http://doi.org/10.35784/iapgos.5572 *received: 15.11.2023 | revised: 16.01.2024 | accepted: 11.03.2023 | available online: 31.03.2024*

REVIEW OF MODELLING APPROACHES FOR WEBSITE-RELATED PREDICTIONS

Patryk Mauer
Opole University of Technology, Opole, Poland

Abstract. *This paper researches various modelling approaches for website-related predictions, offering an overview of the field. With the ever-expanding landscape of the World Wide Web, there is an increasing need for automated methods to categorize websites. This study examines an array of prediction tasks, including website categorization, web navigation prediction, malicious website detection, fake news website detection, phishing website detection, and evaluation of website aesthetics.*

Keywords: machine learning, web sites, prediction methods, classification algorithms

PRZEGLĄD PODEJŚĆ DO MODELOWANIA PRZEWIDYWAŃ ZWIĄZANYCH Z WITRYNAMI INTERNETOWYMI

Streszczenie. *Ten artykuł naukowy przeprowadza analizę różnorodnych metod modelowania stosowanych do prognozowania aspektów witryn internetowych, zapewniając przegląd tej dynamicznie rozwijającej się dziedziny. Podczas gdy Internet nieustannie się powiększa, nabiera wagi potrzeba stosowania automatycznych metod do klasyfikacji nowo powstających stron internetowych. Zbadano metody zastosowane w szerokim zakresie przewidywań, obejmujących kategoryzację witryn internetowych, prognozowanie zachowań nawigacyjnych użytkowników online, identyfikację stron o złośliwym charakterze, wykrywanie fałszywych informacji, rozpoznawanie prób phishingu oraz ocenę estetycznych aspektów witryn internetowych.*

Słowa kluczowe: uczenie maszynowe, witryny internetowe, metody przewidywania, algorytmy klasyfikacji

Introduction

The estimated total number of existing websites in the World Wide Web exceeded 2 billion stated by Statsfind.com [4]. The same platform states that there are 252 thousand websites created everyday worldwide which gives 175 websites every minute. These statistics only showcase the need of automated ways to indexing and categorization of these websites. Content categorization falls into use cases of machine learning or probabilistic models as they can be trained to automatically categorize or classify content into predefined categories based on the content's characteristics, features, or patterns. Techniques needed for achieving a success in these fields have been developed by researchers generally in purpose of cracks and malicious websites detection [2], web navigation prediction [5], fake news detection [1], Search Engine Optimization [11]. All the purposes of websites or generally content classification utilized different methods of data acquisition and preprocessing, feature extraction and machine learning algorithms. This paper covers the need of a structured review of the techniques and approaches used particularly for website categorization.

1. Modelling approaches

The following chapters are structured in the way that each point in the chapters number lists corresponds to a particular approach used by researchers. The numeration is consistent throughout the paper. The table 1 shows the corresponding data.

2. Data acquisition and preprocessing

Data acquisition and preprocessing phases differed significantly from each other and are presented in the following subchapters.

2.1. Data sources and sizes

The data in the examined research papers was gathered from the following sources in the following sizes:

1) Manually labelled set of 450 web pages that was uniformly distributed among six categories. In total, the pages contained 3890 images.
2) Publicly available web log datasets that include Cyber Threat Intelligence (CTI) Dataset: DePaul University web server logs from April 2002, with 13,745 sessions, 683 pages, and 16 categories; Microsoft Anonymous Web Data (MSWEB) Dataset: Microsoft logs from February 1998 with 38,000 random users and 294 unique Vroots (virtual directories – a feature that allows the web server to serve content from a directory other than the server's root directory); BMS WebView1 Dataset: E-commerce website logs from Gazelle.com with 59,601 sequences and 497 distinct items.
3) The experimental data comprised: (i) the "UK Web Archive" that contributed an initial dataset consisting of 14922 categorized websites. (ii) A manually gathered list of 510 URLs containing malware content to enhance the dataset's coverage of malicious sites. (iii) Dataset created utilizing the "Google Safe Browsing" API to classify and label websites, including the identification of "Malware" and other safety-related categories.

Table 1. Corresponding numbers of points in numbered lists to researchers' papers and goals of the prediction. Indication on how to read the further parts of the paper

No.	Goal of the prediction	Researchers
1	Website categorization – the proces of assigning a website to predefined categories.	Nandanwar A., Choudhary J., 2020 [8]
2	Web Navigation Prediction – the anticipation of the next web page or resource that a user is likely to access during their online browsing session.	Jindal H., Sardana N., 2022 [5]
3	Crack and Malicious Website detection – the forecast of whether a website is engaged in distributing illegal software.	Cohen D., Naim O., Toch E., Ben-Gal, 2021 [2]
4	Fake News Website detection – the classification of whether a website falls under the category of authentic or fabricated news providers.	Bozarth L., Budak C., 2019 [1]
5	Phishing Website detection – the categorization of a website as either legitimate or a potential phishing threat	Opara C., Chen Y., Wie B., 2023 [7]
6	Evaluation of website aesthetics – the prediction of the mean subjective user score regarding the design aesthetics of a website.	Delitzas A., Chatzidimitriou K.C., Symeonidis A.L., 2023 [11]
7	Web Page Classification into low, medium and high degree of adjustment to Search Engine Optimization (SEO) guidelines.	Matošević G., Dobša J., Mladenić D., 2021 [10]
8	Predicting users' intention of potential online purchase.	Sakar C.O., Polat S.O., Katircioglu M., 2018 [9]

4) The data source comprises (i) Lists of Fake and Traditional News Sites: Collected by Bozarth et al. in 2019, resulting in 1800 unreliable news sites and 8200 reliable news sites (ii) Homepage and Subpages: Collected using scrapy for 7600 homepages, followed by recursive crawling of subpages, resulting in 2 million distinctive subpages. (iii) News Article Webpages and Tweets: obtained from Bozarth et al. (2019), including 170 thousand unique articles shared in 700 thousand Tweets.
5) Data sources encompass Alexa.com for genuine websites, Phishtank.com for fraudulent web pages, and a set of 22 thousand web pages collected in Korkmaz et al. (2020) [6].
6) A publicly accessible dataset collected by researchers in 2014 includes 398 webpage screenshots with aesthetics ratings on a scale of 1 to 9. Additionally, a second dataset, generated by the authors, consists of 100 webpages randomly selected from Alexa's top 5000 websites and includes pairwise comparisons of web designs obtained through crowdsourcing via a web application.
7) The data in the research is a random sample of 600 pages extracted DMOZ – open directory of web pages maintained by volunteers. These pages were manually classified by three independent SEO experts into low, medium and high adjustment level to SEO guidelines. A Python script was used to extract website features and keywords.
8) The data was curated with emphasis on its diversity in order not to show any tendency related to marketing campaigns or particular days. The dataset consisted of over 12 thousand user navigation sessions with restriction that each session belong to a unique user. Over 15% of the dataset contained user sessions that ended with a purchase. Ten features were numerical that mainly consisted of spent time on defined websites, number of visited pages and bounce and exit rates. The eight categorical features contained information about the user agent, geographical location, and data if the user is a new or returning one.

2.2. Data types and categories

The data collected from previously mentioned sources appeared in the following formats and meaning:

1) Numerical data: The patterns from extracted images from manually labelled websites were used as feature vectors for training classifier models.
2) Text data, Numerical data: Web Log Files – a unit of user activity on a website usually characterized by a sequence of interactions between a web browser and a web server during a specific time period. These web log session files are primarily text-based and include numerical data for various attributes like session size, average session duration, as well as counts of unique items and Vroots.
3) Numerical data: Website design features that encompass attribute values of all elements, both visible and invisible, that are present on a webpage. This data was acquired by employing a web scraper, which was fed with a list of URLs.
4) Text data: Content extracted from homepages and subpages, content of news article webpages, text in Tweets. Categorical data: Lists of websites categorized as either fake news or traditional news, classification of domains as having little content or being for sale, classification of URLs according to the domains in the news sites lists.
5) Text data: Raw URLs and HTMLs were collected and processed to extract relevant information such as metadata, text content, and structural elements. Raw URLs were parsed to identify domains, paths, and parameters, while HTML documents were analysed to extract text, tags, attributes, and the hierarchical structure of web pages.
6) The data types in the paper include numerical data (user ratings, standard deviations), categorical data (user IDs), binary data (pairwise comparisons, control questions), image data (webpage screenshots), and temporal data (timestamps).
7) Numerical data – all 21 independent features, mainly html tags and keywords, were described as numbers of occurrence or length or word/character count. The only categorical data was a dependent variable representing a level of adjustment to SEO guidelines.
8) Numerical data – statistic of users' activity in terms of spent time on website and bounce rate. Text data – the user identification information, including the operating system type, geographical location and traffic source as well as the visitor type determined as either new or returning one.

To summarize, the researchers made their predictions based on the following web related elements : Images, Web Log Files, Web elements and their attributes, raw HTMLs and URLs, user data and user activity, timestamps, Packet Capture Files. The data used in by them can be assigned to a particular goal of the prediction what is depicted in the table 2.

Table 2. Association between data categories and their respective served prediction purposes

Goal of the prediction	Data categories
Website categorization	Images, Web elements and their attributes, Text on websites, Raw HTML and URL, User ratings, Webpage screenshots.
Web Navigation Prediction	Web Log Files, Statistic of users' activity, User identification information.
Evaluation of website aesthetics	User ratings, Webpage screenshots.
Detection of encrypted malicious activity	Packet capture files.
Predicting users' intention	Statistic of users' activity, User identification information.

2.3. Preprocessing

The collected data often contains noise, irrelevant content, might be in an unsuitable format for processing. Preprocessing steps the referenced literature included are:

1) Elimination of images utilized for non-page-descriptive purposes, especially for advertisements and navigation items. The feature vectors were derived from these images using a pre-trained Convolutional Neural Network (CNN) model.
2) Removal of HTTP requests for files, differentiating the traffic to user and spider generated, organizing user interactions into sessions, identifying sessions (as time-stamped user click-streams) and users (assigning unique user IDs to web sessions), generating representations of user navigation patterns, and classifying links between these patterns. This preprocessing was performed to prepare the data for building a Markov Model.
3) The values of the collected attributes, such as element areas, text lengths and rates, colours, etc., were calculated. These values were then subjected to statistical functions including maximum, minimum, summation, average, count, colour classification, standard deviation, normal distribution, interquartile range, and normalization.
4) Domains with little content and links directing to external websites were filtered out. URL Characteristics: suffix count, domain length, top-level domain. Homepage Auxiliary Data: Quality, content, and specific links. Homepage Style and Scripts: HTML tags for scripts, style, metadata. Homepage Link Categories: Routine links. Homepage HTML Tags and Path: DOM tree elements, tag counts, characteristics. Network Characteristics: Structure, connectivity metrics. Motifs: Recurring subgraph patterns.
5) Removal of HTTP:// and HTTPS:// prefixes from the URLs. Creation of dictionary of words out of the HTML documents, where all punctuation characters were treated as separate tokens. Then word-level corpus of HTMLs and character-level corpus of URL were tokenised to a one-dimensional digital vector. Equalization of the length of the URLs – filling the lacking characters with a token of no significance. Concatenation of URL and HTML embedded matrices into a two-dimensional matrix.

6) As the crowdsourced data was generated as pairwise comparison, the date preprocessing phase included creating a count matrix that captures the number of times one webpage is preferred over another. Then by applying the Bradley-Terry model the authors estimated the aesthetics ranking out of the calculated probabilities of preferred websites and normalized the score to a range of 1 to 10 to match the first dataset.
7) The researchers did not mention any significant preprocessing steps that were needed to be taken on the dataset. It may be assumed that the nature of data (small numerical data representing the number of occurrence or length of html tags) did not require any manipulation. One action, directed to decrease computational complexity and provide uniformity to a spectrum of algorithms used in the paper, was to use Min-Max Scaler that preserve the original distance and relationship between the datapoints.
8) The preprocessing steps contained encoding the categorical variables with one-hot encoding and standardizing (centralizing) the numerical data the way that mean of the values is equal zero and standard deviation to one. The researchers applied filter-based feature selection methods and compared them to feature extraction. The latter which involves transforming features into linear combinations of attributes, was considered impractical due to the need to track and update features during user interactions. They employed techniques such as Correlation, Mutual Information (MI), and the mRMR algorithm for feature selection. MI, which measures mutual dependence between variables and captures both linear and nonlinear relationships, was used to rank features. Continuous variables were discretized to apply MI effectively. The mRMR algorithm was utilized to select a subset of features that maximized relevance to the class variable while minimizing redundancy among selected features. The study aimed to find the most informative features for classification, ultimately enhancing the system's performance. After first evaluation of the classification model, the authors detected an imbalance problem – the dataset contained much more sessions where user does not make purchase comparing to where the purchase was made. The oversampling approach was applied and resulted in adding more purchase-made session to the data set.

2.4. Dataset size overview

The size of dataset is an important factor in obtaining high scores in models' evaluation metrics. The data sizes and its meaning in the papers is shown in table 3.

Table 3. Data size and meaning of datasets in corresponding papers

No.	Data size and meaning
1	450 Web pages, 6 categories, 3890 images
2	Three datasets were utilized with (number of sessions: number of different pages): (13 754: 683), (38 000: 294), (59 600: 497).
3	15432 URLs**,** 2522 features for each webpage.
4	1800 unreliable news sites, 8200 reliable news sites, 7627 homepages, 2 million subpages, 174 thousand articles.
5	45373 phishing and benign instances that contain URLs and HTMLs. 321 thousand unique words from the set of HTML codes.
6	As the dataset was relatively small – 400 screenshots of web pages rated by 40000 users and 100 screenshots of pages rated by 174 users, the authors used transfer learning methods to increase the efficiency of the score predictions.
7	The dataset consisted of randomly sampled 600 websites from a public directory of websites. Out of these websites 21 independent features were extracted that stored small numerical values describing mainly html tags.
8	Over 12 thousand unique user navigation sessions that end with either purchase or leaving a website.

3. Models used for predictions

The researchers utilized the following models in their experiments:

1) Transfer learning method was used that utilized already trained Deep Convolutional Network VGG-19 model that extracted feature vectors out of images. Subsequently the vectors were used as independent variables for Logistic regression, Support Vector Machine (SVM), K-nearest neighbour (KNN) and Naïve Bayes machine learning algorithms.
2) The models used for web navigation prediction were two threshold-based All-Kth Modified Markov Models which are probabilistic models. One of them with geometric and second with branching factor thresholds. Geometric employs properties like states, outlinks, sessions, and transition. In the contrast, branching factor bases on a ratio between a number of outlinks and number of states.
3) Five well-known machine learning models are employed for website categorization in the second experiment. These models include logistic regression, k-nearest neighbours (KNN), an artificial neural network (ANN), adaptive boosting, and a CART decision tree.
4) The researchers combined custom structure-based models with NELA and RDEL content based classifying models. This way they updated the baseline models by enhancing them with structural features resulting in NELA+ and RDEL+ models.
5) The WebPhish model designed for a binary classification task, aiming to predict between two distinct classes: legitimate or phishing. Consists of two convolutional layers and two FC layers that apply ReLu activation function.
6) The authors decided to use a deep neural network with an architecture inspired by AlexNet that is known to perform well in aesthetics assessment tasks involving web and photo images. The original AlexNet architecture was designed for classification tasks, but in this study, it was adapted for regression. Specifically, the authors replaced the original output layer with a single neuron output for regression. The goal was to predict an aesthetic score for each webpage screenshot. Since the datasets used in the study were relatively small, the authors took steps to prevent overfitting – they decreased the neuron count in the fully connected layers of the network, which improved the model's ability to generalize effectively when working with a limited amount of data.
7) The researchers used five machine learning methods: Logistic Regression, KNN, SVM, Naïve Bayes and J48 Decision Trees. The models had their hyperparameters tuned with Grid Search approach. The researchers put emphasis on statistical tests and techniques of the dataset, i.e. Fleis Kappa to assess reliability of agreement between the SEO experts and correlation analysis between the independent variables. This emphasis aimed to reveal the importance of particular features in the SEO guidelines adjustment level classification.
8) The comparison of three models has been conducted that included: Multilayer Perceptron (MLP) employed with a single hidden layer with resilient backpropagation; SVM algorithm, found its place in the study, despite lacking a mechanism for continuous updates with objective to identify an optimal hyperplane for effective class separation, and the authors utilized both linear and radial basis function kernels to handle linear and nonlinear data relationships. The authors also encompassed the C4.5 algorithm for decision tree construction. This choice was made due to C4.5's ability to handle numerical attributes, address missing values, and perform tree pruning. Additionally, the study explored the Random Forest algorithm, which involves constructing an ensemble of decision trees using bagging resampling and aggregating their predictions via a voting mechanism.

4. Evaluation metrics

The choice of evaluation metric is critical as it directly affects performance assessment. It must align with specific goals, consider trade-offs, adapt to context, reflect real-world impact, address bias and imbalance, and suit the task type (regression or classification) and multiclass complexities. Some domains may require specialized metrics tailored to unique challenges. The papers authors in their models used the following metrics:

1) F1-Score, Accuracy, Precision and recall, Confusion matrix. Accuracy assesses the overall correctness by determining the proportion of accurate predictions among all the predictions made. Precision measures the accuracy of positive predictions, finding the ratio of true positives to all positive predictions. Recall assesses the model's ability to identify all actual positives, calculated as the ratio of true positives to all actual positives. F1-Score is a balance between precision and recall, providing a single score that considers both false positives and false negatives.
2) Coverage – average number of predictions possible for the given test dataset – the ratio of total possible predictions from each session to total number of session in the testset; Prediction Accuracy – a measure calculated by dividing the number of correct predictions by the total number of sessions in the test set.
3) The proposed approach's performance was evaluated using the following metrics: Classification Accuracy: These measure how accurately the classifier predicted website categories. It's computed by dividing the count of correctly classified instances by the total number of instances. True Positive Rate (Recall), True Negative Rate, Precision and F1-Score.
4) The study primarily uses the ROC AUC metric to evaluate the performance of the classifiers. This metric evaluates the classifiers' capacity to differentiate between positive and negative classes, while considering the class imbalance between mainstream and fake news. Additionally, the study conducts an error analysis based on websites attributes such as age, popularity, and ideological leaning to further assess model performance.
5) Confusion matrix was generated to describe the performance of the classification model that allowed to visually evaluate the model, as well as calculate F1-Score, Accuracy, Precision and Recall.
6) The evaluation metrics mentioned include Precision, Recall, F1-Score, and Accuracy. All the metrics were provided for each of the trained classifiers with an additional differentiation for the applied method of image feature extraction.
7) The models were evaluated on Accuracy of the predictions with used of both Holdout and 10-Folds Cross Validation methods. The researchers also evaluated the suitability of the dataset size by plotting the learning curve of each model. To prove the difference between pre-set baseline model, the McNemar and Wilcoxon tests were performed to determine the level of similarity of the predictions.
8) The researchers provided a comprehensive classification report that contained F1-Score, Accuracy, True Positive Rate, True Negative Rate determined for each class in the classification.

5. Conclusions

This paper has examined a spectrum of prediction tasks within the subject of website-related predictions. These tasks encompass website categorization, web navigation prediction, crack and malicious website detection, fake news website detection, phishing website detection, and the assessment of website aesthetics.

Researchers carried out acquisition of data from various sources, which range from manually labelled sets to publicly available datasets or combination of both. Furthermore, the datasets themselves exhibit significant diversity in terms of size and format, with some encompassing hundreds of web pages and images, while others include massive collections of URLs, web content, or session logs. In cases where dataset sizes are constrained, transfer learning and model adaptation have proven beneficial, enhancing model performance. In essence, data handling in the context of web-related predictions is not a one-size-fits-all endeavour but rather a dynamic process that necessitates adaptability to accommodate the unique characteristics and requirements of each task.

Modelling approaches also exhibit a wide range, spanning across traditional machine learning algorithms such as SVM, KNN, Naïve Bayes, and logistic regression, along with probabilistic models like Markov Models. Deep neural networks inspired by architectures like AlexNet have been employed, as well as customized structural models tailored to specific prediction tasks.

The evaluation of model performance relies on a suite of metrics, including F1-Score, accuracy, precision, recall, ROC AUC, and more. These metrics encompass a spectrum of measures, including but not limited to the F1-Score, which balances precision and recall, accuracy, which quantifies overall correctness, precision, highlighting the proportion of correctly predicted positive instances among all instances predicted as positive, recall, indicating the proportion of actual positive instances that were correctly predicted, and (ROC AUC), which evaluates the model's ability to distinguish between classes.

The dynamic nature of the web imposes adaptability in data acquisition, preprocessing, and modelling. There are various ways of achieving similar results and only continuous creation and assessment of the models can reveal the best state-of-art approach to predicting diverse, only growing web-related dependencies. It is specifically the tempo of the growth of the web along rising users' activity that drives the scientific research in the fields of website categorization, identification of malicious content and user navigation prediction.

References

[1] Bozarth L., Budak C.: Lay it Out: Detecting Fake News Publishers through Website Structure Data, 2019 [http://doi.org/10.2139/ssrn.3419781].
[2] Cohen D. et al.: Website categorization via design attribute learning. Computers & Security 107, 2021, 102312 [http://doi.org/10.1016/j.cose.2021.102312].
[3] Delitzas A., Chatzidimitriou K. C., Symeonidis A. L.: Calista: A deep learning-based system for understanding and evaluating website aesthetics. International Journal of Human-Computer Studies 175, 2023, 103019.
[4] How many websites are there in the world? – A Daily Calculator [https://www.statsfind.com/how-many-websites-are-there-in-the-world-a-daily-calculator/] (available: 13.02.2024).
[5] Jindal H., Sardana N.: Web navigation prediction based on dynamic threshold heuristics. Journal of King Saud University–Computer and Information Sciences 34(6), 2022, Part A, 2820–2830 [http://doi.org/10.1016/j.jksuci.2020.03.004].
[6] Korkmaz M. et al.: Deep neural network based phishing classification on a high-risk URL dataset. International Conference on Soft Computing and Pattern Recognition. Springer International Publishing, Cham, 2020.
[7] Matošević G., Dobša J., Mladenić D.: Using Machine Learning for Web Page Classification in Search Engine Optimization. Future Internet. 13(9), 2021.
[8] Nandanwar A., Choudhary J.: Web Page Categorization based on Images as Multimedia Visual Feature using Deep Convolution Neural Network, 2020, 619–625.
[9] Opara C., Chen Y., Wei B.: Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics. Expert Systems with Applications 236, 2024, 21183 [http://doi.org/10.1016/j.eswa.2023.121183].
[10] Sakar C. O. et al.: Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput & Applic 31, 2019, 6893–6908.
[11] Shaffi S. S., Muthulakshmi I.: Search Engine Optimization by using Machine Learning for Web Page Classification. International Conference on Augmented Intelligence and Sustainable Systems – ICAISS, 2022, 342–349.

M.Sc. Eng. Patryk Mauer
e-mail: patryk.mauer@student.po.edu.pl

In 2022 he started his doctoral studies at the Opole University of Technology in the field of technical informatics and telecommunications.
Research interests: business process automation, web scraping, artificial intelligence.

https://orcid.org/0000-0003-4173-0424