Data Min Knowl Disc
DOI 10.1007/s10618-015-0428-8

Exploiting link structure for web page genre identification

Jia Zhu[1] · Qing Xie[2] · Shoou-I Yu[3] · Wai Hung Wong[4]

Received: 1 December 2013 / Accepted: 20 June 2015

Abstract As the World Wide Web develops at an unprecedented pace, identifying web page genre has recently attracted increasing attention because of its importance in web search. A common approach for identifying genre is to use textual features that can be extracted directly from a web page, that is, On-Page features. The extracted features are subsequently inputted into a machine learning algorithm that will perform classification. However, these approaches may be ineffective when the web page contains limited textual information (e.g., the page is full of images). In this study, we address genre identification of web pages under the aforementioned situation. We propose a framework that uses On-Page features while simultaneously considering information in neighboring pages, that is, the pages that are connected to the original page by backward and forward links. We first introduce a graph-based model called GenreSim, which selects an appropriate set of neighboring pages. We then construct a multiple classifier combination module that utilizes information from the selected neighboring pages and On-Page features to improve performance in genre identification. Experiments are conducted on well-known corpora, and favorable results indicate that our proposed framework is effective, particularly in identifying web pages with limited textual information.

Responsible editor: Thomas Seidl.

✉ Jia Zhu
jia@intelligentforecast.com; jzhu@m.scnu.edu.cn

[1] School of Computer Science, South China Normal University, Guangzhou, China

[2] Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

[4] School of Decision Sciences, Hang Seng Management College, Hong Kong, China

Published online: 07 July 2015

Keywords Genre identification · Neighboring pages selection · Multiple classifiers

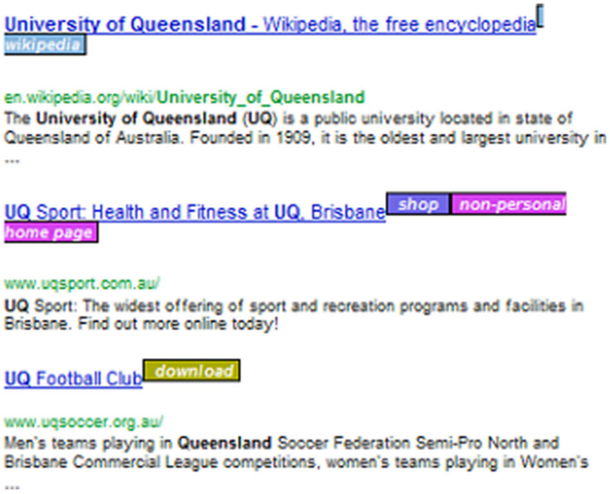
1 Introduction

As the World Wide Web develops at an unprecedented pace, the importance of categorizing web pages to improve web search has increased substantially. Web page categorization can be divided into two main sub-problems: topic categorization and genre categorization. Our work will address the problems of genre categorization (also known as genre identification). Unlike topic categorization that identifies information in the detailed contents of web pages (Stein and zu Eissen 2006), genre categorization focuses on the functional purposes of web pages and discovers groups of texts that share a general transmission form, purpose, or discourse properties. As such, genre can be defined as a term for the category or form of a set of web pages. For example, a website such as www.news.com.au has many pages about different topics, including sport news and finance news; however the genre of the pages in this website is *news*. Genre identification has recently attracted increasing attention because it can be used to significantly improve the quality of web and document search results (Abramson and Aha 2012; Bjroneborn 2011; Pritsos and Stamatatos 2013; Zu Eissen and Stein 2004). Through genre identification, a search engine can label the genre of each retrieved result, which allows users to save time, as demonstrated in Fig. 1 using WEGA Firefox plugin.[1]

The majority of existing approaches to genre identification are limited to using information on a page, namely, On-Page features. On-Page features include textual information, such as bag-of-words (BOW) and term frequency (Kessler et al. 1997; Stamatatos et al. 2000), and structural information Boese and Howe 2005; Zu Eissen and Stein 2004), such as HTML tags. Textual information represents web page content whereas structural information specifies how content is presented. In a particular web page, however, these features are sometimes missing or unrecognizable because of various reasons. For example, some web pages contain a large number of images or videos and few textual content as shown in Fig. 2. In this case, a classifier that depends solely on On-Page textual features will find it difficult to correctly identify the genre of a web page.

Considering the aforementioned limitations, we propose a framework that uses additional information from the neighboring pages of a given web page to improve genre identification. Neighboring pages are web pages that are connected to a given web page by either forward or backward links. The main motivation for using additional information from such pages is that their genre is typically similar to that of the original page, and thus they contain discriminative textual information to recognize the genre of the given web page. Therefore, using linkage information to harvest additional textual features can provide additional references for classifiers to learn an improved classification model (Chen and Choi 2008; Mehler et al. 2007; Qi and Davison 2008). One problem in using information from neighboring pages is that a given page may have numerous neighboring pages that likely belong to different web gen-

[1] http://www.uni-weimar.de/en/media/chairs/webis/research/projects/wega/.



University of Queensland - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/University_of_Queensland
The **University of Queensland** (UQ) is a public university located in state of Queensland of Australia. Founded in 1909, it is the oldest and largest university in ...

UQ Sport: Health and Fitness at UQ, Brisbane

www.uqsport.com.au/
UQ Sport: The widest offering of sport and recreation programs and facilities in Brisbane. Find out more online today!

UQ Football Club

www.uqsoccer.org.au/
Men's teams playing in **Queensland** Soccer Federation Semi-Pro North and Brisbane Commercial League competitions, women's teams playing in Women's ...

Fig. 1 Sample of labeled search results

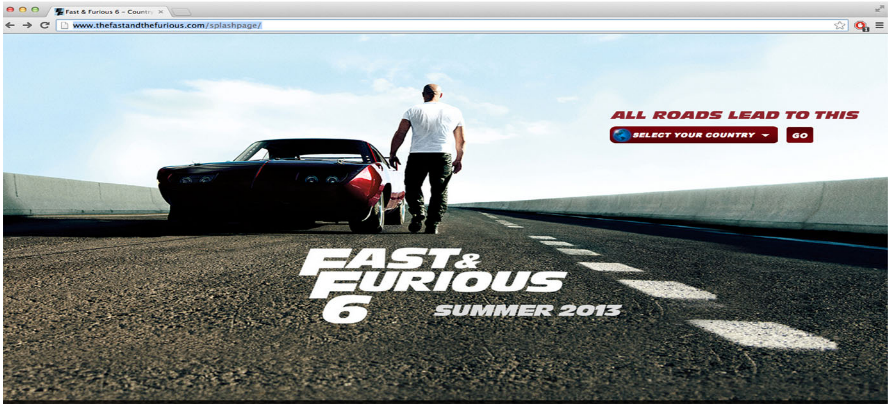


Fig. 2 Web page with few textual information

res. However, we are only interested in neighboring pages that are in the same genre as the original page. Therefore, we propose a link-based graph model called GenreSim, which exploits a link structure to select relevant neighboring pages. Compared with existing methods, our model considers as many neighboring pages as possible while greatly reducing noisy information by selecting only an appropriate subset of relevant neighboring pages. To further improve classification performance, we then construct a multiple classifier combination (MCC) module to combine outputs from various classifiers that use different features. The proposed framework is evaluated using benchmark corpora for web page genre identification, and the favorable results indicate the effectiveness of our method.

In summary, our main contributions are threefold compared to a previous work (Zhu et al. 2011):

1. We propose a link-based graph model called GenreSim, which exploits link structure to select relevant neighboring pages. The selected pages are used to improve genre identification of the original web page. We improve the selection process by considering the probability that a page has a given number of backward links.
2. We propose an MCC model that combines outputs of multiple classifiers to improve genre identification.
3. Extensive experiments with statistical tests on standard benchmark corpora demonstrate the effectiveness of our method.

The rest of this paper is organized as follows. In Sect. 2, we discuss related works in web page genre identification. In Sect. 3, we describe the main challenges in genre identification and the details of our framework. In Sect. 4, we present our experiments, evaluation metrics, and the results of the experiments. We conclude the study in Sect. 5.

2 Related works

Previous studies on web page genre identification differ significantly with respect to two factors, namely: (1) the feature set used to represent the content and structure of web pages and (2) the method used to classify web pages based on the chosen feature set. The following review of previous important works is presented in chronological order.

Zu Eissen and Stein (2004) provided a corpus of eight genres following a user study on genre usefulness and built a genre classifier using discriminant analysis. They also examined various feature sets in an attempt to combine different kind of information, including word frequencies, text statistics, and parts of speech frequencies. All these information are textual information that are also used in our framework.

Kennedy and Shepherd (2005) focused on a specific genre and its sub-genres. Using a neural network, they attempted to discriminate between home pages and non-home pages. Their feature set comprises features on content (e.g., common words and meta tags), form (e.g., number of images), and functionality (e.g., number of links and use of scripts). The best reported results are for personal home pages.

Finn and Kushmerick (2006) investigated different feature sets for genre classification, including BOW, parts of speech frequencies, and text statistics. They studied how these features can be combined to maximize accuracy both for a single topic and across topics. Their model was generated by C4.5, and they evaluated how their classifier can perform well either in a single domain or in a domain transfer with ensemble and active learning.

Santini (2006) studied web page genre classification based on three different feature sets, including frequencies of common words, parts-of-speech trigrams, HTML tags, punctuation marks, and so on. She built a corpus and evaluated her approach to predict the genre of unclassified web pages using SVM and naive Bayes classifiers. Afterward, the author further analyzed web page genre classification in terms of genre hybridism and individualization (Santini 2007). Genre hybridism accounts for genre variation within web pages, whereas individualization refers to the absence of any recognized genre in a web page. The author claimed that web pages require a zero-to-multi-genre classification scheme aside from the traditional single-genre classification. We applied this concept in our approach, but select the genre with the highest score for a page rather than mark a page with multiple genres.

Kanaris and Stamatatos (2007) proposed low-level feature sets of variable-length character n-grams and combined this representation with information on the most frequent HTML tags. Based on two benchmark corpora, they demonstrated that an SVM-based classification approach can improve the results in both cases compared with other machine learning approaches. Based on their research outcome, we selected SVM as the classification method for our proposed framework.

Dong et al. (2008) examined the effects of various attributes on four web genres, namely, FAQ, news, e-shopping, and personal home pages, by using a naive Bayes classifier and the information gain measure for feature selection. The results indicated that few features produce high precision, but many features produce high recall. In addition, combined attributes will always perform better than single attributes.

Chen and Choi (2008) defined five top-level genre categories and developed new methods to extract 31 features from web pages, which analyzed both contents and other features, such as HTML tags and Java scripts. Their evaluation results showed that additional features can help a classifier improves its learning of the classification model.

Jebari (2009) proposed a new centroid-based approach for web page genre categorization using a set of genre-labeled web pages. The obtained centroids from these pages are used to classify new web pages. Each web page is assigned to all predefined genres with confidence scores, and the centroids are refined after classification. The feature sets include URL addresses, logical structures, and hypertext structures. The experiments conducted on two known corpora showed that this approach is fast and outperforms other machine learning approaches that require data training.

Mason et al. (2010) extended previous works that use n-gram representation of a web page to automatically classify web pages by genre given that these n-gram based methods have high precision and low recall rates. Their experiments showed that their approach can assign more labels than previous works while maintaining high accuracy. In line with most existing research, we only focus on the accuracy of single-label classification in this study, which indicates that each web page only belongs to one genre.

Sharoff et al. (2010) investigated the performance of several types of features and found that one of the n-gram features performs best in their experiments. However, its performance may not be transferable to a wider web because of the lack of comparability among different annotation labels. Although their experiments are comprehensive, their approach cannot be applied when the amount of text in a web page is limited.

Kim and Ross (2011) reported that term distribution pattern is a better indicator for determining genre class than term frequency. They worked with pdf documents rather than web pages, but some of their research outcomes, such as using the derived document structural information from a scientific article and locating target information to accurately extract text information, can also be used as references in web page genre identification.

Kumari and Reddy (2012) recently proposed an approach called combined stemming. They focused on extracting genre-relevant words based on word level and linguistic features to improve classification accuracy using the techniques of combined stemming and stop-word elimination. Their experimental results showed the superior performance of their proposed method on the test data set.

Pritsos and Stamatatos (2013) focused on using content information with different text representation methods. They also examined SVM-based learners and an ensemble of classifiers. The results demonstrated that their approach can achieve high precision while maintaining relatively high recall. However, their approach still cannot work well when the information in web pages is limited.

All of the aforementioned approaches focus only on On-Page information, which may be ineffective when textual information on the pages is sparse. Abramson and Aha (2012) presented a method that uses information from URLs for web page genre classification because some URLs may contain text that indicates the genre, such as a blog. This approach can partially solve the problem, but it is still not a general approach for all websites. For example, Facebook and Twitter are both social networking websites, yet determining if they should belong to the same genre based on their URLs is difficult.

Compared with the aforementioned works discussed above, our proposed framework does not only use textual and structural information, but also information from selected relevant neighboring pages. We improve classification accuracy by using combinations of multiple classifiers. The details of this framework are introduced in the next section.

3 Proposed framework

Our proposed framework identifies web page genre based on both On-Page features and features from relevant neighboring pages. Features are selected using the feature selection module and are used to train multiple classifiers. During testing phase, the prediction scores acquired from multiple classifiers are inputted into the MCC module to obtain a single prediction score. The proposed framework is illustrated in Fig. 3.

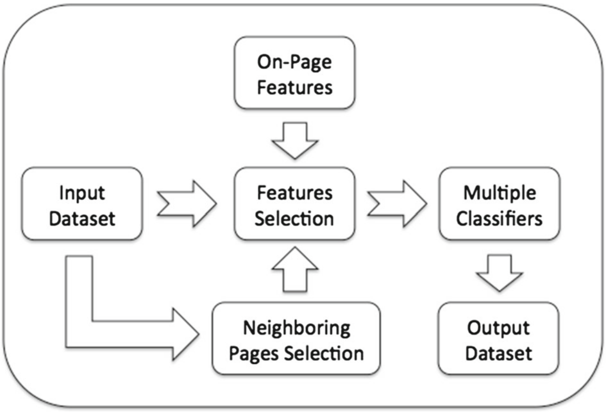


Fig. 3 Proposed framework

3.1 On-Page features

On-Page features capture textual and structural information in a given web page. Structural features mainly indicate the layout of a web page, whereas textual features represent the content of a web page. The textual information used in our method are page URL, title, keywords, headings, and anchors. Similar to many of the existing works discussed in Sect. 2, we do not use text in between ⟨body⟩ tags because the types of information we require are straightforward and rich resources for a classifier to perform genre identification compared with the text in the body. For example, if the URL contains specific words such as "faq,""cv," and "how," then the web genre can be easily recognized because these words provide clear information. Moreover, for those web pages with limited textual information, these words are the most easily accessible sources; thus, we select them as textual features. We apply Term Frequency-Inverse Document Frequency (TF-IDF) to represent all the words we extracted as vectors. In our approach, a term is a synonym of a "word". TF-IDF is one of the most popular statistics used in text mining. The calculated term weight and inverse document frequency are used to score and rank the relevance of a document (Salton and McGill 1986). The TF-IDF score for each word is calculated as follows:

$$TermScore_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \times log \frac{|D|}{|\{d : t_i \in d\}|}, \tag{1}$$

where n_{ij} is the number of occurrences of the word t_i in document d_j, the denominator is the sum of the occurrences of all words in document d_j, $|D|$ is the total number of documents in the corpus, and $|\{d : t_i \in d\}|$ is the number of documents where the word t_i appears. In our case, each document is a web page. We eventually obtain a vector that is a list of TF-IDF-represented words that we have extracted from each web page.

Structural features are utilized by our method by computing the frequency of different HTML tags. Following Dong et al. (2008), Lee and Myaeng (2002), Zu Eissen and Stein (2004), the HTML tags for which frequency is computed for images, links, emails, forms, tables, and div tags. However, according to existing works and our preliminary evaluations, structural information that is used independently cannot obtain sound results, particularly if the data set contains various genres of web pages. Therefore, we combine structural information with textual information to form On-Page features. The structural features we used are listed in Table 1.

3.2 Neighboring pages selection module

On-Page features are insufficient for web page genre identification if the page has limited text information. Therefore, we propose to retrieve additional information from neighboring pages. Such additional information can provide further evidence for classifiers to perform reliable predictions. Some neighboring pages likely do not belong to the same genre as the original page. Therefore, selecting only relevant neighboring pages from a large pool of neighboring pages is crucial. In the next paragraphs, we

Table 1 Structural features

Feature name
Number of images
Number of links
Number of emails
Number of forms
Number of tables
Number of DivTags

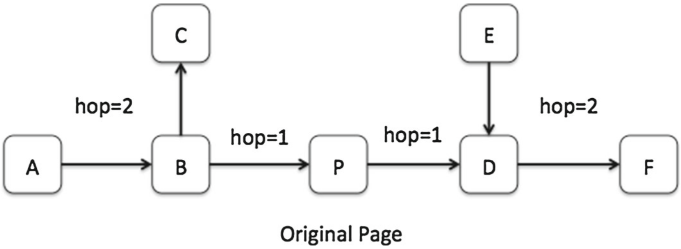


Fig. 4 Neighboring pages of Page P

detail how a set of neighboring pages is created for each web page, and then we will describe how our link-based similarity measure (GenreSim) selects appropriate neighboring pages.

Neighboring pages are web pages connected to the original web page by forward or backward links (Qi and Davison 2008). In particular, level-1 neighboring pages are the web pages that are immediately connected to a given web page by either forward or backward links. Level-2 neighboring pages are web pages that are two hyperlinks away from a given web page. For example, in Fig. 4, P is the original page. B, D, and A, C, E, F are its level-1 and level-2 neighboring pages, respectively. The radius we considered in our model is up to two hops because it is the farthest hop from the backward and forward links of level-1 neighboring pages to the original page. Considering additional pages, for example, three hops, will not only significantly increase the amount of computation, but will also receive more noisy information that will decrease classification performance according to our past observations and experiments. For example, assume that a page has an average of 10 neighboring pages and 50 % of these pages are not of the same genre as the original page. This difference indicates that we need to filter out $10 \times 10 \times 10 \times 0.5 = 500$ useless pages in the case if we consider a case with three hops.

We propose a link-based similarity measure called GenreSim, which is derived from the PageSim similarity measure (Lin et al. 2006), to select relevant neighboring web pages. Compared with other popular link-based similarity measures such as SimRank (Jeh and Widom 2002) and Co-citation (Salton and McGill 1986), PageSim can measure the similarity between any two pages without requiring intermediate pages unlike in SimRank. PageSim can also consider neighboring pages in multiple hops in contrast to Co-citation, which only considers immediate neighboring pages.

We model all the neighboring pages and the original page as a directed graph $G = (V, E)$, with vertices V representing web pages $v_i (i = 0, 1, 2, \ldots, n)$ and directed edges E representing hyperlinks among the web pages. Page v_0 is the original page. Below is a list of definitions for our model.

Definition 1 Path: $p(u, v)$ is a path that denotes a sequence of vertices from vertex u to vertex v through a set of edges $e_i (i = 1, 2, \ldots, m)$, $e_i \in E$, and $E \in G$.

Definition 2 Path set: $PATH(u, v)$ is the set of all possible paths from vertex u to vertex v.

Definition 3 Backward and forward links: $B(v)$ is the set of vertices that represent neighboring pages with a backward link to vertex v, and $F(v)$ is the set of vertices that represent neighboring pages with a forward link to vertex v.

Definition 4 Page score: $Score(v)$ is the query-independent score of page v, which represents the importance of v in the web graph according to its backward and forward links.

The original PageSim similarity measure only uses forward links, whereas in GenreSim, we use both forward and backward links because the number of backward links of a page can provide a general estimate of its prominence on the web graph (Kleinberg 1999). In addition, we implement a weighted method to compute the page score and make it suitable for genre identification. A hyperlink from page u to v can be considered as the recommendation of page v by u, and thus, u and v should have some kind of similarity (Arasu et al. 2001). Recommendation naturally decreases along with the links. We then define the recommendation score $Score(u, v)$ of page u that propagates to v through $PATH(u, v)$ as shown in Eq. (2):

$$Score(u, v) = \begin{cases} \Sigma_{p \in PATH(u,v)} \frac{d \cdot Score(u)}{\prod_{x \in p, x \neq v}(|F(x)|+|B(x)|)}, & v \neq u \\ Score(u), & v = u \end{cases} \tag{2}$$

where $d \in (0, 1]$ is a constant decay factor that makes the score decrease along with the path. Given that we do not concentrate on optimizing d in this work, we simply set the value of the decay factor to 0.5, which is a common setting used by many graph-based approaches. This setting indicates that the score is decreased by half for each hop in the path. $Score(u)$ is the page score for page u, and $u, v \in V$.

To compute $Score(u)$ in Eq. (2), we extend the hypertext-induced topic selection (Kleinbery 1999) algorithm. In the previously defined web page graph G, web pages with many links that point to them are called *authorities*, whereas web pages with many outgoing links are called *hubs*. Authorities are improved by incoming edges from good hubs, and hubs are improved by outgoing edges from good authorities. Let $H(p)$ and $A(p)$ be the hub and authority score of a page p, respectively. These scores are defined such that the following equations are satisfied for all pages p:

$$\begin{cases} H(p) = \Sigma_{u \in V | p \rightarrow u} A(u) \\ A(p) = \Sigma_{v \in V | v \rightarrow p} H(v), \end{cases} \tag{3}$$

where $H(p)$ and $A(p)$ are normalized for all web pages and calculated by the number of links. In our approach, we use the sum of authority and hub scores as the score of a page, that is, $Score(p) = H(p) + A(p)$, because pages with high authority scores are expected to have a relevant content, whereas pages with high hub scores are expected to contain links to a relevant content. However, high-scoring pages with a few backward links but a large number of forward links, which are highly likely to be spam pages because they point to many irrelevant pages but only a few pages refer to themselves. We use the number of backward links to measure similarity because our objective is to find neighboring pages that are similar to the original page. According to Kleinberg (1999), two web pages are similar in terms of prominence if they have similar numbers of backward links, which indicates that similar numbers of web pages refer to them. Therefore, we modify the preceding equation as follows:

$$\begin{cases} H(p) = \Sigma_{u \in S | p \rightarrow u} \omega(p) \cdot A(u) \\ A(p) = \Sigma_{v \in S | v \rightarrow p} \omega(p) \cdot H(v), \end{cases} \tag{4}$$

where $\omega(p)$ is the weighting parameter calculated by the relevance value of a neighboring page to the original page. $\omega(p)$ is defined as follows:

$$\omega(p) = \frac{1}{|\log N - \log N(p)| + 1}, \tag{5}$$

where N is the number of backward links of the original page, and $N(p)$ is the number of backward links of neighboring page p. In this manner, if the number of backward links for page p is similar to that for the original page, then a higher weight will be assigned to the page.

However, this formula does not represent the probability for a page to have a given number of backward links. For example, if an original page has 100 backward links, then the difficulty of finding a neighboring page with a similar number of backward links is higher than that for an original page, which only has 5 backward links. Therefore, we extend Eq. (5) as follows:

$$\omega(p) = \frac{1}{|\log N - \log N(p)| + 1} \cdot N, \tag{6}$$

which indicates that if the original page has a higher number of backward links, then its neighboring pages should also have higher weights to better match the calculation for page score.

We then obtain the similarity score $Sim(u, v)$, which denotes the similarity between u and v and represents their importance compared with other pages as well as the similar recommendation they propagate to other pages in the graph. The graph is built from neighboring pages; thus, most of the pages are from the same domain. Consequently, u and v likely belong to the same web genre. For example, the wiki page of Michael Jordan has numerous backward and forward links, most of which are related to the NBA, whereas its neighboring page, namely, the wiki page of Kobe Bryant, has recommendations similar to other pages in the web graph according to the linking information. The similarity score is defined in the following equation by adopting the Jaccard measure (Jain and Dubes 1988), which is commonly used in information retrieval to measure similarity:

$$Sim(u, v) = \frac{\sum_{i=1}^{n} min(Score(v_i, u), Score(v_i, v))}{\sum_{i=1}^{n} max(Score(v_i, u), Score(v_i, v))}, \tag{7}$$

where $u, v \in V$.

We then select the top K neighboring pages with the highest similarity scores to extract information and construct features similar to the On-Page features we discussed in Sect. 3.1. We use these features from neighboring pages with On-Page features to identify the genre of a web page. The details are provided in the next section.

3.3 MCC module

To achieve good performance in genre identification, we train multiple classifiers using the features acquired from the previous section and combine the classifiers to obtain a single prediction score. Most genre identification algorithms are based on machine learning techniques. SVM (Mitchell 1997) is a powerful learning method based on structural risk maximization theory, which aims to minimize generalization error instead of relying on the empirical error from training data alone. We use three classifiers in our approach, namely, SVM Contents, SVM On-Page Features, and SVM Neighboring Pages. SVM Contents is based on the textual information on the page as described in Sect. 3.1. SVM On-Page Features adds structural information in the feature set. SVMs based on textual and structural information have been used by various researchers in genre identification, such as in Kanaris and Stamatatos (2007), Pritsos and Stamatatos (2013), Santini (2006), as discussed in Sect. 2. SVM Neighboring Pages is based on the textual and structural information in selected neighboring pages. We also design an MCC module to combine the three classifiers because the combination of homogeneous classifiers that use heterogeneous features can improve the final result (Orrite et al. 2008).

Assume that each classifier based on SVM produces a unique decision when we define each web page p as belonging only to one genre. We compare the results among all three classifiers, and the final output depends on the reliability of the decision confidences provided by the participating classifiers. We apply the concept of Decision Template (DT) to avoid the case in which the classifiers make independent errors (Kuncheva et al. 2001). DT uses the outputs from all classifiers to calculate the final support, which is also called the confidence score, based on a matrix for each class.

Assume that each classifier produces the output $E_i(p) = [d_{i1}(p), \ldots, d_{i|G|}(p)]$, where $d_{ij}(p)$ is the membership degree given by the classifier E_i in which a web page p belongs to the genre j, $j = 1, \ldots, |G|$. The outputs of all classifiers can be represented by a decision matrix DP as follows:

$$DP(p) = \begin{pmatrix} d_{11}(p) & \dots & d_{1|G|}(p) \\ d_{21}(p) & \dots & d_{2|G|}(p) \\ d_{31}(p) & \dots & d_{3|G|}(p) \\ & \vdots & \\ d_{N1}(p) & \dots & d_{N|G|}(p) \end{pmatrix}. \tag{8}$$

The membership degree $d_{ij}(p)$ is calculated using the training set T_j, $j \leq |G|$, where T_j is the training set for each genre and $|G|$ is the number of genres.

$$d_{ij}(p) = \frac{Ind(T_j, i)}{|G|}, \tag{9}$$

where $Ind(T_j, i)$ is an indicator function with a value of 1 only if the output from the classifier E_i based on T is the same as the output based on training set T_j and the classifier assigns the page to genre j. Otherwise, we assign 0 to the function. At this stage, we obtain the membership degree for each page p that belongs to a genre j and stored in a matrix $DP(p)$.

We then calculate the confidence score $ConfidenceScore_j(p)$ of page p using various rules from the $DP(p)$ for each genre j, and choose the genre with the highest confidence score. Assuming that N is the number of classifiers, we apply the minimum, maximum, and average rules for the preceding matrix to consider diversity among multiple classifiers as follows:

$$Minimum\,rule : ConfidencScore_j(p) = Min_{i=1}^{N}(d_{ij}(p)), \tag{10}$$

$$Maximum\,rule : ConfidenceScore_j(p) = Max_{i=1}^{N}(d_{ij}(p)), \tag{11}$$

$$Average\,rule : ConfidenceScore_j(p) = Mid_{i=1}^{N}(d_{ij}(p)). \tag{12}$$

4 Evaluations

We discuss the evaluation results in this section. To investigate our approach in depth, we evaluated the performance of individual classifiers and the MCC algorithm, and compared them against existing works on different corpora.

4.1 Corpora and data preparation

In our experiment, we used the updated version from two popular corpora, KI-04 and 7-Web collections (Santini 2006). We updated the web pages using a search engine to obtain the latest contents and their neighboring pages. These corpora are composed of English web pages. Each web page is associated with a specific source URL address and belongs to a single genre class. KI-04 corpus (Table 2) is composed of 1205 HTML web pages, which are divided into eight genres. 7-Web corpus (Table 2) contains 1400 HTML web pages from seven genres.

To evaluate if our approach can perform well in case web pages do not have much textual information, we also manually and randomly collected 2000 such pages

Table 2 KI-04 and 7-Web data sets

Genres (KI-04)	#	Genres (7-Web)	#
ARTICLE	127	BLOG	200
DOWNLOAD	151	E-SHOP	200
DISCUSSION	205	FAQs	200
PORTRAYAL-PRIVATE	126	ONLINE NEWSPAPER FRONTPAGE	200
PORTRAYAL-NON PRIVATE	163	PERSONAL HOME PAGE	200
LINK COLLECTION	127	LISTING	200
HELP	139	SEARCH PAGE	200
SHOP	167		

Table 3 IV-12 data set

Genres	#
Movie homepages	500
Photography websites	500
Video sharing websites	500
Music download websites	500

with four genres from the Internet, called the IV-12 data set as shown in Table 3. The four genres are movie homepages, e.g., Transformers 4,[2] photography websites, e.g., Dreamstime,[3] video sharing websites, e.g., YouTube,[4] and music websites, e.g., Music.com.[5]

We crawled neighboring pages for the web pages in the three corpora by using the "⟨a href⟩" tag for forward links and AHREFS,[6] which is an online tool to analyze websites, for backward links. The latest crawl was conducted in November 2012 with a total of 3,72,325 neighboring pages after removing broken links. We preprocessed all web pages, including neighboring pages, before using classifiers to train. We selected and tokenized text into words, as well as remove numbers, non-letter characters, common stop words,[7] and special characters. We then selected stem terms using the Lovins stemmer (Lovins 1968) and calculated the TF-IDF value of each remaining word for

[2] http://www.transformersmovie.com/.

[3] http://www.dreamstime.com/.

[4] http://www.youtube.com.

[5] http://www.music.com.

[6] http://ahrefs.com/.

[7] http://www.textfixer.com/resources/common-english-words.txt.

each page. All aforementioned steps were implemented using the Word Vector Tool[8] and htmlparser.[9]

4.2 Evaluation results

We evaluated our approach based on the metrics adopted in some previous works (Laender et al. 2008; Pereira 2009; Sebastiani 2002).

Pairwise F1 is defined as the harmonic mean of pairwise precision and pairwise recall. Pairwise precision is the number of true positives divided by the total number of elements labeled that are belonging to the positive class. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class. These variables are defined as $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, and $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$.

TP stands for true positive, which is the number of items correctly labeled as belonging to the positive class; FP stands for false positive, which is the number of items incorrectly labeled as belonging to the class; and FN stands for false negative, which is the number of items that are not labeled as belonging to the positive class but that actually do.

Given that our data is multi-class, evaluation is conducted in two binary classifications: one-class-against-the-rest and pairwise (Bernhard et al. 1998). We used macro-averaging (Sebastiani 2002) to measure the classification accuracy of both methods. Hence, precision and recall are first evaluated "locally" for each class, and then "globally" by averaging the results of different classes. Then, the macro-averaging of precision $Precision_M$ and recall $Recall_M$ is represented as $Precision_M = \frac{\sum_{i=1}^{|C|} Precision_i}{|C|}$, $Recall_M = \frac{\sum_{i=1}^{|C|} Recall_i}{|C|}$, where $|C|$ is the total number of classes.

4.2.1 Evaluations of neighboring pages selection

We first evaluated our neighboring pages selection model by applying different similarity measures in the KI-04, 7-Web, and IV-12 corpora. Given that different originating pages will have a varying number of linked pages, using a fixed number to neighboring pages for the evaluation appears to be a convincing approach. We selected the top K neighboring pages with the highest similarity scores generated by different similarity measures and manually checked the pages to determine selection accuracy. In our experiments, we evaluated two cases, $K = 10$ and $K = 20$. Selection accuracy is measured using $\frac{M}{K}$, where M is the number of neighboring pages with the same genre as the original page. GenreSim achieves better results than the other models in all three corpora, as shown in Fig. 5. In addition, it performs particularly well in the IV-12 corpus, which consists of web pages with insufficient textual information in both $K = 10$ and $K = 20$. Performance deteriorates for all similarity measures as K increases, but the margin of deterioration in GenreSim is smaller than those of

[8] http://ostatic.com/wvtool.

[9] http://htmlparser.sourceforge.net/.

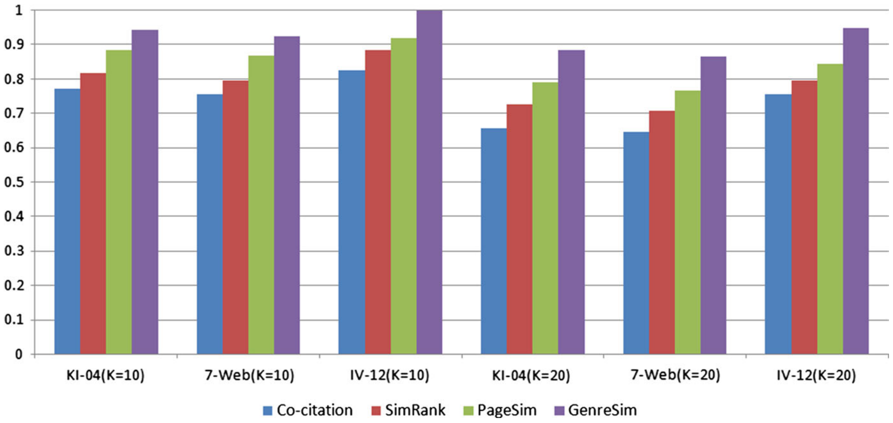


Fig. 5 Performance of different similarity measures for neighboring pages selection

the others. Given that performance is better when $K = 10$ than when $K = 20$, we decided to use only the top 10 neighboring pages to identify the genre of web pages.

4.2.2 Comparisons of individual classifiers

This section presents the comparisons between individual classifiers. As discussed earlier, we implemented SVM Contents and SVM On-Page Features classifiers, which represent existing works, to compare with our approach based on neighboring pages, namely, the SVM Neighboring Pages classifier. In addition, to evaluate the performance of our model in selecting the most relevant neighboring pages for genre identification, we also generated another classifier with neighboring pages by using the Google Similar Pages function[10] to return the first 10 neighboring pages. All classifiers were implemented by Weka API,[11] and we used RBFKernel Vapnik (1995) for SVM with default Weka' settings. In the learning phase, we adopt 10-fold cross validation to split the data sets for classifier training and testing.

Figure 6 shows the macro-averaging of four classifiers for the pages in the test data sets KI-04, 7-Web, and IV-12 for pairwise classification, whereas Fig. 7 shows the macro-averaging of four classifiers for the pages in the test data sets KI-04, 7-Web, and IV-12 for one-class-against-the-rest classification. As shown in the results in both tables, no significant difference is observed between the SVM Contents classifier and the SVM On-Page Features classifier in the KI-04 and 7-Web data sets because the dimension of textual information feature is significantly higher than that of structural information feature and dominated by textual feature in the On-Page features classifier. The classifier based on the feature of neighboring pages generated from our algorithm outperforms other classifiers, except in the 7-Web data set in the case of pairwise classification, in which the SVM Contents classifier yields better precision result. This finding is attributed to the 7-Web data set consists of web pages with

[10] http://www.google.com/help/features.html.

[11] http://www.cs.waikato.ac.nz/ml/weka/.

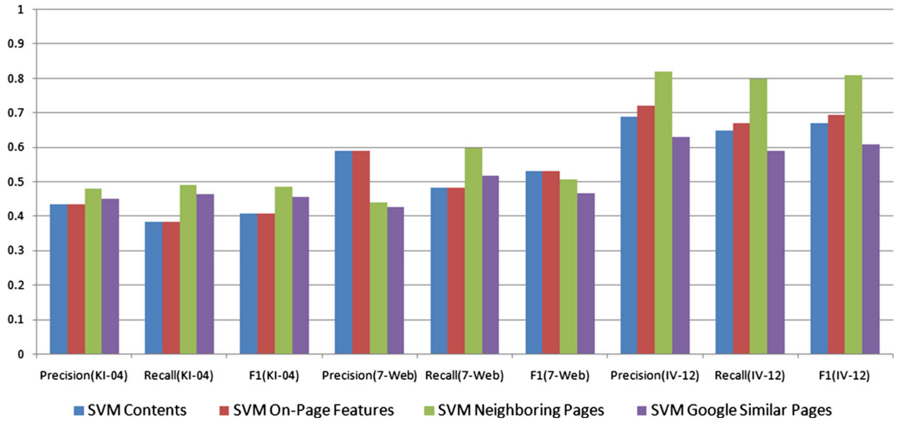


Fig. 6 Macro-averaging of individual classifiers in pairwise classification

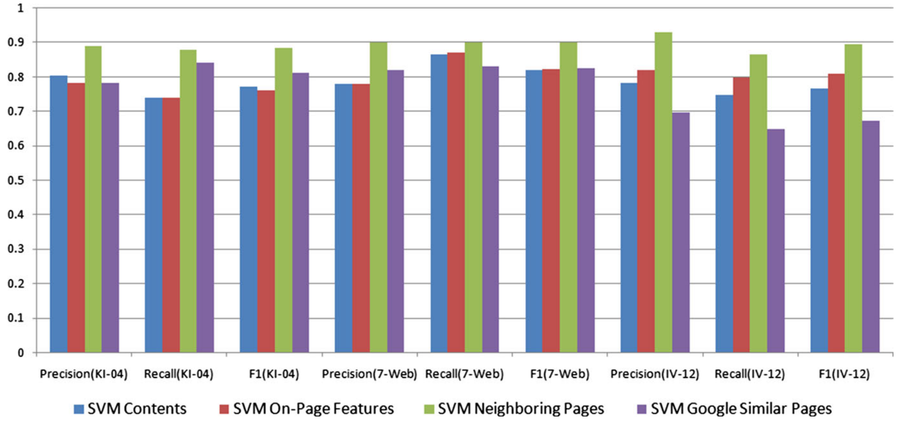


Fig. 7 Macro-averaging of individual classifiers in one-class-against-the-rest classification

rich textual/structural information but with relatively insufficient or irrelevant textual/structural information in neighboring pages. In the case of IV-12, for the data set contains web pages without much textual information, SVM Contents and SVM On-Page Features achieve better results than the other two data sets. However, the three corpora are different from one another and should be treated independently because they have varying sizes and are different types of web page. The key finding in this test case is that the SVM Neighboring Pages classifier performs better than the other classifiers, exhibiting approximately 15 % improvement from the traditional On-Page method. Its superiority is beyond the cases in the other two corpora. This finding validates our conclusion that information in neighboring pages can significantly improve genre identification performance when textual information is limited and insufficient. In addition, the classifier based on neighboring pages generated by our model outperforms the one generated by the Google Similar Pages function in all cases, which also supports the favorable results of our neighboring pages selection approach.

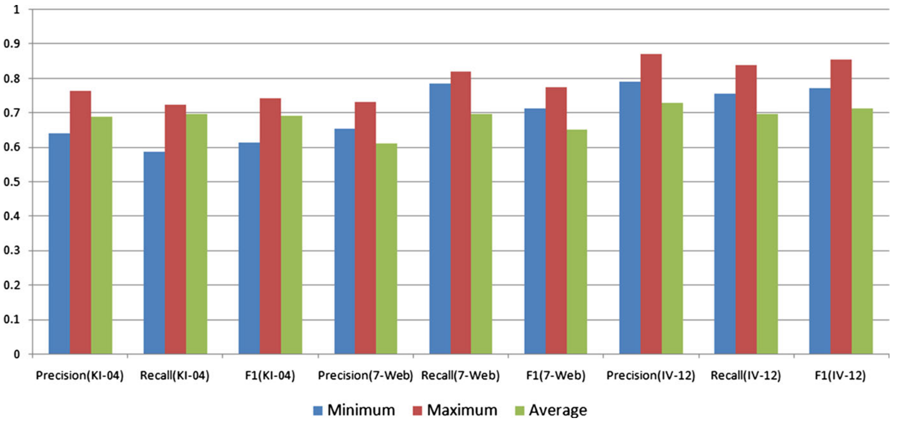


Fig. 8 Macro-averaging of MCC in pairwise classification

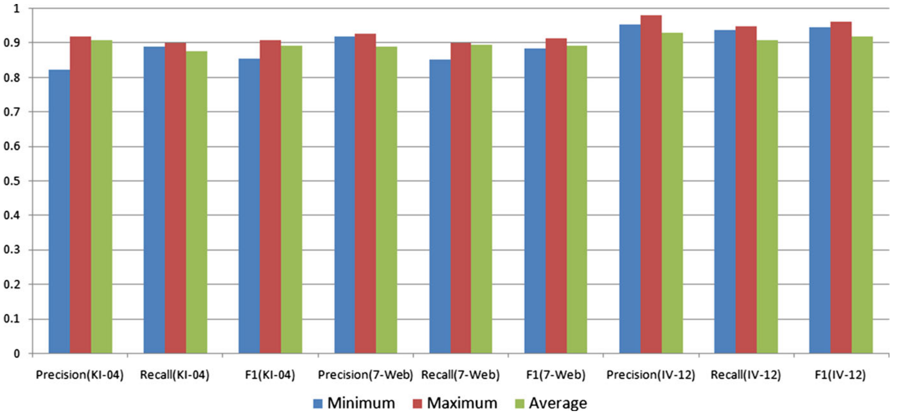


Fig. 9 Macro-averaging of MCC in one-class-against-the-rest classification

4.2.3 Performance of MCC

We evaluated MCC in both pairwise classification and one-class-against-the-rest classification. Figures 8 and 9 show the macro-averaging results of MCC by applying different rules to calculate the confidence score as described in Sect. 3. As shown in the tables, the maximum rule produces better results than the other two rules in all three data corpora. Based on the results, we discover that the identification results can be evidently increased by integrating all accessible information. In particular, for the 7-Web data set, the result of MCC outperforms those by SVM Contents and SVM On-Page Features, which indicates that the information extracted from neighboring pages can effectively enrich knowledge on target web pages and improve the final identification results.

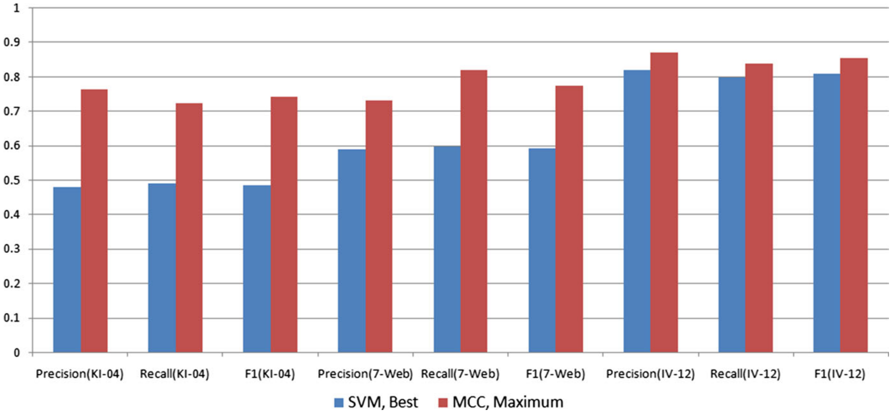


Fig. 10 MCC vs. individual classifiers in pairwise classification

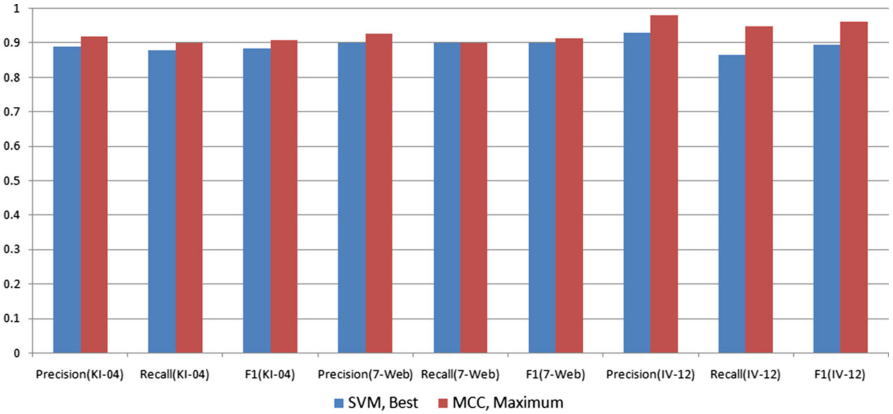


Fig. 11 MCC vs. individual classifiers in one-class-against-the-rest classification

4.2.4 MCC vs. individual classifiers

We also compared the results between MCC and individual classifiers, as shown in Figs. 10 and 11. We selected the results of the maximum rule in MCC and the best results from the three features produced by SVM for comparison. The results show that MCC improves by approximately 20 % in KI-04 and 7-Web in pairwise classification and outperforms SVM in one-class-against-the-rest classification at an average of 4 %, which indicates that MCC can improve considerably in the case of multi-class classification.

4.2.5 MCC vs. other ensemble methods

To evaluate the performance of MCC, in addition to using the common majority voting (MV) Lam and Suen (1996) algorithm as the baseline method for comparison, we also implemented the random feature subspacing ensemble (RFSE) algorithm based on the

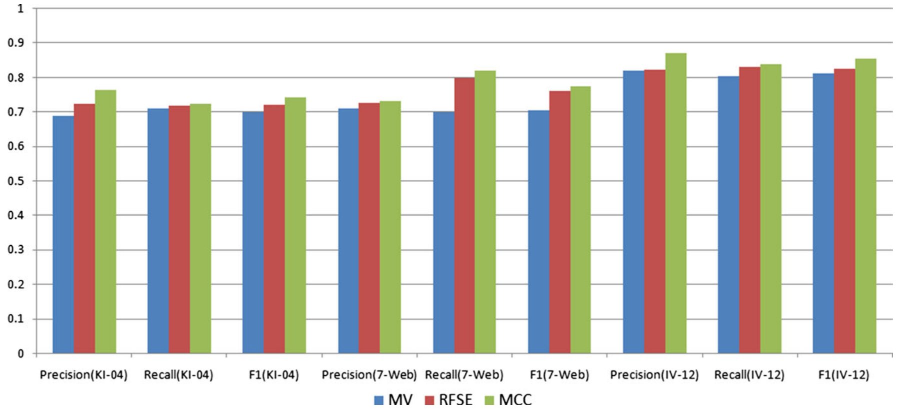


Fig. 12 MCC vs. other ensemble methods in pairwise classification

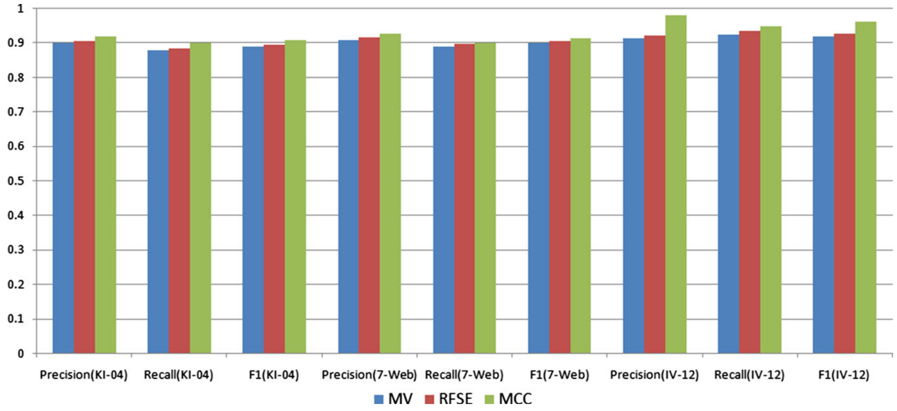


Fig. 13 MCC vs. other ensemble methods in one-class-against-the-rest classification

latest work (Pritsos and Stamatatos 2013) without considering the case of an unknown genre to suit our objective. We used the same features in all three methods. With respect to the parameters of RFSE, we applied the standard setting to the values of $k1$ and $k2$, which are 10 and 100, respectively. Figs. 12 and 13 show that all three methods achieve better results than the individual classifiers. The performance of MCC is only slightly better than those of MV and RFSE except in the IV-12 data set, which shows an average improvement with approximately 3 %. This result indicates that our proposed method is suitable for web pages without much textual and structural information.

4.2.6 Significance tests

We also designed two significance tests to compare the performance of two systems. The first is the *s* test designed for micro-level analysis (Yang and Liu 1999) and the second is the standard t test[12] designed for macro-level analysis. We used the following notations for the s test:

- N is the number of decisions by each system,
- n is the number of times that the decision by two systems are different, and
- k is the number of times that the decision by one system is correct and the decision of the other system is incorrect.

The P value (1-sided) is computed using the following binomial distribution:

$$P(Z) = \begin{cases} \sum_{i=k}^{n} \binom{n}{i} \times 0.5^n, & Z \geq k \\ \sum_{i=0}^{k} \binom{n}{i} \times 0.5^n, & Z < k, \end{cases} \tag{13}$$

where $Z = \frac{k-0.5n}{0.5\sqrt{n}}$.

Tables 4, 5, and 6 summarize the results of the statistical significance tests on the $F1$ score on the three corpora, where "$>$" or "$\gg$" indicates a better classifier. We only performed significance tests on pairwise classification because its $F1$ value as is in apparent as in the case of the one-class-against-the-rest classification among classifiers. Not surprisingly, MCC dominates both micro and macro level significance tests. It is better than the other two ensemble methods and the individual classifiers.

With regard to the individual classifiers, we cannot say that the SVM Neighboring Pages classifier is significantly better than the others in the KI-04 and 7-Web corpora because their outcomes in the s test and t test disagree. However, based on the results in Table 6, we found that the SVM Neighboring Pages classifier outperforms all individual classifiers, which proves our method can improve web page genre identification performance when the textual information of web pages is limited.

4.2.7 Comparisons using full text

To make our work more comprehensive, we also compared the SVM Neighboring Pages classifier, which is our main contribution to one of the representative existing works (Sharoff et al. 2010) using full text information of web pages. We used 4-grams in this experiment, which is the best feature found by the authors. The details are shown in Figs. 14 and 15. From the results, we can see that the performance of both methods are in the same level in the KI-04 and 7-Web corpora but SVM Neighboring Pages is better in IV-12. For the KI-04 and 7-Web corpora, we noticed that the original web pages already contain sufficient textual feature; thus, the SVM 4-grams algorithm can work well. However, our SVM Neighboring Pages classifier can still achieve similar performance by applying information from neighboring pages. This result can evidently support the effect of the selected neighboring pages. Furthermore, for the IV-12 corpus that consists of web pages with limited textual information, our approach

[12] http://en.wikipedia.org/wiki/T-test.

Table 4 Statistical significance test results on KI-04

sysA	sysB	*s* test	*t* test
MCC	RFSE	$>$	$>$
MCC	MV	$>$	$>$
MCC	SVM Neighboring Pages	$\gg$	$\gg$
SVM Neighboring Pages	SVM On-Page Features	$>$	$>$
SVM Neighboring Pages	SVM Contents	$>$	$>$
SVM Neighboring Pages	SVM Google Similar Pages	$\sim$	$>$
SVM On-Page Features	SVM Contents	$\sim$	$\sim$
SVM Contents	SVM Google Similar Pages	$<$	$<$

$\gg$ or $\ll$ means P value ≤ 0.01,
$>$ or $<$ means $0.01 < P$ value < 0.05,
$\sim$ means P value > 0.05

Table 5 Statistical significance test results on 7-Web

sysA	sysB	*s* test	*t* test
MCC	RFSE	$>$	$>$
MCC	MV	$>$	$>$
MCC	SVM Neighboring Pages	$\gg$	$\gg$
SVM Neighboring Pages	SVM On-Page Features	$\sim$	$<$
SVM Neighboring Pages	SVM Contents	$\sim$	$<$
SVM Neighboring Pages	SVM Google Similar Pages	$\sim$	$>$
SVM On-Page Features	SVM Contents	$\sim$	$\sim$
SVM Contents	SVM Google Similar Pages	$>$	$>$

$\gg$ or $\ll$ means P value ≤ 0.01,
$>$ or $<$ means $0.01 < P$ value < 0.05,
$\sim$ means P value > 0.05

Table 6 Statistical significance test results on IV-12

sysA	sysB	*s* test	*t* test
MCC	RFSE	$>$	$>$
MCC	MV	$>$	$>$
MCC	SVM Neighboring Pages	$>$	$>$
SVM Neighboring Pages	SVM On-Page Features	$>$	$>$
SVM Neighboring Pages	SVM Contents	$>$	$>$
SVM Neighboring Pages	SVM Google Similar Pages	$\gg$	$\gg$
SVM On-Page Features	SVM Contents	$\sim$	$\sim$
SVM Contents	SVM Google Similar Pages	$>$	$>$

$\gg$ or $\ll$ means P value ≤ 0.01,
$>$ or $<$ means $0.01 < P$ value < 0.05,
$\sim$ means P value > 0.05

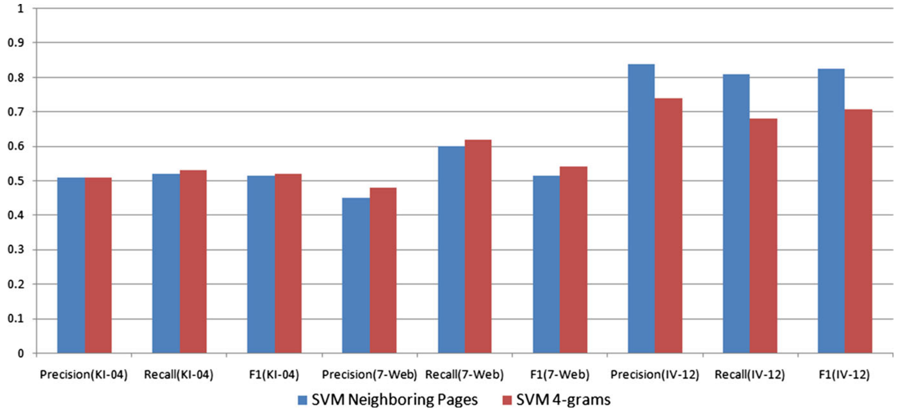


Fig. 14 SVM Neighboring pages vs. SVM 4-grams in pairwise classification

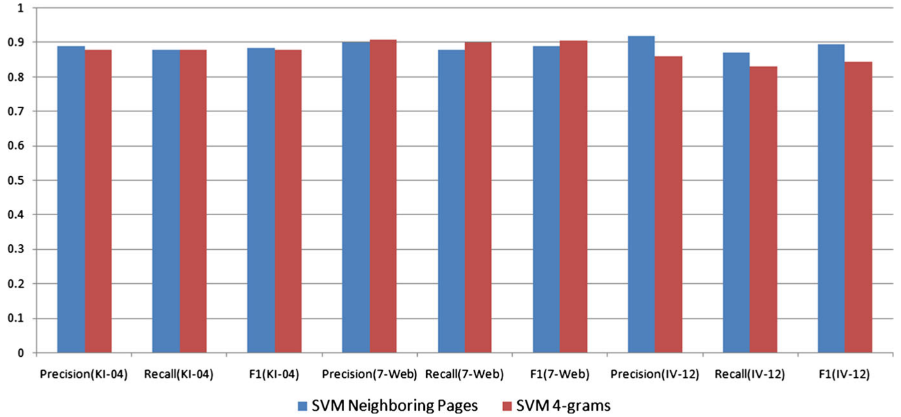


Fig. 15 SVM Neighboring pages vs. SVM 4-grams in one-class-against-the-rest classification

can outperform the method uses only textual features on the original pages, which is not surprising. This approach benefits from auxiliary information from neighboring pages.

4.2.8 Error analysis

Lastly, we analyzed the errors of the SVM Neighboring Pages classifier and the MCC method, and categorized these errors in Tables 7 and 8. The figure presented in the tables is the average percentage of errors from pairwise classification and one-class-against-the-rest classification. In this case, “error” is defined as the classifier or method that has not classified the web page in the right class. In Tables 7 and 8, “Only Classifier” or “Only Method” means that the errors are only made by the SVM Neighboring Pages classifier or MCC but not the others. “Less Information” means the average percentage of errors in the case in which neighboring pages have less information or their genres

Table 7 Error statistics of SVM Neighboring Pages classifier

Errors	KI-04 (%)	7-Web (%)	IV-12 (%)
Only classifier	8.5	19.4	2.9
Less information	62.6	78.5	64
All classifiers	43.5	54.4	62.1

Table 8 Error statistics of MCC

Errors	KI-04 (%)	7-Web (%)	IV-12 (%)
Only method	3.5	3.4	3.8
Less information	46.5	48.5	52
All methods	88.5	85.4	82.2

are different from the original page should be classified. “All Classifiers” or “All Methods” means that errors are made by all approaches.

From Table 7, we learned that the SVM Neighboring Pages classifier heavily relies on information from selected pages. For example, in the 7-Web data set, 19.4 % of errors made by the SVM Neighboring Pages classifier do not occur with other classifiers. In other words, if we use information from the original page, then these mistakes can be avoided. The main reason that the classifier made these mistakes is, the insufficient of useful information in selected neighboring pages. For example, we found that for 78.5 % errors in the 7-Web data set, the neighboring pages have less information than the original page. For example, no backward and forward links is found for the “PHP_025.htm”, and we obviously cannot get any information from neighboring pages. None of the classifiers can identify this kind of page because textual or structural information is limited in the page. Another error example is “PHP_028.htm”, that is, although we can obtain a few neighboring pages, most are linked pages that are irrelevant to the page. We also found a pattern in which errors frequently happened in a personal homepage, which is one of the genres in the 7-Web data set. According to our observations, personal homepages generally contain a few external links or their linked pages belong to different genres, which is a limitation in using information from neighboring pages.

According to the results in Table 8, given that MCC and the other two methods are combined by all classifiers, no obvious pattern is observed in all three corpora. In addition, the percentage of errors with less information is approximately 50 %, which indicates that the cause of errors does not depend on the amount of information we obtain.

5 Conclusions and future work

In this study, we addressed the challenge in the field of web page genre identification, which significantly benefits the improvement of web search. We focused on the situation wherein web pages only contain limited textual information, in which conventional approaches based on only On-Page features may not correctly identify the genre of these pages because of the insufficient available information. To address this problem, the idea of using neighboring web pages was proposed to potentially enrich useful feature information. For a web page to be identified, several of its neighboring pages are of the same genre as the original page. Therefore, we can select these pages and collect useful textual information from them, which can help identify the genre of the given web page. Such mechanism is particularly useful when the textual information of the given web page is insufficient. Based on these considerations, we designed our framework to select appropriate neighboring pages and effectively identify the genre of a given web page. To filter out the most appropriate neighboring pages, the GenreSim model was designed according to a graph-based similarity measure. To maximize the use of neighboring pages, we designed a MCC mechanism to integrate the information into neighboring pages with On-Page features. This framework successfully improved web page genre identification performance, particularly for cases in which the textual information is limited for the web pages to be identified. However, given that our solution heavily relies on information obtained from neighboring pages, if the neighboring pages of the given web page also contain minimal textual information, or if most of them are irrelevant to the original page, then our solution may not work as expected. Nevertheless, the proposed framework is an essential and valuable complement for the research on web page genre identification, and it provides an effective alternative solution to deal with extreme situations, such as insufficient textual information provided. The advantages of our proposed framework were proven and supported by extensive experimental evaluation on different well-known corpora with error analysis. In the future, we will focus on investigating dynamic weighting for individual classifiers to further improve the ensemble performance of multiple classifiers.

Acknowledgments This work was supported by the Youth Teacher Startup Fund of South China Normal University (No. 14KJ18) and the National High Technology Research and Development Program of China (863, No. 2013AA01A212).

References

Abramson M, Aha DW (2012) What's in a URL? genre classification from URLs. In: Workshops at the 26th Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence, pp. 1–8

Arasu A, Cho J, Garcia-Molina H, Paepcke A, Raghavan S (2001) Searching the web. ACM Trans Internet Technol 1(1):2–43

Bernhard S, Burges JC, Smola AJ (1998) Advances in kernel methods: support vector learning. The MIT Press, Cambridge

Bjroneborn L (2011) Genre connectivity and genre drift in a web of genres. In: Genres on the Web: Computational Models and Empirical Studies, pp. 255–274

Boese E, Howe A (2005) Effects of web document evolution on genre classification. In: Proceedings of the ACM 14th Conference on Information and Knowledge Management, pp. 632–639

Chen G, Choi B (2008) Web page genre classification. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 2353–2357

Dong L, Watters C, Duffy J, Shepherd M (2008) An examination of genre attributes for web page classification. In: Proceedings of the 41th Annual Hawaii International Conference on System Sciences, pp. 129–138

Finn A, Kushmerick N (2006) Learning to classify documents according to genre. J Am Soc Inf Sci Technol 57(11):257–262

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs

Jebari C (2009) A new centroid-based approach for genre categorization of web pages. J Lang Technol Comput Linguist 24(1):73–96

Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543

Kanaris I, Stamatatos E (2007) Web page genre identification using variable-length character n-grams. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol 7(1), pp. 3–10

Kennedy A, Shepherd M (2005) Automatic identification of home pages on the web. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 99–108

Kessler B, Nunberg G, Shutze H (1997) Automatic detection of text genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 32–38

Kim Y, Ross S (2011) Formulating representative features with respect to genre classification. Genres Web Comput Model Empir Stud 42:129–147

Kleinberg JM (1999) Hubs, authorities, and communities. ACM Comput Surv 31(4es):5

Kleinbery JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

Kumari KP, Reddy A (2012) Performance improvement of web page genre classification. Int J Comput Appl 53(10):24–27

Kuncheva LI, Bezdek JC, Duin RP (2001) Decision templates for multiple classifier fusion. Pattern Recognit 34(2):299–314

Laender AHF, Goncalves MA, Cota RG, Ferreira AA, Santos RLT, Silva AJC (2008) Keeping a digital library clean: new solutions to old problems. In: Proceedings of the 8th ACM Symposium on Document Engineering, pp. 257–262

Lam L, Suen CY (1996) Majority vote of even and odd experts in a polychotomous choice situation. Theory Decision 41(1):13–36

Lee Y, Myaeng S (2002) Text genre classification with genre-revealing and subject-revealing features. In: Proceedings of the 25th ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, pp. 145–150

Lin Z, King I, Ly MR (2006) Pagesim: a novel link-based similarity measure for the World Wide Web. In: Proceedings of the 5th International Conference on Web Intelligence, pp. 687–693

Lovins J (1968) Development of a stemming algorithm. Mech Transl Comput Linguist 11:22–31

Mason JE, Shepherd M, Duffy J, Keselj V, Watters C (2010) An n-gram based approach to multi-labeled web page genre classification. In: Proceedings of the 46th Hawaii International Conference on System Sciences, pp. 1–10

Mehler A, Gleim R, Wegner A (2007) Structural uncertainty of hypertext types. an empirical study. Proceedings of the International Workshop:Towards Genre-Enabled Search Engines: The Impact of NLP, pp. 13–19

Mitchell T (1997) Machine learning. McGraw-Hill, New York

Orrite C, Rodriguez M, Martinez F, Fairhurst M (2008) Classifier ensemble generation for the majority vote rule. In: Proceedings of the 13th Iberoamerican Congress on Pattern Recognition, pp. 340–347

Pereira DA, Ribeiro BN, Ziviani N, Alberto HF, Goncalves AM, Ferreira AA (2009) Using web information for author name disambiguation. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 49–58

Pritsos D, Stamatatos E (2013) Open-set classification for automated genre identification. In: Proceedings of the 35th European Conference on Information Retrieval Research, pp. 207–217

Qi X, Davison B (2008) Classifiers without borders: incorporating fielded text from neighboring web pages. In: Proceedings of the 31st Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development on Information Retrieval, pp. 643–650

Salton G, McGill MJ (1986) Introduction to modern information retrieval. McGraw-Hill, New York

Santini M (2006) Common criteria for genre classification: Annotation and granularity. In: Workshop on Text-based Information Retrieval. In Conjunction with the 21st European Conference on Artificial Intelligence(ECAI), pp. 1–6

Santini M (2007) Characterizing genres of web pages: Genre hybridism and individualization. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences, pp. 71–80

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv (CSUR) 34(1):1–47

Sharoff S, Wu Z, Markert K (2010) The web library of babel: evaluating genre collections. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, pp. 3063–3070

Stamatatos E, Fakotakis N, Kokkinakis G (2000) Text genre detection using common word frequencies. In: Proceedings of the 18th Internation Conference on Computational Linguistics, pp. 808–814

Stein B, zu Eissen SM (2006) Is web genre identification feasible? In: 17th European Conference on Artificial Intelligence (ECAI 06), pp. 815–816

Vapnik V (1995) The nature of statistical learning. Springer, New York

Yang Y, Liu X (1999) A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, pp. 42–49

Zhu J, Zhou X, Fung G (2011) Enhance web pages genre identification using neighboring pages. In: Proceedings of the 12th International Conference on Web Information System Engineering, pp. 282–289

Zu Eissen SM, Stein B (2004) Genre classification of web pages: user study and feasibility analysis. In: 27th Annual German Conference on AI (KI 04), pp. 256–269