An Empirical Study to Classify Website Using Thresholds from Data Characteristics

Ruchika Malhotra and Anjali Sharma

Abstract The advent of web had resulted in a plethora of information and data. However, its volume heterogeneity and unstructured organization makes information retrieval difficult. To the existing practice where website categorization is largely based on style rather than text, addition of an extra dimension in form of genre is expected to significantly improve the search outcome. Keeping this in view, we attempt to build a novel classification model to categorize websites into genres using thresholds of the web metrics. Statistical measures of central tendency are assumed to render a value that distinguish websites from a sample space containing News, Travel and Tourism, Entertainment and Social media. Through the statistical analysis of the data we find that the data distribution of all metrics which constitute the website properties are highly skewed. Hence, conventional analysis based on normal distribution statistics fails to apply. Adopting to a systematic empirical approach, we find that the classification performance measure identified through the Area Under the Curve is maximized around a threshold value which is twice the value of the "median-absolute-deviation" of the web metrics.

Keywords Web genre · HTML metrics · Threshold · Median-Absolute-Deviation · Naive Bayes

1 Introduction

The semi-structured, dynamic and heterogeneous nature of websites make information classification increasingly challenging [1, 2]. As a result, even the most versatile search engines provide lesser accurate results to very specific information sought on the web. In fact, the current search engines return the ranked list of documents

R. Malhotra (✉) · A. Sharma
Delhi Technological University, Bawana Road, New Delhi 110042, India
e-mail: ruchikamalhotra2004@yahoo.com

A. Sharma
CSIR-NPL, Dr K S Krishnan Marg, New Delhi 110012, India

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_39

depending upon textual similarity, together with an independent measure of each web page's importance [3]. In such cases, the search outcome is a myriad of web pages, which then further requires more user defined search filters, thereby making the search task more difficult. For more complex cases, it may even require that the user would need to visit each web page and apply a manual search for the desired information.

However, dealing with genres pose two important aspects, (i) What are most important and relevant genres to be considered? and (ii) and what factors explicitly help distinguish the genres. In general, the complexity of a web page is heterogeneous in nature, i.e., despite prescription of what is genre, it may be found that genres tend to overlap and mix [4, 5]. For instance, the Google search engine has tabs for 'Maps', 'News', 'Images', 'Videos' etc., to segregate content according to presentation style, rather than topic. However, the web page information are not classified according to genres, and as a result the user is unable to specify the category of search content style like 'Report' or 'Wikipedia' or 'e-Commerce', which may be more specific in accordance to the search interest itself. Thus, classification of web pages into genre is a challenging task as the information sought is specific to certain appropriate features that describe the web page in context of a genre.

However, beset with difficulties the need has been realized to have an extra classification scheme, particularly in terms of genre. As of date, the web genre classification distinguishes between pages by means of their style, presentation layout, form and meta-content features rather than topic. Genre adds an extra dimension to the classification of web pages, with improved search results. In order to classify the web pages we require specific features through which the genres can be classified. This forms the objective of the current work.

Assuming that a set of independent HTML metrics can be identified and that a combination of these metrics can aptly describe the genre of a given website, we propose that there exists a threshold value associated with each metric, on the basis of which one can classify the web pages. Provided that there exists such a scheme, and that the underlying methodology is simple, it may be then guaranteed that information retrieval from web would be fast and accurate and be quite simpler to handle large databases. As a case study, we try to identify websites related to "Travel and Tourism" and "Social Media" from a sample space containing genres such as "E-Commerce", "Social Media", "Entertainment", and "News". Using statistical analysis of the data distribution, a systematic methodology is being developed for identifying the thresholds of the web metrics by exploiting inherent internal characteristics of the HTML metrics. The thresholds of selected web metrics are then utilized for building models to predict genre of website using machine learning techniques.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the related works in the estimation of thresholds in software engineering. Section 3 explains our threshold-based feature selection methodology. In Sect. 4, we show our results which include the statistical description of the metric data distribution, the classifier performance values against a range of values. From the trend that follows, we find that a reasonable threshold associated with each web category is the value corresponding to twice the median absolute deviation. In Sect. 5 we present

the limitations of the work. Finally, we conclude the work in Sect. 6 and suggest improvements for future studies.

2 Literature Survey

The web genre classification depends upon the input metrics, ma- chine learning technique and the genre class. In earlier research work, web pages were considered to represent a single web genre [6, 7] for web genre classification. In contrast with this assumption, many scholars and researchers argued that a single genre classification scheme is inappropriate for web pages [8–11].

In one of the original attempts, Crowston and Williams [12] from a set of randomly selected web pages, with certain set of objectives identified, proposed four types of genres - Reproduced genres, Adapted genres, Novel genres and Unclassified web pages. The study revealed that genres cannot be simply cached and stored in a repository, but evolves. Similar credence was supported by Shepherd and Watters [6, 7]. The latter authors introduced a new terminology "Cybergenre", which is currently popular as "web genre". Accordingly, the genre is characterized by a base triplet namely, {<content>, <form>, <functionality>}. While both <content> and <form> represented the traditional genres, <functionality> defined the capabilities offered by the web. Significantly, along with the functionality genre, an important attribute was soon realized, i.e., based on the use of hypertext and/or HTML. Each hypertext corresponded to a genre. It may be noted that although a website may be a collection of web pages, the genre analysis is basically done for the entire website [13, 14]. A super-genre classification of websites [15] was done by using structure, content and their combination to improve the classification accuracy.

Overall, the research work discussed above is in agreement that structure and functionality attributes of a web page represent useful information which can be used to identify the genre of a website. Therefore, we have focused on the quantitative web metric set of <Structural> and <Functionality> attributes represented by text formatting, navigation and external object HTML tags.

In general, the threshold in software systems can be estimated from statistical deductions and mathematical models. The statistical methodology provides qualitative thresholds, however, to improve the validity of the results it is important to study the relation between the data characteristics, underlying assumptions and nature of the problem. Here we briefly review the threshold estimation studies based on statistical inference for software.

The study conducted by Erni and Lewerentz [16] estimated the threshold to be in the range of statistical mean (μ) and standard deviation (σ), represented as $T \pm = \mu \pm \sigma$, assuming data to be distributed normally. However, the technique assumed the input metric data to be normally distributed. Usually such distributions are seldom common in software projects and hence the applicability of the technique is limited. The work by French [17] included Chebyshev's inequality theorem along with μ and σ for threshold calculation but distribution nature of data was again not

considered and the methodology suffered to the data outliers. The recent work of de Siqueria et al. [18], have suggested three similarity thresholds, using arithmetic or the weighted mean, *k*-means clustering and silhouette coefficient maximization, for the genre aware focused crawling.

Shatnawi [19] used ROC characteristics to identify threshold values and analyzed its association with different error severity levels. The relevant threshold values were found for high and medium risk categories of ordinal classification but could not find practical threshold values for binary classification. In a following study [20], the author calculated the thresholds corresponding to the C&K metrics using Bender's approach [21] based on logistic regression and it was found that risk levels can be used to identify metric thresholds. Similarly, Malhotra et al. [22] also used Bender's approach to calculate the metrics threshold and determined the effects of threshold on change prediction with inter-project studies. Their results showed that the transferability of the threshold is limited rather to a narrow confidence interval in inter-project comparisons. In a more recent work, Shatnawi [23] proposed data transformation method to reduce skewness in the data and the threshold values were estimated using the statistical parameters such as μ and σ, similar to the works of Erni et al. [16]. However, what limits the underlying methodology is the shift of values by a constant value prior to the data transformation. Alves et al. [24] investigated data distribution properties of object oriented metrics to derive threshold values and the estimated metrics threshold values were insensitive towards data outliers. Similarly, Ferreira et al. [25] statistically analyzed the data to calculate the threshold range of certain metrics. The authors found that most of the metrics followed a heavy-tailed distribution and argued that a general threshold could not be applied to the object oriented software projects. On the other hand, Hussain et al. [26], compared the effect of thresholds derived using Bender's approach and those mentioned by Alves et al. [24] and concluded that thresholds cannot be generalized for all the systems due to variation in data characteristics.

The studies discussed above have emphasized the importance of the data characteristics and statistics to be considered before estimating the thresholds. Hence, in this work we estimate the threshold of web metrics using the statistical measures of central tendency after analyzing the data distribution. The threshold estimates are used for categorizing the websites according to their genre.

3 Methodology

The methodology we follow in this work is schematically shown in Fig. 1. The Web Metric Collection and Reporting System (MCRS) [27], crawls URL to collect HTML, NLP and text metrics for web genre classification. The HTML metric collector extracts all the links in the web page and collects various HTML web metrics namely, Text Formatting tags, document structure tags, external object tags, instruction and navigation tags. As highlighted in [28], the combination of lexical, functional and structural attributes shall be used for genre classification. Therefore,

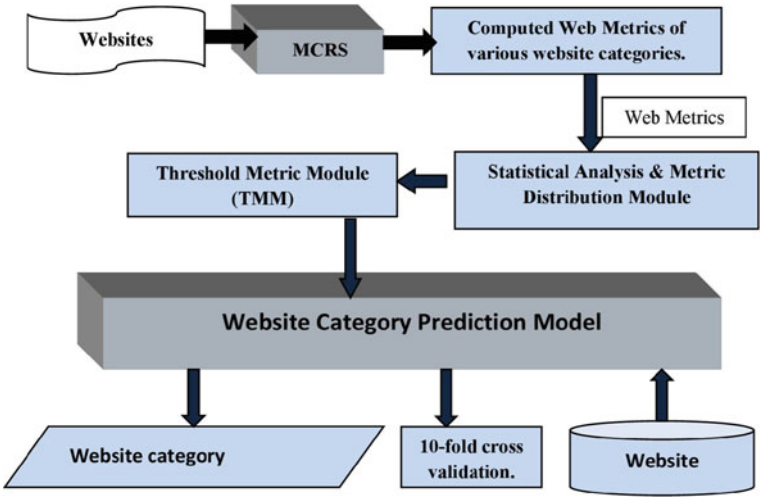

Fig. 1 Schematic representation of the methodology adopted in the study

we have used the "Structural" features of web page represented by text formatting
, , <li>, <p>, , <ul> and navigation <a> tags, while the external object tags <img>, <script> are used to define "Functionality". Therefore, these nine web metrics, listed in Table 1, constitute the independent variables in the study, which are used to categorize the website as "Travel and Tourism", "E-Commerce", "Social Media", "Entertainment" or "News".

The set of nine metrics and website category serve as input to the statistical analysis and metric distribution module. The statistical parameters of central tendency for sample space including range ($R_{\max}$, $R_{\min}$), mean (μ) and median (x_{m}) are calculated in this module. Also the histogram plots are investigated to identify the distribution characteristics of the input metric space. These statistical parameters along with the sample space serve as input to the Threshold Metric Module (TMM), which estimates the threshold values for the "Travel and Tourism" and "Social Media" website category.

The website category prediction model is built using Naive Bayes classifier, with input from the TMM and renders the classification performance measure of the web category in terms of the AUC values. The selection of the Naive Bayes algorithm is not only because of the common use in data mining applications, but also due to its reliable performance for small dataset [29]. The default parameter settings were used for the learners as specified in Weka. A priori, AUC is chosen as the performance measure, due to the inherent class imbalance observed in the dataset. By definition, AUC is the probability a classifier ranks a randomly chosen positive instance higher than its negative instance counterpart. In the Receiver Operating Characteristics (ROC), the magnitude of AUC varies in between 0 and 1. Note that ROC analysis helps in decision making, by relating the performance and non-performance of a classification model.

Table 1 The HTML metrics used in the study

Table	HTML tag	Description
Break	 	It breaks the line and bring the statement in next line. It is an empty tag which means it has no end tag
Division		It defines a section in html tag and act as a container for other elements to style them using CSS and perform various tasks
List		It is used in ordered list, menu or unordered list to list various items
Paragraph	<p>	To write any paragraph we use <p> tag. Spacing is provided automatically by browser before and after paragraph
Span		It provides no change by itself and is used to group elements in a document
Unordered list		tag is used with tag to create unordered list. It defines the list with bullets
Image	<image>	It defines image in html. It adds image to the html page by using two attributes 'src' (source) and 'alt' (alter). Src provides the path of the image. Alt alters the size of the image
Script	<script>	It defines client side scripting. It contains scripting statements and also it points external file through "src" (source) attribute
Anchor	<a>	This tag is used to include hyperlinks to the html page. It links one page to another. The main attribute of <a> tag is <href> which include link destination

4 Results and Discussion

4.1 Data Characteristics

In Table 2 we describe the statistics of the metric data, the latter which includes all five web categories. The range of the metrics with its upper limit, designated as $R_{\max}$ are shown along with the measures of central tendency, i.e., the mean (μ) and median (x_{m}).

It is evident from Table 2 that the range of the metrics are quite different. We also find significant difference in the mean and median values thereby inferring a non-normal distribution of the metrics. All metrics distribution are found positively skewed, since $\mu > x_{\mathrm{m}}$. Empirically, by considering the difference ($\mu - x_{\mathrm{m}}$), as a measure of skewness, the data reveal that the distribution associated with the and <a> metrics are relatively more skewed, while <script> is least skewed.

We first attempt to construct the threshold parameters for the "Travel and Tourism" web category. For the same, we first analyze the statistical parameters associated with the web category with respect to the sample space. In Table 2, we show the statistical description of the metrics associated with the "Travel and Tourism" web category. The overall characteristics of the sub-space remains similar to that of the sample space, i.e., the metrics distribution are skewed with mean being greater than the

Table 2 The statistical description of the selected HTML metrics for the sample space, travel and tourism and social media categories

Metrics	Sample space			Travel and tourism				Social media			
	R_{max}	x_m	μ	R_{max}	x_m	μ	MAD	R_{max}	x_m	μ	MAD
 	113	4	12.24	94	5	13.23	19.5	113	2	10.2	2
	1229	90	155.03	472	56	71.97	65	1229	65	147.8	43
	907	48	97.97	315	36	51.34	72	652	11	60.3	11
<p>	203	6	20.65	153	3.5	11.71	9	203	3	20.4	3
	506	29	68.72	403	11.5	37.47	39	376	10	43.8	9
	190	8	17.45	56	6.5	9.36	14.5	122	3	11.2	3
<image>	179	22	34.01	90	15	20.34	24.5	179	8	27.9	8
<script>	108	16	19.46	42	10	13.81	14.75	66	12	17.7	7
<a>	958	101	175.55	353	76	88.81	69	760	44	111.3	35

The R_{max}, x_m, μ, MAD values represent the Upper, Median, Mean and Median-Absolute-Deviation values, respectively. The Lower value (R_{min}) for all the metrics is 0

median. Besides, we also note that the category resides well inside the sample space with no range maximum of any of its nine metrics with that of the sample itself.

The basic statistical description of the data pertained to the "Social Media" web category is shown in Table 2. It is found that the distribution of data are very different from <script> metrics and spans the entire range in the "Social Media" category, which is not the case for "Travel and Tourism". The metrics range from [0, 472] in "Travel and Tourism", but shows a wider range of [0, 1229] in the "Social Media" category. We also note that four out of nine metrics, namely
, , <p> and <img> representing the "Social Media" category are spread all across the entire range, which is in contrast to the data distribution associated with "Travel and Tourism" category.

For a better understanding of the category wise metric distribution with respect to the sample, and also among the five categories, we analyzed the data in terms of frequency plots as shown in Figs. 2 and 3. The wide difference among the web categories in the metric space is very evident. Not only that we find the frequencies associated with the metric values to be very different, but also that certain metrics distribution were found to be continuous for some categories, while for others it looked non-uniform and discontinuous. For instance, for in "Travel and Tourism" the frequency of data in the range [0, 50] was found to be 17, while in "Social Media" it was determined to be 30. On the other hand, both metrics <li> and <ul> shows a continuous and decreasing trend with increasing range in the "Social Media" category, while in "Travel and Tourism", the distribution is discontinuous. In fact, based on our inter-quartile analysis, we find that the data in the "Travel and Tourism" category for metric <li> in the range [300, 350] and that for <ul> in [50, 60] are representation of being outliers. Thus, statistical analyses show that the web categories in the sample space are widely different in terms of the metrics that define each category.

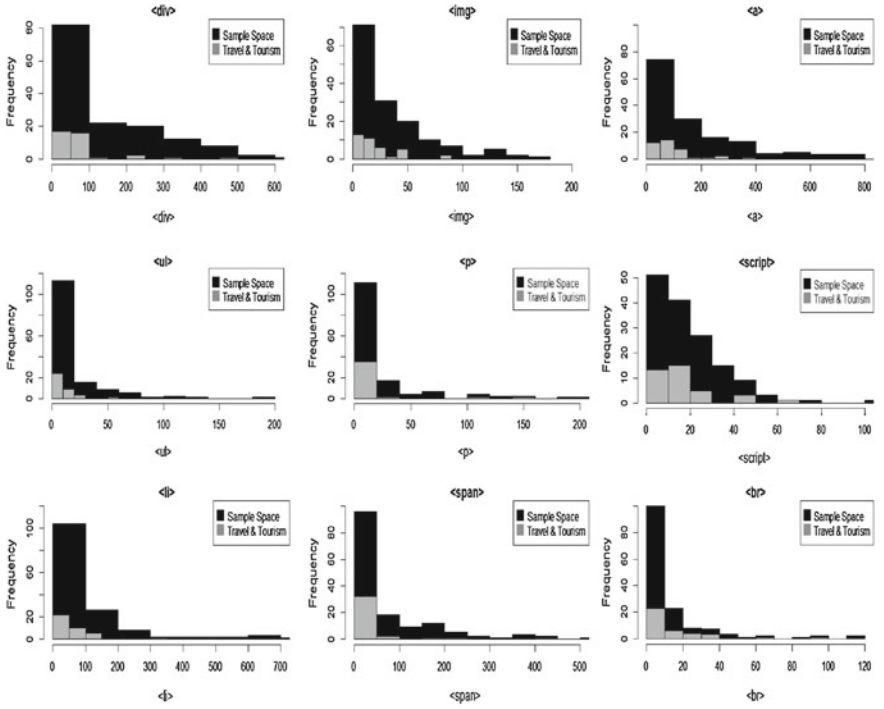

Fig. 2 Histogram representation of the sample space (shaded black), with data on "Travel and Tourism" (shaded grey) projected, of the nine metrics used in the study

These statistical observations of the metric data incite to look for a threshold value, or a set of values, which differentiate each web category in a given sample space.

4.2 Threshold Calculation

The problem at hand, therefore is to determine the threshold upon which one can classify the web categories from a given sample space. For this, one need to have an initial guess to the threshold value, upon which the performance measure of certain chosen category can be calculated. Further assuming that there exists a unique set of threshold parameters to determine the threshold, we vary the guess parameter in increments so as to obtain an optimal value of performance. As obvious, the range is quite different for all metrics within a given web category, and also among various categories. Thus, the minimum and maximum values of a metric distribution are not very good statistical parameter to use, since they can fluctuate greatly from sample to sample. Besides, the distribution as mentioned above have significant deviations from that of a normal distribution.

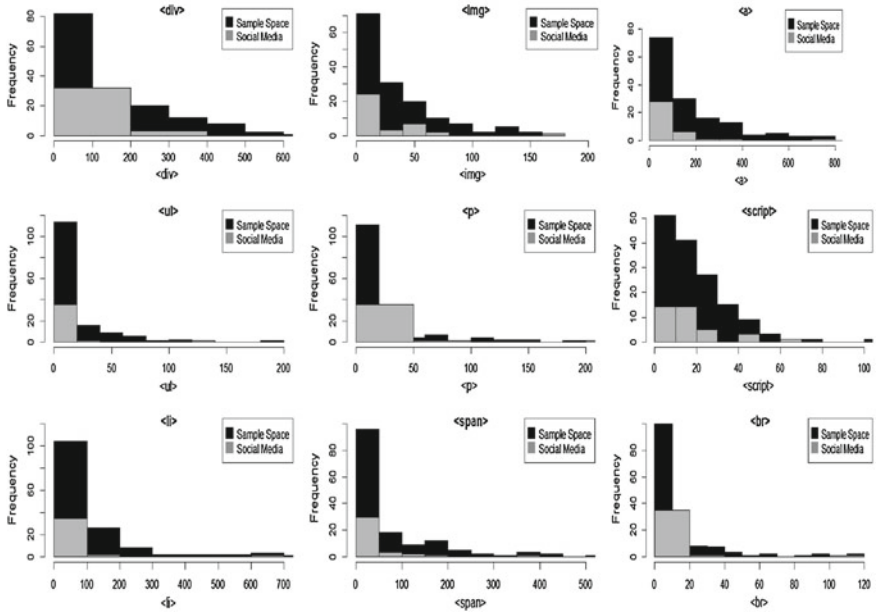

Fig. 3 Histogram representation of the sample space (shaded black), with data on "Social Media" (shaded grey) projected, of the nine metrics used in the study

As a result, we argue that neither the limiting range parameters nor the mean value of the distribution is a good choice to be considered as an initial guess for the threshold. For a simple, nonparametric statistic to represent variability of a skewed data, we therefore consider median (x_{m}) as our reference measure of central tendency, which also forms as our initial guess to the threshold value. In analogy to the role of standard deviation (σ) in normal distribution. A widely used parameter for variance in skewed dataset is the statistical quality referred as "Median Absolute Deviation" (MAD). Much similar to the relevance of $\mu \pm 2\sigma$ in normal statistic dataset, here we use $x_{\mathrm{m}} \pm (2 \times \mathrm{MAD})$ as a range for the calculation of the threshold value. Mathematically, MAD is defined in Eq. (1) as,

$$\mathrm{MAD} = \mathrm{median} \times (|X_i - \mathrm{median}_j(X_j)|) \tag{1}$$

By definition, MAD represents a measure of statistical dispersion. For non-normal dataset, MAD is a robust estimator of scale than the conventional variance or standard deviation. MAD also is a much better statistical quantity for distributions that have neither mean nor variance, such as that for Cauchy distribution, and thus includes as a universal statistical quantity for any metric space irrespective of its nature. Furthermore, an advantage of using MAD as a statistical estimator is due to its insensitiveness towards outliers. We define threshold as a boundary which differentiates radically different regions. In this context, we anticipate a change in the variation of AUC as a

Table 3 The performance of the web category prediction model with AUC measure, with and without threshold

	Travel and tourism	Social media
Without threshold	0.71	0.74
x_m	0.66	0.94
$x_m + \delta$	0.55	0.94
$x_m + 2\delta$	0.62	0.89
$x_m + 3\delta$	0.74	0.89
$x_m + 4\delta$	0.97	0.87
$x_m + 5\delta$	0.92	0.87
$x_m + 6\delta$	0.92	0.87

Here x_m represents the median value, and $5\delta = (2 \times \text{MAD})$

function of $x_m + n\delta$, where δ is an increment and "n" a positive integer. To determine the maximum range up to which $n\delta$ values will be varied, $2 \times$ MAD is considered.

4.3 Website Category Prediction Model Using Threshold

The median and $2 \times$ MAD values of all metrics are first calculated (Refer Table 2). To calculate the performance of web category prediction model using thresholds, we transform each metric into a binary form. The metric values below threshold are transformed to "zero" and those above as "one". Thereafter, the binary transformed dataset is fed as input to the Naive Bayes algorithm with stratified remove folds as class balance technique. The corresponding AUC value is computed which are shown in Table 3.

The AUC with median as threshold for the transformed data is determined to be 0.66, which is relatively lower to the AUC computed for the original (untransformed) dataset, i.e., 0.71. Note that for the value of $x_m + n\delta$, which shows a characteristic change in the AUC value would be referred as the threshold value associated with the web category.

In Fig. 4 we show the stacked line graph of the performance measure variation with respect to $x_m + n\delta$, obtained for both "Travel and Tourism" and "Social Media". The results are shown in this form mainly because stacked line representation enable us to capture the trend in the variation with variable threshold range assumed in the calculation. Besides, since such a graph is cumulative at each point the data does not overlap. It is very evident that the performance measures of both web categories are very similar. With increase in the δ, the graph initially decreases by 16 and 7% for "Travel and Tourism" and "Social Media", respectively. Thereafter, we find a gradual increase in the performance measure up to $x_m + 4\delta$, beyond which the values saturate. Behold, the definition of threshold as the boundary which separate two region in the variation of performance, we find that the boundary in this study points to $x_m + 5\delta$. Interestingly, this also correspond to the $(2 \times \text{MAD})$ value. Thus, with the

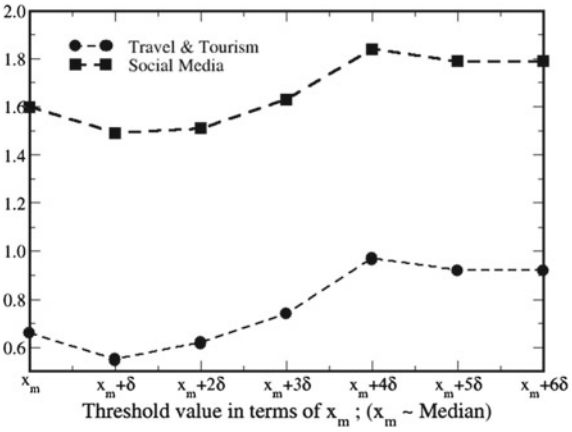

Fig. 4 The stack plot showing the variation of relative performance measure of "Travel and Tourism" and "Social Media". Note that for the exact AUC measure corresponding to "Social Media", the value corresponding to a given $x_m + n\delta$, has to be subtracted from the "Travel and Tourism" value

observation that two very different skewed metric data distribution associated with "Travel and Tourism" and "Social Media", not only exhibiting similar trend but also inferring (2 × MAD) as threshold help us formulate the following hypothesis; i.e., there exists a close relationship between the threshold value and statistics in the web category distinction, with (2 × MAD) value corresponding to the threshold itself.

5 Threats to Validity

One of the basic question is how well the present experiment has been done. In this perspective, one of the confounding issue is in the selection of the input metrics. Whether or not the chosen metrics forms a complete metric space, and/or whether there exists a linear dependence between the metrics is an internal threat to the validity of the results. As a check, it would suffice to use feature selection techniques and calculate the optimal threshold value.

In this study, we have used the Naive Bayes algorithm. For a wider understanding, the use of a single machine learning algorithm could be a possible threat to the conclusion validity of this study. However, As mentioned above Naive Bayes has been found to yield reliable results for smaller dataset and also yet being a simple model the algorithm has found numerous applications providing high performance for a large variety of datasets. However, as a future work we will be evaluating the performance with several other machine learners such as Bagging and Boosting algorithms.

The observations following the study is limited in generalization as to similar studies which may span other networking sites. This pose a possible external validity threat. For instance, the design of websites can significantly depend on the culture and tradition of various communities and public across the globe. To minimize these

local effects, it is important to collect data from various other networking sites across the globe and investigate to what extent the results can be generalized.

6 Conclusion

Identification of proper web genres are expected to ease classification at both organizational and at user level. Given that the evaluation of web sites would thereby become plausible at lower cost, development of web genres are becoming increasingly important for the developers to adopt measures so as to ease search queries. In this regard, we propose a model based on threshold to distinguish various web categories based on the statistical measures of central tendency. Since the metrics that define the metric space are found skewed, we guess that the threshold would be more related to the median value than the more widely used mean. Setting the definition of threshold as the boundary that differentiates the classification performance rendered by a machine learning algorithm, we vary the threshold value in increments, from median towards the skewed part of the spectra. The trend as captured by the AUC values clearly shows that beyond certain optimum value, the magnitude of the performance measure saturates. We argue that the set of metric values that put the magnitude of the performance in saturation can be termed as the threshold. In statistical realms, our study shows that the threshold is ($x_{\mathrm{m}} + 2 \times \mathrm{MAD}$), where x_{m} and MAD represents the median and "median absolute deviation", respectively. In analogy with the standard deviation which is commonly used for dataset with normal distribution, we conclude that the proposed threshold estimate evaluated lies within 95% confidence interval. The use of Median-Absolute-Deviation has never been proposed in any earlier works related to threshold determination in website categorization, and hence require more experiments over a wider range of website classifications.

References

1. Chetry, R.: Web genre classification using feature selection and semi-supervised learning (2011)
2. Gatto, M.: Web as Corpus: Theory and Practice. Bloomsbury Academic, London (2014)
3. Stein, B., Zu Eissen, S.M., Lipka, N.: Web genre analysis: use cases, retrieval models, and implementation issues. In: Genres on the Web, pp. 167–189. Springer, Dordrecht (2010)
4. Ponzanelli, L., Mocci, A., Lanza, M.: Summarizing complex development artifacts by mining heterogeneous data. In: Proceedings of the 12th Working Conference on Mining Software Repositories, pp. 401–405. IEEE Press, New York (2015)
5. Wu, L., Du, L., Liu, B., Xu, G., Ge, Y., Fu, Y., Li, J., Zhou, Y., Xiong, H.: Heterogeneous metric learning with content-based regularization for software artifact retrieval. In: IEEE International Conference on Data Mining (ICDM), pp. 610–619. IEEE, New York (2014)
6. Shepherd, M., Watters, C.: The functionality attribute of cybergenres. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, HICSS-32, p. 9. IEEE, New York (1999)

7. Shepherd, M., Watters, C.: Identifying web genre: hitting a moving target. In: Proceedings of the WWW Conference. Workshop on Measuring Web Search Effectiveness: The User Perspective, vol. 18, New York (2004)
8. Rosso, M.A.: Using genre to improve web search. Doctoral dissertation, University of North Carolina, Chapel Hill (2005)
9. Williams, K.C.M.: Reproduced and emergent genres of communication on the World Wide Web. Inf. Soc. **16**, 201–215 (2000)
10. Santini, M.: Characterizing genres of web pages: genre hybridism and individualization. In: HICSS 40th Annual Hawaii International Conference on System Sciences, pp. 71–71. IEEE, New York (2007)
11. Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X.: Genre based navigation on the web. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences, p. 10. IEEE, New York (2001)
12. Crowston, K., Kwasnik, B.H.: A framework for creating a facetted classification for genres: addressing issues of multidimensionality. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, p. 9. IEEE, New York (2004)
13. Copestake, A.: Errors in wikis. In: Proceedings of the Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources (2006)
14. Mehler, A.: Text linkage in the wiki medium: a comparative study. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 3–7, 2006 (EACL 2006): Workshop on New Text—Wikis and blogs and other dynamic text sources, pp. 1–8 (2006)
15. Lindemann, C., Littig, L.: Classification of web sites at super-genre level. In: Genres on the Web, pp. 211–235. Springer Netherlands (2010)
16. Erni, K., Lewerentz, C.: Applying design-metrics to object-oriented frameworks. In: Proceedings of the 3rd International on Software Metrics Symposium, pp. 64–74. IEEE, New York (1996)
17. French, V.: Establishing software metric thresholds. In: Proceedings of the 9th International Workshop on Software Measurement (1999)
18. de Siqueira, G.O., de Assis, G.T., Almeida Ferreira, A., Mangaravite, V., Cardeal P´adua, F.L.: Strategies for automatic determination of similarity threshold for genre-aware focused crawling processes. In: IADIS International Journal on WWW/Internet, vol. 15 (2017)
19. Shatnawi, R., Li, W., Swain, J., Newman, T.: Finding software metrics threshold values using ROC curves. J. Softw. Maint. Evol.: Res. Pract. **22**, 1–16 (2010)
20. Shatnawi, R.: A quantitative investigation of the acceptable risk levels of OO metrics in open-source systems. IEEE Trans. Softw. Eng. **36**, 216–225 (2010)
21. Bender, R.: Quantitative risk assessment in epidemiological studies investigating threshold effects. Biom. J.: J. Math. Methods Biosci. **41**, 305–319 (1999)
22. Malhotra, R., Bansal, A.J.: Fault prediction considering threshold effects of object-oriented metrics. Expert. Syst. **32**, 203–219 (2015)
23. Shatnawi, R.: Deriving metrics thresholds using log transformation. J. Softw.: Evol. Process. **27**, 95–113 (2015)
24. Alves, T.L., Ypma, C., Visser, J.: Deriving metric thresholds from benchmark data. In: IEEE International Conference on Software Maintenance (ICSM), pp. 1–10. (2010)
25. Ferreira, K.A., Bigonha, M.A., Bigonha, R.S., Mendes, L.F., Almeida, H.C.: Identifying thresholds for object-oriented software metrics. J. Syst. Softw. **85**, 244–257 (2012)
26. Hussain, S., Keung, J., Khan, A.A., Bennin, K.E.: Detection of fault-prone classes using logistic regression based object-oriented metrics thresholds. In: IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 93–100 (2016)
27. Malhotra, R., Sharma, A.: A web metric collection and reporting system. In: Proceedings of the Third International Symposium on Women in Computing and Informatics, pp. 661–667. ACM, New York (2015)
28. Malhotra, R., Sharma, A.: Quantitative evaluation of web metrics for automatic genre classification of web pages. Int. J. Syst. Assur. Eng. Manag. **8**, 1567–1579 (2017)

29. Frman, G., Cohen, I.: Learning from little: comparison of classifiers given little training. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 161–172. Springer, Berlin (2004)