3), 2010, pp. 435 - 461.

Dinev, T. “Why Spoofing is Serious Internet Fraud,” *Communications of the ACM*, 49(10), 2006, pp. 76-82.

Drost, I. and Scheffer, T. “Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam,” *In Proceedings of the European Conference on Machine Learning*, 2005, pp. 96-107.

Ester M., Kriegel H., and Schubert M. “Web Site Mining: A New Way to Spot Competitors, Customers, and Suppliers in the World Wide Web,” *In Proceedings of the 8th ACM SIGKDD*, 2002, pp. 249–258.

Grazioli, S. and Jarvenpaa, S. L. “Perils of Internet Fraud: An Empirical Investigation of Deception and Trust with Experienced Internet Consumers,” *IEEE Transactions on Systems, Man, and Cybernetics Part A,* 20(4), 2000, pp. 395-410.

Li, X., Chen, H., Zhang, Z., Li, J., and Nunamaker Jr., J. F. “Managing Knowledge in Light of its Evolution Process: An Empirical Study on Citation Network-based Patent Classification,” *Journal of Management Information Systems*, 26(1), 2009, pp. 129-153.

Lim, C. S., Lee, K. J., and Kim, G. C. “Multiple Sets of Features for Automatic Genre Classification of Web Documents,” *Information Processing and Management*, 41, 2005, pp. 1263-1276.

Orlikowski, W. J. and Yates, J. “Genre Repertoire: The Structuring of Communicative Practices in Organizations,” *Administrative Sciences Quarterly*, 33, 1994, pp. 541-574.

Rosso, M. A. “User-Based Identification of Web Genres,” *Journal of the American Society for Information Science and Technology*, 59(7), 2008, pp. 1053-1072.

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., and Liu, X. “Genre Based Navigation on the Web,” *In Proceedings of the 34th Hawaii International Conference on Systems Sciences*, 2001.

Zhang, Y., Egelman, S., Cranor, L. and Hong, J. “Phinding Phish: Evaluating Anti-phishing Tools,” *In Proceedings of the 14th Annual Network and Distributed System Security Symposium* (NDSS), 2007.