An \textit{n}-gram Based Approach to the Classification of Web Pages by Genre

Jane E. Mason
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
jmason@cs.dal.ca

Abstract
The extraordinary growth in both the size and popularity of the World Wide Web has created a growing interest not only in identifying Web page genres, but also in using these genres to classify Web pages. The hypothesis of this research is that an \textit{n}-gram representation of a Web page can be used effectively to automatically classify that Web page by genre. This research involves the development and testing of a new model for the automatic identification of Web page genre; classification results using this model compare very favorably with those of other researchers.

1. Introduction
In the book \textit{Genre Analysis} [23], Swales observes that it is a component of human behavior to organize and categorize communicative events, at least in part, through the use of groups of genres. As Shepherd and Watters [20] note, the use of genre for classification allows the recognition of items that are similar, even in the midst of great diversity. It is no surprise then, that the unprecedented expansion of the World Wide Web has triggered an increasing awareness of Web page genres, and a corresponding interest in using these genres to classify Web pages.

As early as 1997, Crowston and Williams [6] conducted a survey of 100 Web pages, looking for both reproduced and emergent genres. They found that analyzing Web usage through genres was very effective, and encouraged Web site designers to take users’ expectations about familiar genres into consideration. That same year, Chekuri et al. [5] automatically classified Web pages into pre-specified categories, with the goal of increasing the precision of Web searches. The following year, Roberts [15] demonstrated that personal homepages represent a distinct genre, and Shepherd and Watters [20] examined the emergence of what they termed cybergenres. They noted that some genres seem to emerge spontaneously, such as hotlists and homepages, while others evolve from existing genres, such as online dictionaries and newspapers.

Web page genre classification is a potentially powerful tool in filtering the results of online searches. Categorizing Web pages by genre can allow a user to specify the genre that is of interest with regard to a particular query. For example, depending on the query, the user may be interested in retrieving only FAQ pages or only personal home pages. A user could also choose to filter online searches by excluding a particular genre from retrieved pages, such as retrieving all genres except shopping pages.

The hypothesis of the research discussed in this paper is that an \textit{n}-gram representation of a Web page can be used effectively to automatically identify the genre of that Web page. A new model for the automatic identification of Web page genre is developed, using \textit{n}-gram representations of Web pages and Web page genres. The proposed research includes an exhaustive examination of the problems associated with developing the model, including the size and number of the \textit{n}-grams with which each Web page and Web page genre is represented, the method of computing the distance (dissimilarity) between two \textit{n}-gram representations, and the feature selection method with which to choose these \textit{n}-grams. The effect of preprocessing the data, such as removing JavaScript code and HTML tags from each Web page, is also studied. The new model is evaluated on several data sets, including data sets that are multi-labeled and/or highly unbalanced. The results are compared with those of other researchers for the same data sets.

The remainder of the paper proceeds as follows. Section 2 describes the initial phase of research in which models for representing the Web pages and Web page genres are developed and tested. Section 3 discusses the selection of a distance function for determining the similarity between two \textit{n}-gram profiles, and Section 4 relates the results of the investigation of feature selection measures for choosing the \textit{n}-grams with which to represent each Web page. Section 5 discusses the effects of preprocessing the data set, and Section 6 describes the development of thresholds to allow Web pages to be labeled as belonging to more than one genre. Section 7 summarizes the results to date and discusses the direction of future work.

2. Representing Web Pages for Classification
In order to classify Web pages by genre, it is necessary to identify features that effectively characterize each Web
page and genre. Web pages can be represented much like documents that are used in text classification, however the representations may also include information such as URL information or HTML tags. For example, Rehm [14] uses linguistic features combined with HTML meta data and presentation related tags, while Meyer zu Eissen and Stein [13] combine genre-specific vocabulary and closed-class word sets with text statistics, part-of-speech information, and HTML tags. Boese and Howe [2] use a bag of words representation augmented with other information that includes text statistics, HTML tags, and URL information. Jebari [8] combines two centroid-based classifiers, one of which uses structural information from the document, while the other uses URL information. Stein and Meyer zu Eissen [22] give a detailed chronological overview of the document representations used for genre classification on Web-based corpora.

The ongoing research discussed in this paper represents Web pages using n-grams. An n-gram can be thought of as the contents of a fixed-size sliding window moved through the text. Byte n-grams are raw character n-grams in which no bytes are ignored, including the whitespace characters, therefore byte n-grams capture some of the structure of a document. Character n-grams, on the other hand, use letters only and typically ignore digits, punctuation, and whitespace. The use of n-grams has been common in language modeling since at least 1948 when Claude Shannon, considered the father of information theory, investigated the question of determining the likelihood of the next letter in a given sequence of characters [19]. Since that time, n-grams have been widely used in natural language processing and statistical analysis. For example, Cavnar and Trenkle [3] use n-gram representations of documents for text classification, as do Kešelj et al. [10] in their work on authorship attribution.

The research for this thesis began with an evaluation of the potential of using n-grams to represent Web pages during the genre classification process. As a preprocessing step, HTML tags and JavaScript code were removed from the Web pages. In this initial study, the textual content of Web pages was used to create Web page profiles, each consisting of the most frequent byte n-grams in the Web page, and their associated frequencies. These n-gram profiles were produced using the Perl package Text::Ngrams\(^1\). Three classification models were tested using these n-gram profiles, each of which used a distance function measure to determine the similarity between two profiles. The models differed in their representations of the Web page genres. For each model, 60 trials were performed, each with a different combination of n-gram length and Web page profile size. The n-gram length ranged from 2 to 7 in increments of 1, and the Web page profile size ranged from 500 to 5000 in increments of 500.

All of the experiments were run using 10-fold cross validation to increase robustness against overfitting.

These experiments were run using the popular 7-Genre data set\(^2\), which contains 1400 Web pages and is evenly balanced, with 200 Web pages in each genre. Based on these experiments, we were able to conclude that the n-gram representation of Web pages and Web page genres does indeed have good potential for use in genre classification. We also concluded that increasing the number of most frequent n-grams used from each Web page beyond a base number of 500 does not appreciably affect the precision, recall, and classification accuracy. The most robust of the classification models was selected for further investigation. This model represents each Web page genre by combining the Web page profiles from each Web page of that particular genre (in the training set) to form a centroid profile to represent that genre. Each Web page profile from the test set is compared with each centroid genre profile from the training set. The Web page is assigned the label of the Web page genre profile to which its profile is closest (most similar), according to a distance measure. Table 1 gives a comparison of the best results obtained in these experiments with those of three other researchers using the same corpus; see Mason et al. [12] for a more detailed discussion of these classification models, experiments, and results.

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santini [17]</td>
<td>0.906</td>
</tr>
<tr>
<td>Kanaris and Stamatos [9]</td>
<td>0.965</td>
</tr>
<tr>
<td>Dong et al. [7]*</td>
<td>0.965</td>
</tr>
<tr>
<td>Mason et al. [12]</td>
<td>0.946</td>
</tr>
</tbody>
</table>

Table 1. Best classification accuracy results for the 7-Genre corpus. (*The results reported in [7] refer to a sub-corpus of four genres.)

3. Measuring the Similarity Between a Web Page and a Web Page Genre

The first phase of this research determined the similarity of a Web page profile and a genre profile using a distance function suggested by Kešelj et al. [10] in their paper on the use of n-gram profiles for authorship attribution. The second phase of this research consisted of experiments to test 14 other distance measures for use in the n-gram classification model. These distance functions were selected based on those examined by Tomović et al. [24], and include measures based on the arithmetic mean, the geometric mean, the quadratic mean, and the Euclidean mean.

\(^1\)http://users.cs.dal.ca/~vlado/srcperl/Ngrams/

\(^2\)http://www.itri.brighton.ac.uk/~Marina.Santini
distance. Somewhat surprisingly, none of the distance functions tested in these experiments out-performed the initial distance measure in terms of precision, recall, and classification accuracy. Based on these results, the initial distance function was selected for use in subsequent work. We therefore define the distance between two n-gram profiles as

\[
d(P_1, P_2) = \sum_{m \in P_1 \cup P_2} \left( \frac{2 \cdot (f_1(m) - f_2(m))}{f_1(m) + f_2(m)} \right)^2,
\]

where \(f_1(m)\) and \(f_2(m)\) are the frequencies of n-gram \(m\) in the two profiles, \(P_1\) and \(P_2\) respectively.

4. Feature Selection Measures

The term *curse of dimensionality* was coined by Richard Bellman [1] to describe the problem that occurs when searching in high dimensional spaces. As the dimensionality of the input data space increases, it becomes exponentially more difficult (more computationally complex) to fit models for the parameter space. In practice, it is often necessary to use feature selection techniques to select a subset of relevant features, because most standard machine learning techniques cannot be directly applied when the dimensionality is very high. Yang and Pedersen [26] provide a comparative study of the traditional feature selection techniques in text classification. In their experiments, they found that Information Gain and the Chi-square test were the most effective measures for feature selection. Focusing specifically on the classification of Web pages by genre, Dong et al. [7] compared the performance of Information Gain, Mutual Information, and the Chi-square statistic for selecting features. They found that although all three of the selection measures were capable of detecting small sets of discriminating features, when feature sets were as small as 5, only Information Gain and the Chi-square statistic were able to successfully select features that gave good performance.

Although high classification accuracy was achieved in the initial Web page genre classification experiments discussed in Section 1, using frequency as a feature selection measure [12], it was hypothesized that a more theoretically sound feature selection measure could be more effective. Thus, the next phase of research investigated not only frequency, but also the Chi-square statistic and Information Gain as feature selection measures for choosing the n-grams with which to represent each Web page. The Chi-square statistic is a statistically based measure, while Information Gain is probability based. See, for example, Yang and Pedersen [26] for further details.

The feature selection experiments were run on the 7-Genre data set, as well as on the KI-04 data set\(^3\). The KI-04 data set, constructed by Meyer zu Eissen and Stein [13] has eight genres that were suggested by participants in a user study on the usefulness of Web page genres. The original corpus includes some empty Web pages, and so we follow the lead of Santini [17] and Chaker and Habib [4] in using the 1205 non-empty pages. The number of Web pages per genre ranges from 126 to 205. In these experiments the n-gram length was varied from 2 to 10 in increments of 1, and for each n-gram length, the number of n-grams used in the Web page profiles was varied from 5 to 500, in increments of 5 from 5 to 50, and in increments of 25 from 50 to 500. As before, all of the experiments were run using 10-fold cross validation.

This examination of the effect of feature selection measures on the automatic classification of Web pages by genre showed that the type of feature selection method, the n-gram length, and the Web page profile size each have a significant effect on the performance of the classification model, measured in terms of mean classification accuracy. The Chi-square statistic was the most effective feature selection method, achieving very high accuracy with Web page profiles as small as 5. Although high classification accuracy was achieved with each n-gram length, from 2 to 10, the best classification accuracy on each data set was found using n-grams of length 2, indicating that this n-gram length could be a good first choice for use with other data sets. The results indicated that small profile sizes of between 10 and 40 are excellent choices for use with this model, particularly when coupled with the use of n-grams of length 2.

As shown in Table 2, each of the feature selection measures allows the Web page genre classification model to achieve a high mean classification accuracy for n-grams of length 2 to 10 with Web page profile sizes of 5 to 500. This gives very strong support for the use of n-gram representations of Web pages and Web page genres. Using the Chi-square statistic as a feature selection measure gives the best performance on each of the data sets.

<table>
<thead>
<tr>
<th>Feature Selection Measure</th>
<th>7-Genre</th>
<th>KI-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.828</td>
<td>0.900</td>
</tr>
<tr>
<td>Information Gain</td>
<td>0.887</td>
<td>0.918</td>
</tr>
<tr>
<td>Chi-square Statistic</td>
<td>0.955</td>
<td>0.969</td>
</tr>
</tbody>
</table>

\(^3\) http://www.itri.brighton.ac.uk/~Marina.Santini
5. Data Preprocessing

The experiments discussed thus far in this paper were all run on data sets in which the Web pages had been preprocessed to remove both the HTML tags and JavaScript code, however the question arose as to whether such preprocessing steps were beneficial or detrimental to the classification accuracy achieved by the model. In order to investigate the effects of typical data preprocessing steps when using an n-gram approach to Web page representation, a preliminary study was conducted in which frequency was used as the feature selection measure. The experiments were run with three different levels of preprocessing on the data sets; these levels were no preprocessing, removing only the JavaScript code, and removing both the HTML tags and the JavaScript code. The experiments were run on both the 7-Genre and KI-04 data sets, using 10-fold cross validation in each case.

The results of these experiments were somewhat inconclusive: the same level of preprocessing was not found to be superior on both data sets. However, the classification model was able to achieve very high classification accuracy using byte n-gram Web page profiles even when no preprocessing of the data set was performed. See Mason et al. [11] for a more detailed discussion of these experiments and results.

After the completion of the experiments on feature selection measures, discussed in Section 4, the preprocessing experiments were run again. For this set of experiments, the Chi-square statistic (rather than frequency) was used as the feature selection measure for choosing the n-grams with which to represent each Web page. In these experiments the length of the n-gram was varied from 2 to 7, and for each n-gram length, the number of n-grams used in each Web page profile was varied from 5 to 50, in increments of 5. As in the previous experiments, all of the experiments were run using 10-fold cross validation.

The results for these experiments, as shown in Table 3, showed that the mean classification accuracy on each data set was the same whether no preprocessing was done, or whether the JavaScript code was removed. However, each of these two levels of preprocessing gave better results than removing both the HTML tags and the JavaScript code. This is a very interesting finding, because it is an indication that the presence of the HTML tags is beneficial to the classification model. Although the presence of the JavaScript code is not beneficial to the classification accuracy, neither is it detrimental, therefore we concluded that preprocessing the Web pages by removing the JavaScript code is not necessary with this classification model.

6. Multi-labeling of Web Pages

Most research on the classification of Web pages by genre has focused on labeling each Web page as belonging to a single genre. However, researchers who have conducted surveys and user studies concerning Web page genre have acknowledged the difficulty of assigning a single genre label to a Web page [6, 13, 16]. In response to this problem, Santini has proposed a zero-to-multi genre classification scheme [17, 18]. Santini's user studies indicate that, at least from a user's perspective, a single-genre classification scheme is too narrow. Although some Web pages may fit neatly into a single genre, others do not fit into any genre, while still others may fit appropriately into several genres. Santini suggests designing genre classification models that not only allow Web pages to have more than one genre label, but that also allow a Web page to have a zero-genre label. This latter would be for the case in which a Web page does not match the conventions of any known genre. Labeling Web pages in this manner is not only a more flexible approach from the users' perspective, but also better reflects the complex nature of the mixture of pages found on the World Wide Web.

In order to classify a Web page as belonging to more than one genre, or as being of no known genre, the classification model that has been developed to date was modified to include thresholds that are computed for each genre. If the distance between the Web page profile and a genre profile is less than the threshold, then the Web page is labeled as belonging to that genre. The following two methods of setting the genre thresholds were investigated.

One method of determining the threshold for each Web page genre is to base the threshold on the distribution of the Web pages that belong to that genre, in the training set. For each genre, the distance is computed between the genre profile and the profile of each Web page of that genre in the training set. If this set of distances has a normal distribution, the genre threshold is set at the \(85^{th}\) percentile (one standard deviation around the mean), such that 85% of the Web pages belonging to that genre (in the

<table>
<thead>
<tr>
<th>Data Preprocessing Level</th>
<th>7-Genre</th>
<th>KI-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML tags &amp; JavaScript removed</td>
<td>0.981</td>
<td>0.980</td>
</tr>
<tr>
<td>JavaScript removed</td>
<td>0.995</td>
<td>0.992</td>
</tr>
<tr>
<td>No preprocessing performed</td>
<td>0.995</td>
<td>0.992</td>
</tr>
</tbody>
</table>

Table 3. Mean classification accuracy for data preprocessing levels, averaged over n-gram lengths of 2 to 7 and Web page profile sizes of 5 to 50.
The data set contains 1539 Web pages, each with one or and Andrej Bratko, and described in Vidulin et al. [25]. Because each genre threshold is set based on the distribution of the Web pages belonging to the genre, we will refer to this as the distribution method.

Another method of setting each genre threshold is to first order all of the Web pages in the training set according to their distance from a particular genre profile, in ascending order. This ordered list of Web pages from the training set is then stepped through one Web page at a time, such that at each step, one Web page is labeled as belonging to the genre in question, and the accuracy of the classification is computed. In this manner, the optimal threshold for that genre, based on the training data, can be determined; we therefore refer to this as the optimal threshold method.

The proposed methods for setting genre thresholds were tested on the 20-Genre data set. This data set is a multi-labeled Web page collection constructed by Mitja Luštrek and Andrej Bratko, and described in Vidulin et al. [25]. The data set contains 1539 Web pages, each with one or more genre labels. Of the 1539 Web pages, 1059 have one genre label, 438 have two genre labels, 39 have three labels, and 3 have four labels. This gives a total of 2064 labels. The number of Web pages belonging to each genre ranges from 55 to 227, making this a highly unbalanced data set.

In these experiments the length of the $n$-gram was varied from 2 to 5, and for each $n$-gram length, the number of $n$-grams in each Web page profile was varied from 5 to 50, in increments of 5. All of these experiments were run using 3-fold cross validation. As shown in Table 4, the proposed optimal threshold method outperformed both the proposed distribution method for setting genre thresholds, and the bagging algorithm used by Vidulin et al. [25]. Although the precision of the classification is relatively high, the recall is much lower; the classification model is assigning fewer genre labels than were given by the human classifiers who originally labeled the Web pages.

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vidulin et al. [25]</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>Bagging algorithm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mason Distribution method</td>
<td>0.93</td>
<td>0.36</td>
</tr>
<tr>
<td>Mason Optimal threshold method</td>
<td>0.98</td>
<td>0.72</td>
</tr>
</tbody>
</table>

7. Conclusions and Future Work

This research is ongoing; the current phase of research compares the classification results of the proposed Web page classification model, which uses a distance measure approach to compare $n$-gram profiles, with the classification results of two well-known machine learning methods, the $k$-nearest neighbor ($k$nn) method, and the support vector machine (SVM) method. In each case, no preprocessing of the data set is performed, and the Web pages are represented by $n$-gram profiles in which the Chi-square statistic is used as a feature selection measure.

The final phase of this research will test the proposed classification model on a highly unbalanced data set to which noise has been added, and if necessary, refine the classification model to better deal with such real world conditions. For the purposes of this study, noise is defined as any Web page belonging to a Web page genre that is not one of the genres in the data set. The optimal threshold method, discussed in Section 6, will be used to identify noise Web pages during the classification process.

The direct contributions of this thesis will be primarily in the area of Web page genre classification, however by developing a new approach to Web page genre classification, this thesis will provide advances in the state of the art that will be of interest to both the text classification and Web page classification communities. By evaluating the proposed models on established data sets and making public the results, the resources available to other researchers are extended. This research includes developing a Web page genre classification approach that is effective not only on balanced data sets, but also on unbalanced and multi-labeled data sets. Classification on unbalanced data sets is an extremely challenging problem; therefore this work will contribute to the field of classification in general, and be of interest to researchers performing classification on unbalanced data sets of any type. Because this research involves developing $n$-gram representations for the Web pages and Web page genres, it will also contribute to any field which $n$-gram representations can be used. The use of $n$-grams is language independent, even going beyond traditional languages to those of, for example, music and genetic sequences, therefore the fields of research that could potentially benefit from this work are many.

Acknowledgements

This research has been supported by the Killam Trust and the Natural Sciences and Engineering Research Council of Canada (NSERC), and supervised by Dr. Michael Shepherd.
Bibliography


