Performance Improvement of Web Page Genre Classification

K. Pranitha Kumari
Assistant Professor
Department of Computer Science and Engineering
University college of Engineering
Osmania University
Hyderabad, A.P., India.

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

ABSTRACT

The dynamic nature of web and with the increase of the number of web pages, it is very difficult to search required web pages easily and quickly out of thousands of web pages retrieved by a search engine. The solution to this problem is to classify the web pages according to their genre. Automatic genre identification of web pages has become an important area in web page classification, because it can be used to improve the quality of web search results and also to reduce the search time. In this paper, a Combined Stemming Approach (CSA) is proposed to extract genre relevant words and to classify web pages by genre (non- topical) based on word level and linguistic features. Experiments were performed on 7-genre corpus. In order to improve the accuracy of the results, we applied combined stemming and stop word elimination techniques. The proposed approach of extracting features discriminates web pages by genre. The classification results obtained using Random Forest classifier was compared with the results of other researchers, who worked on the same corpus. It is shown that the method proposed is superior in performance in terms of accuracy.

General Terms

Classification, Stemming

Keywords

Web page classification, Genre, Corpus, Feature Extraction, Combined Stemming Approach

1. INTRODUCTION

As the World Wide Web is both fast-paced and dynamic in nature, web page classification becomes increasingly important in web searching. Web page classification is one of the essential techniques for mining the web. Web page classification assigns a web page to one or more predefined class labels. According to the type of category, classification can be divided into sub-problems such as topic classification, sentiment classification, genre classification, and so on Xiaoguang QI et al. [1]. Web page Classification based on genre and genre analysis is a promising research area. More attention has been given to automatic genre identification of web pages because it can be used to improve the quality of web search results and to implement genre based search Mehler, A et al.[2].

Genre is one of the basic properties of web page. The term *genre* comes from the Greek word *genus*, meaning *kind* or *sort.* There is no definitive understanding among researchers as to what is meant by the term genre A. Finn et al. [3]. The effective way of retrieving a web page, depends on its genre rather than the content and is useful in many areas. For example, a retrieval query about a certain topic such as “Oracle” gets many documents related to the company “Oracle” from an Internet search engine, but these documents belong to different genres such as company homepage, product specification, product advertisement and review of a certain product. If genre could be integrated into a content-based search system, the classification of documents would be totally different and more helpful for the user. In genre based classification, genre revealing features are very important to discriminate web pages belongs to different genres. In this paper, a Combined Stemming Approach (CSA) is designed to extract genre based features and stem words along with stylistic features.

Genre classification is applied in many ways, such as (1) Implementation of genre based search engines, (2) Creating and developing directory structures, (3) Profiling is another area of research that could benefit from genre analysis, (4) Query reformation is a technique that can be used by search engines to help refine a user’s query for improving results and (5) Filtering has become another important aspect in the electronic world, with needs for spam filters as well as filters for children.

The issues of web page genre classification are discussed in Boese E.S[4], Santini M.[5] such as corpus related issues, classification model related issues and feature related issues. Corpus related issues are related to the size of the corpus and ambiguity of data in each genre of different corpus. Feature related issues are related to feature extraction of web pages referred in Lijuan J. et al. [6]. Classification model related issues gives different performance results on different corpora. Some other issues are the (1) lack of an established genre list, (2) the unclear relation between traditional and web genres, (3) the need to classify large quantities of web pages quickly, (4) the design of the genre inventory, and (5) the identification of emerging genres in the web.

This paper is organized as follows: related work is presented in section 2, Section 3 describes the extraction of content features as well as the extraction of linguistic features, section 4 introduces the corpora used in this study, describes the various performance measures and also presents the experimental results and finally section 5 concludes the paper.

2. RELATED WORK

In order to classify web pages by genre using a machine learning approach, it is necessary to identify features that effectively characterize each web page by genre. According to Shepherd et al. [7], the main goal is to extract a set of features, which will allow the classifier to discriminate the genres and assign the correct genre label to each Web page. The representation of web pages used in the genre classification task tend to be based on those used in text classification, however the web page representations may be augmented with low level as well as high level information such as n-grams, HTML tags, JavaScript code, URL information and Part-of-Speech (POS) tags.

Genres are represented in different ways by different researchers, Lei D et al. [8] represented the genre based on content, style, and functionality. According to Lei the genre is represented as a triple <content, form, functionality> known as cyber genre. Content is a text visible to the user via the web browser. Form can be HTML tags such as title, head, font, bullet, style, table, tr, td. Functionality can be HTML tags such as applet, script, jsp, link, form, select, option, text area, input. Lei has considered the above features and performed experiments on four web genres, and got an accuracy of 96.5%.

Santini M.[9] considered combinations of feature types such as most common words in english, POS tags, punctuation symbols, genre-specific facets, HTML tags and URL based tags. SVM classifier, KNN classifier and an inference model were used for automatic genre identification, and got an accuracy of 90.6%, 67% and 86% respectively using 7-genre corpus. Santini M.[9][10]][11][12][13] examined classification problems in terms of two broad textual phenomena: genre hybridism and individualization. *Genre hybridism* describes multi-genre variation within the individual web page, while *individualization* refers to absence of any recognized genre in a web page, this phenomena also called as zero to multi genre classification . The identification of these two phenomena helps to find the range of flexibility of an automatic classification system. Boese E.S. [4] used three different classes of features, which are named as stylistic features, form features and content features, such as style comes to the structure and readability of the web page, form refers to the presentation layout or format of the web page such as text statistics, HTML analysis and content off the web page consists of bag-of-words (BOW), words in HTML, title tag and URL, number types, closed-world sets and punctuations.

The Porter stemming algorithm Willett P. [14] considers common features of English, rare suffixes are not included, there is no equivalent of Lovins' transformation rules, other than rule (1), the undoubling of terminal double letters. Secondly, it removes suffixes only when the remainder stem is fairly substantial. Some suffixes are removed only when at least one syllable is left, and most are removed only when at least two syllables are left. The Porter stemming algorithm is traditional in its removal of suffixes. Thirdly, it removes suffixes in a series of steps, often reducing a compound suffix to its first part, so a step might reduce “ibility” to “ible”, where “ibility” is thought of as being “ible + ity”. Although the description of the whole stemming is little bit complicated. For some words like “happy” it will give “happi”. “Happi” is not having semantic meaning. Lovin's stemming algorithm Lovins, J.B. [15] will give root words, but some genre relevant words, for example “stories”, “assistance” and “directories”, lovin's stemmer is giving stemm words “st”, “as” and “direct” respectively instead of “story”, “assist” and “directory”.

Genre organization is either flat or hierarchical. Hierarchical organization genres are either super genres or sub genres as reported in Lindemann [16], Santini M.[9], Shepherd[17][7], Ezeiza Ramos J [18]. Lindemann [16] proposed, combination of classification by structural and content features and shown that the significant improvement of the overall accuracy of classification.

3. PROPOSED APPROACH

In order to characterize the web page based on genre, a combined stemming approach (CSA) is proposed, which is a combination of both porter stemming and lovin's stemming techniques. The limitations those were present in both the techniques to extract genre relevant words, are removed using CSA. Pre-processing is a basic step in web page classification. Preprocessing was performed to eliminate stop words, HTML tags, and java script code. CSA is applied to extract all root words, along with genre relevant words. To reduce the dimensionality of feature set correlation feature selection (cfs) subset evaluation, term frequency and term rank were applied. Web page is represented by content features such as set of words and stylistic features such as part-of speech (POS) tags. CSA is a simple and reliable technique to extract features when compared to earlier methods that were used in the literature. POS tags were extracted using Stanford penn tree bank tagger [19] and the content features of the web page were extracted using word frequency and word ranking feature selection measures.

3.1 Classification method

The classification model used in this study is Random forest classifier. The Random forest classifier is an ensemble supervised machine learning classifier. It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier. It runs efficiently on large data sets. It can handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. Random forest classifier along with the feature extraction method and attribute selection (i.e. cfssubset evaluation) technique gave better results compared with the existing results of other researchers as shown in Table-2, Table-3 and Table-4 respectively.

4. EXPERIMENTS

4.1 Corpus

The 7-genre corpus used in this study was developed by Santini M[9][20]. This corpus is composed of 1400 English web pages. These web pages are evenly balanced with 200 web pages in each of the seven genres as shown in Table 1. The granularity of the collection is consistent, with the exception of the LISTING genre, which can be decomposed into the subgenres Checklist, Hotlist, Sitemap, and Table of contents.

Table 1: Composition of the 7-Genre corpus

Web Genres	Number of web pages
Blog	200
Eshop	200
Faq	200
Online news paper frontpage	200
Listing	200
Php	200
Spage	200

4.2 Performance evaluation measures

The performance measures used in this study to evaluate the result of the classifier are 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.

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

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

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

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

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

4.3 Experimental results

The experiments were performed on 7-genre corpus, 6-genre corpus (excluding listing super genre) and 4-genre corpus (eshop, faq, php and news) using Random forest classifier along with feature extraction and attribute selection techniques respectively. The results obtained show improved accuracy when compared with Santini M[9] and Lei D et al. [8] as shown in Table-2. Table 3 shows the performance in terms of F-measure, the results are depicted in figure-1. The proposed approach (CSA) also yields high precision and recall results, indicating that the genre labels assigned by the classifiers are quite accurate as shown in Table-4 and the results are illustrated in Figure-2 and Figure-3.

Table 2: Comparison of results in terms of accuracy

Web genres	Santini M	Lee Dong	CSA
7-genre	90.4%	--	91.5%
6- genre	94.5%	--	95.6%
4-genre	--	96.5%	97.4%

Table 3: Performance of the classifier in terms of F-measure

Web Genres	**Santini**	**CSA**
Blog	0.938	0.946
Eshop	0.894	0.905
Faq	0.948	0.956
Online news paper frontpage	0.983	0.987
Listing	0.813	0.823
Php	0.865	0.874
Spage	0.9	0.915

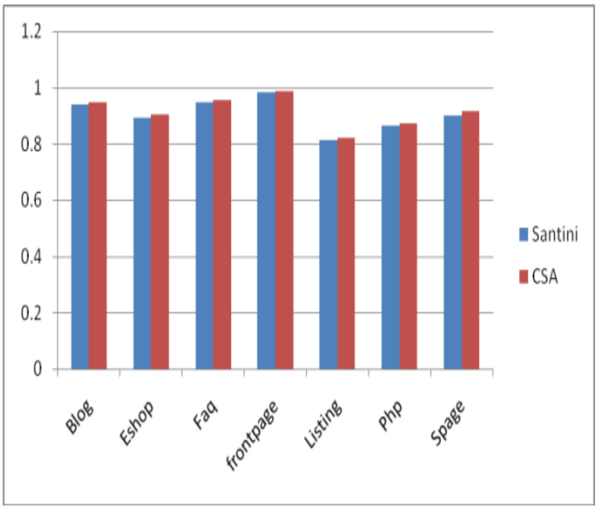

Figure-1: Comparison of result in terms of F-measure

Table 4: Comparison of CSA and Lei D. in terms of precision and recall

Web Genres	Lei D		CSA	
	Precision	Recall	Precision	Recall
Eshop	0.920	0.902	0.955	0.955
Faq	0.992	0.894	0.995	0.99
Php	0.863	0.939	0.965	0.955
news	0.978	0.987	0.995	1.000

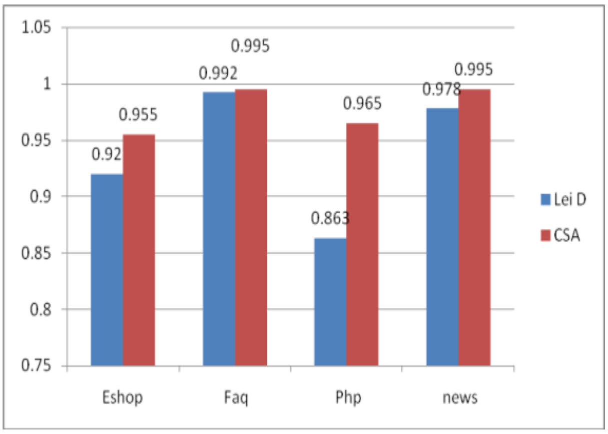

Figure-2: Comparison of CSA and Lei D. in terms of precision

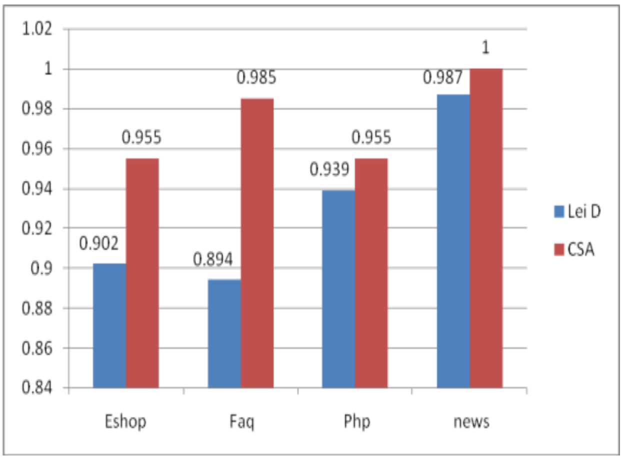

Figure-3: Comparison of CSA and Lei D. in terms of recall

5. CONCLUSION

In this paper, a Combined Stemming Approach (CSA) is proposed for improving the performance of web page genre classification. This approach implements stemming technique that combines porter stemming and lovin's stemming approaches to overcome the limitations that were present in both the techniques. Experiments were run on 7-genre corpus using Random forest classifier. The results obtained in this study show that the proposed approach (CSA) has improved the performance of web page genre classification in terms of both precision and recall, which is very important for increasing the quality of search results. The proposed approach is language dependent in nature and is applicable to English language only.

As the web is multilingual, there is a need to elaborate this work to a larger scale in which a language independent approach is required to extract features of such web pages.

6. REFERENCES

[1]. Xiaoguang QI, and Davison, B. D. Web page classification: Features and algorithms. *ACM Computer. Survey.vol* 41, 2, Article 12 February 2009.

[2]. Mehler, A., Sharoff, S., and Santini, M., Genres on the Web: Computational Models and Empirical Studies. Springer, Berlin/New York, 2009.

[3]. A. Finn and N. Kushmerick. Learning to Classify Documents According to Genre. *Journal of American Society for Information Science and Technology*, 2006.

[4]. Boese E.S. Stereotyping the web: genre classification of web documents, 2005, Citeseer.

[5]. Santini M. "Some issues in Automatic Genre Classification of Web Pages", 2006, Proc. of the Journées Internationales d'Analyse Statistique des Données Textuelles (JADT), Besançon France.

[6]. Lijuan J. and Liping, F Improvement of Feature Extraction in Web Page Classification, 2010 IEEE 2nd International Conference on e-Business and Information System Security (EBISS).

[7] M. Shepherd, C. Watters, and A. Kennedy. Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering*, 3(3&4):236-251,2004.

[8]. Lei D, Carolyn Watters, Jack Duffy, Michael Shepherd An Examination of Genre Attributes for Web Page Classification , 2008,IEEE Proceedings of the 41st Annual Hawaii International Conference on System Sciences.

[9]. M. Santini. Automatic Identification of Genre in Web Pages, PhD thesis, University of Brighton, 2007.

[10]. Santini M. "Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis". Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 2005). Manchester, (UK). 2005.

[11]. Santini M. Characterizing Genres of Web Pages: Genre Hybridism and Individualization Proceedings of the 40th Hawaii International Conference on System Sciences – 2007

[12]. Santini M. Zero, Single, or Multi? Genre of Web Pages Through the Users' Perspective. Information Processing and Management, 2008, pp. 702-737.

[13]. Santini M., Georg Rehm, Serge Sharoff and Alexander Mehler. "Automatic Genre Identification:Issues and Prospects". Journal for Language Technology and Computational Linguistics, JLCL ISSN 0175-1336 Volume 24, 2009.

[14]. Willett P. The Porter stemming algorithm: then and now. Program: electronic library and information systems, 40 (3). pp. 219-223, 2006.

[15]. Lovins, J.B. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11, 22–31 1968.

[16]. Lindemann, C. and Littig, L., Classification of web sites at super-genre level, 2011, Springer journal Genres on the Web pages pages 211—235.

[17]. Kennedy A. and Shepherd M. Automatic Identification of Home Pages on the Web IEEE,Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.

[18]. Ezeiza Ramos J. , Epelde Pagola I., Elordui Urkiza U., Payá Ruiz X. TOWARDS A volumen 6 año 2011.

[19]. http://nlp.stanford.edu/software/tagger.html.

[20]. Santini M. and Sharoff S. "*Web Genre Benchmark Under Construction*". Journal for Language Technology and Computational Linguistics (JLCL) 2009, volume 25, number 1 -- Special Issue: Automatic Genre Identification: Issues, and Prospects".