Web Page Genre Classification: Impact of n-Gram Lengths

K. Pranitha Kumari
Assistant Professor
Department of Computer Science and Engineering
Osmania University
Hyderabad

A.Venugopal Reddy, Ph.D
Professor
Department of Computer Science and Engineering
Osmania University
Hyderabad

S.Sameen Fatima, Ph.D
Professor
Department of Computer Science and Engineering
Osmania University
Hyderabad

ABSTRACT
Web pages are discriminated based on their topic and genre. Web page genres are capable to improve the modern search engines to focus on the user's information need. In this paper, web pages are represented using character n-grams. Character n-gram representation is language independent and allows automatic extraction of features from a web page. Character n-gram representation of a web page can be used efficiently to classify a web page by genre. Support Vector Machine (SVM) classification model is used for classification and experiments were carried out on 7-Genre corpus by varying the length of n-grams. It is observed that the performance in terms of F-measure improves as n-gram lengths are varied from 3 to 5 and it is also observed that performance degrades as the n-gram length is further increased.

General Terms
Genre classification, corpus

Keywords
Character n-gram feature extraction, n-gram length, web page representation, SVM classifier and term frequency

1. INTRODUCTION
Web page genre classification is helpful in filtering the results of online searches. In order to classify Web pages by genre, it is necessary to extract features that effectively separate each web page and genre. In web page genre classification, web pages are represented through different ways such as keywords, n-grams, POS tags, HTML tags, URL contents, document level features, genre relevant words etc.

An n-gram is defined as an n-word, n-character or n-byte substring of a given string. Word n-grams are simply sequences of n consecutive words. Character n-grams use letters only and a non-letter character is substituted by a space, and two or more consecutive spaces are replaced by a single space. An underscore character is used to represent the space. Byte n-grams are unprocessed form of character n-grams in which no bytes are excluded, including the whitespace characters.

There are several advantages of n-gram feature extraction method over other feature extraction methods such as 1) N-grams capture the stems of the words automatically. For example, the words extraction, extracting and extracted are different, but they have some common n-grams, 2) N-grams are capable of dealing with OCR (Optical Character Recognition) generated errors, or orthographic errors done by humans and 3) N-grams are language independent. The key limitation of n-gram feature extraction is, as the length of the n-gram increases, dimensionality of feature set will increase [20].

An n-gram representation of web page is language independent [7][8][9]. An n-gram method can be comfortably applied to any language and also to non-language data such as music and DNA [16][17]. It sufficiently obtains the characteristics of web page and can be easily adapted to the properties of the emerging web genres in the World Wide Web. An n-gram feature extraction doesn't require any text preprocessing or higher level processing, so it avoids the requirement for use of taggers, parsers and stemmers in the feature selection, or any other language-dependent and non-trivial NLP tools.

In this paper, web pages are represented using fixed-length character n-grams. Character n-grams of different lengths (n=3 to 8) are considered. Experiments were run on 7-Genre corpus [12]. SVM classifier is applied for classification of web genres. The classification results show that n-gram of length five gave better results. The results obtained were compared with the existing work.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work on webpage genre classification and web page representation. Section 3 describes the extraction of character n-gram features and the corpus used in this work. In Section 4, the results of the experiments and comparison to previous work are discussed. Finally, Section 5 concludes the paper with future work directions.

2. RELATED WORK
This section deals with work related to n-gram representation of web page and web page genre classification. Web page genre classification is described in terms of three factors: the set of genres, the features used to represent web pages in each genre and the classification algorithm used to distinguish genres.

Web genre palette consisting of eight genres: article, download, link collection, portrayal-private, discussion, help, portrayal-non private and shop called as KI-04 corpus [10] based on web genre usefulness. In [13], a collection of 20 genre labels: Adult, Blog, Children's, Commercial, Community, Content delivery, Entertainment, Error message, FAQ, Gateway, Index, Informative, Journalistic, Official, Personal, Poetry, Prose fiction, Scientific, Shopping and User input were allowing multiple labels to be assigned to one webpage. Although there are certain resemblances in genre

corpora (e.g. personal home page, listing etc.) it is not clear how each genre of one corpus is associated to the genres of another corpus. Home pages and non-home pages are distinguished in [4] and classified as personal home pages, corporate home pages and organizational home pages genres. [13] Focused on PDF web documents and a collection of six genres: academic monograph, business report, book of fiction, minutes, periodicals, and thesis. Issues related to corpus, features and classification algorithms discussed in [14] drawn conclusions from experiments on a focused genre palette may be misleading since features found to be useful in some genres are not equally effective in other genres.

Generally, the features used to represent webpage in webpage genre classification are a combination of different features. Bag-of-words, part-of-speech tags, various punctuation symbols, and function words, document level such as: sentence length, number of words and word length are used as feature set in [1], URL features [2], HTML tags [3] and Bag-of-words [6].

Combined stemming approach (CSA) is proposed in [15] to extract root words and genre relevant words. Random Forest machine learning algorithm was used on 7-Genre corpus for classification. To identify home pages on the Web [4] used features as triple: content, form and functionality. Content, form, functionality and positioning features are used by [5] to identify front pages of news paper on the web. Kanaris [9] proposed a function called as 'glue' that sticks the characters together within an n-gram. The proposed function considered 3-grams and 5-grams against 4-grams, since they can capture both sub-word and inter-word information. The n-gram features can reduce problems derived from typographical errors and spelling mistakes done by humans [19]. Byte n-grams by Mason J. in [18], and the combination of bag of words, genre relevant words, Part-of-speech (POS) tags and HTML tags by Santini M. [12] were considered as features and genre classification was performed by using SVM classifier.

3. EXPERIMENTS

3.1 Web genre corpus

7-Genre corpus was built in early 2005, comprising seven genres identified by Santini M. reported in [12] and consists of 1,400 English web pages as shown in Table 1. It is a balanced corpus in which each genre comprised of equal number of web pages.7-Genre corpus construction followed the criteria of annotation by objective sources and consistent genre granularity except listing.

Table 1. 7-Genre corpus

Genres	WebPages
Blog	200
Eshop	200
Faq	200
frontpage	200
Listing	200
Php	200
Spage	200

3.2 Classification model

The classification model used in this paper is Support Vector Machine (SVM). The SVM method is a popular and well known supervised machine learning method which performs classification by constructing an *N*-dimensional hyper plane that optimally separates the data. The vectors near the hyper plane are the support vectors.

3.3. Character n-gram extraction

The representation of web page using character n-grams is mainly dependent on the length of the n-gram. Character n-grams are sequence of 'n' characters where 'n' is equal to the length of n-gram. Character n-grams of length 3 to 8 were extracted and the n-gram feature set was selected using term frequency (TF). TF was calculated based on the number of occurrences of each n-gram in a web page.

3.4 Performance evaluation measures

The classifier performance was evaluated in this paper by using classifier performance measures precision, recall, F-measure and accuracy. Precision is the number of correctly classified true positive instances by the number of instances labeled by the system as positive. Recall is the number of correctly classified true positive instances divided by the number of positive instances in the data. F-measure is the harmonic mean of precision and recall.

$$Recall = \frac{TP}{TP + FN}$$

$$Precsion = \frac{TP}{TP + FP}$$

$$\boldsymbol{F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}}$$

$$Accuracy = \frac{Number\ of\ correctly\ classified\ documents}{Total\ number\ of\ documents}$$

TP: True Positive
FN: False Negative
FP: False Positive

4. RESULT ANALYSIS

This section describes the results of the experiments carried out on 7-genre corpus. Character n-grams of different lengths (n=3 to 8) were tested on 7-genre corpus and 10-fold cross-validation was used for the experiments with this corpus. The classification was done using SVM classifier and the classification results are shown in Table 2 and Figure 1 respectively. The experimental results obtained on the 7-Genre corpus in terms of F-measure are shown in Table 2. By analyzing the results of Table 2, it is found that the performance improved by incrementing the value of 'n' from 3 to 5 and started declining from 6 onwards. For n-gram of length 5, better classification results were obtained, as 5-gram captures most of the root words in English web pages.

Genre wise comparison of mean precision, recall and F-measure are shown in Table 4, which indicates that as the n-gram length is increased from 3 to 5, the precision, recall and F-measure are improving and from 6 onwards it degrades. It is concluded that using an n-gram of length 5 gave better classification results in terms of precision, recall and F-measure as it captures most of the English words and stems. The advantage of 5-gram is it maintains the dimensionality of

the problem in a reasonable level when compared to n-gram of length greater than 5

Table 2: Classification results obtained on the 7-Genre corpus for n-gram lengths 3 to 8

N-gram length	F-measure
3	80.8%
4	94.9%
5	95.8%
6	92.7%
7	90.5%
8	88.7%

Table 3 summarizes the comparison of classifier performance with existing research [9], [12], [15] and [18]. Santini M. [12] considered the combination of bag of words, genre relevant words, Part-of-speech (POS) tags and HTML tags as features and classification was done using SVM classifier. Kanaris [9] considered the combination of character 3-grams and 5-grams

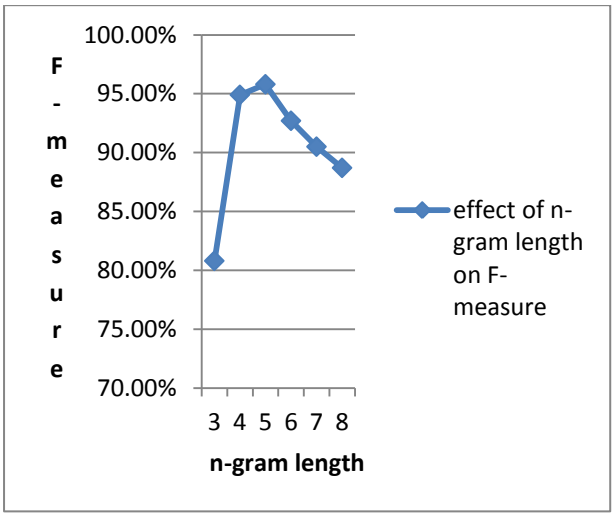

Figure 1. Classification results obtained on the 7-Genre corpus for n-gram lengths 3 to 8

respectively and SVM model was applied to build the webpage genre classifier. Byte n-grams were considered as features by Mason J. [18], and classification was performed using optimal threshold method. Combined stemming approach (CSA) [15] used genre relevant words as features and Random Forest machine learning algorithm was used for classification. From Table 3 and Figure 2, it is clear that, by considering the character n-grams of length 5 gave better classification performance results on 7-genre corpus.

Table 3. Comparison of classifier performance with existing research

Researchers	Performance
Santini M.	90.6%
Kanaris	96%
Mason J	94.6%
CSA	91.5%
Our result	95.78%

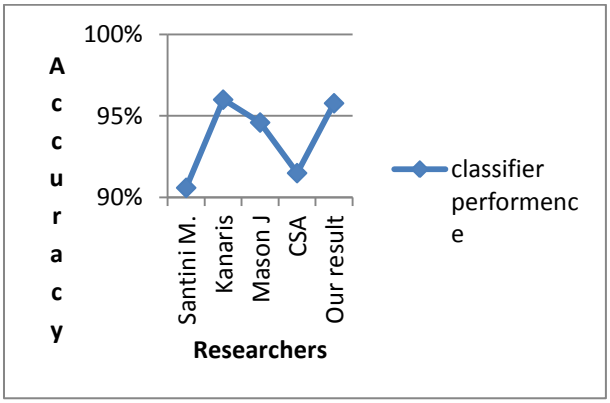

Figure 2. Comparison of classifier performance with existing research

Table 4. Average precision, recall and F-measure over character *n*-gram lengths of 3 to 8

n-gram length	Performance measures	Blog	Eshop	Faq	frontpage	Listing	Php	Spage	Average
3	Precision	0.802	0.701	0.923	0.881	0.919	0.763	0.679	0.81
	Recall	0.79	0.75	0.905	0.85	0.91	0.725	0.72	0.807
	F-measure	0.796	0.725	0.914	0.865	0.915	0.744	0.699	0.808
4	Precision	0.965	0.913	0.995	0.971	0.995	0.863	0.957	0.95
	Recall	0.97	0.895	0.985	1	0.99	0.91	0.895	0.949
	F-measure	0.963	0.904	0.99	0.985	0.992	0.886	0.925	0.949
5	Precision	0.98	0.941	1	0.985	1	0.865	0.949	0.959
	Recall	0.995	0.875	0.995	1	0.995	0.92	0.925	0958.
	F-measure	0.998	0.907	0.997	0.993	0.997	0.887	0.937	0.958
6	Precision	0.955	0.856	1	0.985	1	0.86	0.831	0.927
	Recall	0.965	0.86	0.985	0.985	0.995	0.86	0.835	0.926
	F-measure	0.96	0.858	0.992	0.985	0.997	0.86	0.833	0.927
7	Precision	0.95	0.855	0.995	0.985	0.939	0.854	0.827	0.915
	Recall	0.911	0.856	0.95	0.946	0.92	0.86	0.835	0.896
	F-measure	0.93	0.855	0.97	0.965	0.93	0.854	0.831	0.905
8	Precision	0.927	0.82	0.962	0.979	0.89	0.84	0.82	0.891
	Recall	0.916	0.849	0.942	0.936	0.887	0.84	0.83	0.885
	F-measure	0.921	0.834	0.951	0.957	0.888	0.84	0.824	0.887

5. CONCLUSION

A low-level feature extraction approach called as character n-grams were used for webpage genre classification. The classification results obtained show that, n-gram length is important in the web page genre classification. An n-gram of length five gave better results on 7-genre data set, as 5-gram captures most of the root words in English web pages. Applying n-gram method on Indian languages in web page genre classification can be considered as a future work. N-grams are much less sensitive to OCR-generated errors. When dealing with languages that require non-trivial feature extraction, n-grams provide a language-independent feature extraction technique.

6. REFERENCES

[1] Aidan Finn and Nicholas Kushmerick, Learning to classify documents according to genre, Journal of the American Society for Information Science and Technology, volume 57, pages 1506-1518,2006.

[2] Vidulin V., Lustre, M., Gams M., "Training the Genre Classifier for Automatic Classification of Web Pages", in Proceedings of the 29th International Conference on Information Technology Interfaces, pp.93-98, 2007.

[3] Akira Maeda and Yukinori Hayashi, Automatic Genre Classification of Web Documents Using Discriminant Analysis for Feature Selection.

[4] Alistair Kennedy and Michael Shepherd, Automatic Identification of Home Pages on the Web, Proceedings of the 38th Hawaii International Conference on System Sciences – 2005.

[5] Carina Ihlström and Maria Åkesson, Genre Characteristics - a Front Page Analysis of 85 Swedish Online Newspapers, Proceedings of the 37th Hawaii International Conference on System Sciences – 2004.

[6] Jebari Chaker and Ounelli Habib, Genre categorization of web pages, Seventh IEEE International Conference on Data Mining – Workshops, 2007.

[7] P Majumder, M Mitra, B.B. ChaudhuriN-gram: a language independent approach to IR and NLP

[8] German Aquino, Waldo Hasperue1, Cesar Estrebou1 and Laura Lanzarini, A Novel, Language-Independent Keyword Extraction Method, 2013.

[9] Ioannis Kanaris and Efstathios Stamatatos, Learning to Recognize Webpage Genres, 2009.

[10] Meyer zu Eissen, S. and B. Stein "Genre Classification of Web Pages: User Study and Feasibility Analysis". In Biundo S., Fruhwirth T. and Palm G. (eds.). KI 2004: Advances in Artificial Intelligence, Springer, pp. 256-269, 2004.

[12] Sanitni, M. Automatic Identification of Genre in Webpages. Ph.D. Thesis, University of Brighton, 2007.

[13] Kim Y. and Ross S. "Examining Variations of Prominent Features in Genre Classification, In Proc. of the 41st Annual Hawaiian International Conference on System Sciences (HICSS), 2008.

[14] Boese, E and A. Howe, "Effects of Web Document Evolution on Genre Classification", Proc. of the ACM 14th Conference on Information and Knowledge Management, 2005.

[15] K. Pranitha Kumari and A. Venugopal Reddy, Performance provement of Web Page Genre Classification, International Journal of Computer Applications (0975 – 8887) Volume 53– No.10, September 2012.

[16] M. Nelson and J.S. Downie. Informetric Analysis of a Music Database. *Scien-tometrics*, 54(2):243{255, 2002.

[17] I.S.H. Suyoto and A.L. Uitdenbogerd. Simple e±cient n-gram indexing for effective melody retrieval. In *Proceedings of the First Annual Music Information Retrieval Evaluation eXchange*, September 2005.

[18] Mason, J.E., M. Shepherd, and J. Duffy (2009). "An N-gram Based Approach to Automatically Identifying Web Page Genre". In *Proc. of the 42nd Hawaii International Conference on System Sciences*.

[19] Pollock, J.J. and Zamora, A. : System design for detection and correction of spelling errors in scientifc and scholarly text, *Journal of American Society for Information Science*, 35 (1984)104-109

[20] Artur ·Silic, Jean-Hugues Chauchat, Bojana Dalbelo Basic, and Annie Morin, N-grams and Morphological Normalization in Text Classification: a Comparison on a Croatian-English Parallel Corpus, Progress in Artificial Intelligence, 671—682, Springer, 2007.