TERM WEIGHTING BASED ON INDEX OF GENRE FOR WEB PAGE GENRE CLASSIFICATION

Sugiyanto[1], Nanang Fakhrur Rozi[1], Tesa Eranti Putri[2], Agus Zainal Arifin[2]
[1] Informatics Department, Institut Teknologi Adhi Tama Surabaya
[2] Informatics Department, Institut Teknologi Sepuluh Nopember Surabaya
Email: sugianto@itats.ac.id[1], nanang@itats.ac.id[1], tesa12@mhs.if.its.ac.id[2], agusza@cs.its.ac.id[2]

ABSTRACT

Automating the identification of the genre of web pages becomes an important area in web pages classification, as it can be used to improve the quality of the web search result and to reduce search time. To index the terms used in classification, generally the selected type of weighting is the document-based TF-IDF. However, this method does not consider genre, whereas web page documents have a type of categorization called genre. With the existence of genre, the term appearing often in a genre should be more significant in document indexing compared to the term appearing frequently in many genres despites its high TF-IDF value. We proposed a new weighting method for web page documents indexing called inverse genre frequency (IGF). This method is based on genre, a manual categorization done semantically from previous research. Experimental results show that the term weighting based on index of genre (TF-IGF) performed better compared to term weighting based on index of document (TF-IDF), with the highest value of accuracy, precision, recall, and F-measure in case of excluding the genre-specific keywords were 78%, 80.2%, 78%, and 77.4% respectively, and in case of including the genre-specific keywords were 78.9%, 78.7%, 78.9%, and 78.1% respectively.

***Keywords**:* genre, web page classification, term weighting.

1 INTRODUCTION

Nowadays, World Wide Web is evolving very rapidly, making the number of web pages increases. This results in the difficulty for search engines to find a web page that is needed easily and quickly out of thousands of web pages [1]. One of the solutions to this problem is to classify web pages according to the genre of web pages [2]. Automating the identification of the genre of web pages has become an important area in the classification of web pages [3], as it can be used to improve the quality of web search results and to reduce the search time [4].

Genre is one of the basic properties of web pages [4]. The term genre is derived from the Greek word, 'genus', meaning type or kind. Until now, there has been no definitive understanding among researches about what is mean by the term genre. Using genre is a more effective way to retrieve a web page rather than its content, this method is proven useful in many areas [5],[6]. For example, a search query on a particular topic such as "Oracle" results in many documents related to the "Oracle" company from Internet search engines, but these documents have different genres, such as company websites, product specifications, product advertisings, and product reviews [4].

In order to classify the genre of web pages by using a machine learning approach, it is necessary to identify effective features of each web page based on the genre first [7],[8],[9]. Its main purpose is to extract a set of features, which enables the classifier to distinguish the genre and set the correct label genre for each web page.

Classification in information retrieval system requires a good term weighting [10]. Santini [11] used the TF-IDF term weighting to perform genre classification of web pages. TF-IDF combines the Term Frequency (TF), which measures the density of terms in a document, multiplied by the Inverse Document Frequency (IDF), which measures how informative a term is (its scarcity in the entire corpus) [12],[13].

The document has a type of categorization called genre [11], and not much have yet explore this. Thus, the result of classification between genre and not genre is just the same. A term that often appears in a genre should be important in document indexing, compared to the term that appears frequently in many genres, although its TF-IDF value is high. Weighting a term using TF-IDF is based on the document, in a purpose of determining the index of it [14], [15]. The accurate indexing also depends on how informative a term is to a genre (its scarcity in the entire genre). Term that often appears in many genres should not be an important term despite its high TF-IDF value.

Some genre index should also be a key term for the documents in that genre. In addition, how informative a term is to a genre (its scarcity in the entire genre) is also noteworthy. Some terms that often appear in a genre will certainly have high TF-IDF value, but the term cannot necessarily be said as a key term before its scarcity is calculated based on whole genre. Terms that often appear in many genres

should not have a high value, because it does not reflect the genre index.

Therefore, in this paper, we propose a new weighting method based on the genre of a web page document called Inverse Genre Frequency (IGF). This method is based on a genre that is the manual categorization done by the author in semantics. With this method, term that often appears on many genres will have a small value. This method is expected to have more accuracy, more precision, higher recall and F-measure than TF-IDF.

2 MODEL, ANALYSIS, DESIGN, AND IMPLEMENTATION

2.1 Dataset

The dataset used in this paper is a 7-genre corpus used in the research developed by Santini [11]. This corpus is consisted of 1.400 web pages in English language. These web pages are divided evenly by 200 pages to each of 7 genres, as shown in Table 1.

Table 1 Seven web genre classification

Web genre	Number of web pages
Blog	200
E-shop	200
FAQ	200
Online Newspaper frontpage	200
Listing	200
Personal Home Page	200
Search Page	200

2.2 Term Weighting

Document classification uses vector space model representation of a dataset. The document in vector space model is represented in a matrix containing weights of each term in a document [12], [13]. The weight indicates the significance of a term in a document and in collection of documents. The significance of a term can be seen from its frequency of occurrence. Usually, different terms have different frequencies.

Term Frequency (TF) is the simplest method to weight each term. Each term is assumed to have a significance that is proportional to the number of occurrence of terms in documents. The weight of term *t* on document *d* is calculated using equation (1).

$$\mathrm{TF}(\mathrm{d},\mathrm{t}) = \mathrm{f}(\mathrm{d},\mathrm{t}), \quad (1)$$

where *f(d,t)* represents the frequency of occurrence of term *t* on document *d*.

For genre classification, each term is assumed to have a significance that is proportional to the number of occurrences of terms in genres. The weight of term *t* on genre *g* is calculated using equation (2).

$$\mathrm{TF}(\mathrm{g},\mathrm{t}) = \mathrm{f}(\mathrm{g},\mathrm{t}), \quad (2)$$

where *f(g, t)* represents the frequency of occurrence of term *t* on genre *g*.

If the term frequency takes notice on the occurrences of terms in a document, the IDF takes the occurrences of terms in collection of documents into account. The background of this type of weighting is to value a rarely appearing term in the whole collection of documents more. The significance of each term is assumed to have the opposite proportion to the number of documents containing the particular term. The IDF factor of term *t* is calculated using equation (3).

$$\mathrm{IDF}(\mathrm{t}) = 1 + \log(\mathrm{Nd} / \mathrm{df}(\mathrm{t})), \quad (3)$$

where *Nd* is total number of documents and *df(t)* represents number of documents containing term *t*.

IGF on the other side is taking account on the occurrences of terms in a collection of genres. A term rarely appearing in a collection of genres is considered valuable. The significance of each term is assumed to have the opposite proportion to the number of genres containing the term. The IGF factor of term *t* is calculated using equation (4).

$$\mathrm{IGF}(\mathrm{t}) = 1 + \log(\mathrm{Ng} / \mathrm{gf}(\mathrm{t})), \quad (4)$$

where *Ng* is total number of genres and *gf(t)* represents number of genres containing term *t*.

2.3 Feature Selection Method using TF-IDF mean and TF-IGF mean

For each term in all documents, calculate the weight using TF-IDF and TF-IGF. The weights then averaged by the number of documents [16] and the number of genres respectively. Equation 5 and 6 are the formulas for doing the weight averaging process.

$$\mathrm{TF-IDF}(\mathrm{t}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{TF-IDF}(\mathrm{d_i},\mathrm{t})}{\mathrm{n}}, \quad (5)$$

$$\mathrm{TF-IGF}(\mathrm{t}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{TF-IGF}(\mathrm{g_i},\mathrm{t})}{\mathrm{n}}, \quad (6)$$

where d_i was *i*-th document and g_i was *i*-th genre. The *n* in TF-IDF formula was the number of documents and the number of genres for TF-IGF formula respectively. Then, the weights are sorted in descending mode and the top ranked terms chosen for classification purpose. We chose 100, 500, and 1000 best terms for that purpose. Figure 1 shows the diagram of the feature selection scheme.

2.4 Genre-specific Keywords

Santini [11] proposed 137 specific keywords that can be used to identify the genre of a web page, as shown in the **Error! Reference source not found.**. Santini [11] did automatic extraction of the most common content words per genre to select these sets of specific keywords. Then she measured the coverage of these specific keywords over the web pages belonging to a single genre. This was done to obtain lemmas from adjectives, nouns, and verbs occurring on 80% of web pages with genre. She then

created a new list based on qualitative analysis of the web genre, in order to make the specific keywords more comprehensive and general. Keywords such as post, day, and comment are example of good keywords since blog genre often contains sentences like “posted by”, “comment” etc.

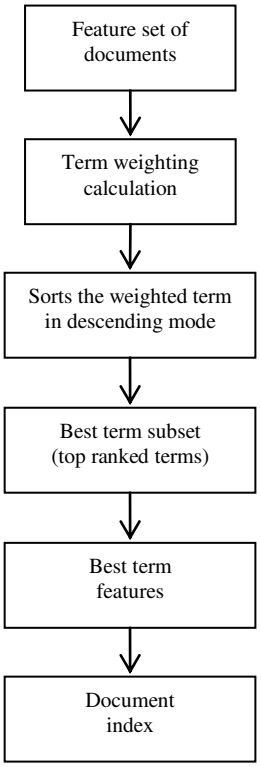

Figure 1 Feature selection scheme

2.5 Random Forest

Random Forest is a Decision Tree-based method which utilizes the concept of bagging [17, 18] in which it uses many trees in the classification process. There are two stages in the Random Forest, the tree growing stage and the classification stage. The first stage begins with the bootstrapping process, i.e. sampling N-cases at random from the training set. From the sampled version of the training set, random sub-setting is conducted for growing the tree which selects some of the input attributes randomly with the provisions of $m \prec d$, where m is the number of selected attributes and d is the total number of attributes. The bootstrap and random sub-setting process is done repeatedly until k-number of tree has been grown. Then, in the classification stage, majority vote of those trees’ classification result is conducted in order to obtain the majority class, which is the final output of the Random Forest as shown in Figure 2.

It is known that the candidate attributes used in the earlier growing process of each tree were not the whole attribute but only partial. Thus, all the grown trees were tree, which differs in shape and size. The expectation of such mechanism is the ability of the trees to minimize its correlation from each other [19] so that it will make the bagging mechanism running more robust than if we only use the standard bagging mechanism [20].

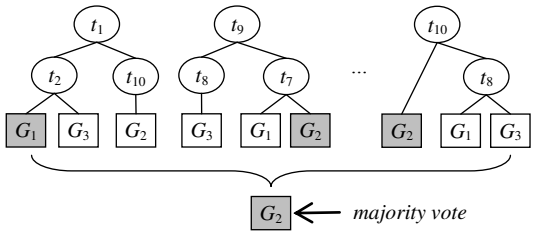

Figure 2 Random Forest Classification Scheme

Figure 2 shows the classification process in Random Forest. It built more than one tree randomly and made decision by doing the majority vote from all of the trees’ classification result.

3 EXPERIMENT

Th experiments were conducted to compare the performance of each term weighting method TF-IDF and TF-IGF for classifying web pages genre. The performance of each term weighting method and its effect are presented as accuracy, precision, recall, and F-measure values of the web pages genre classification results. Accuracy, precision, recall, and F-measure measures are calculated using equation (7), (8), (9), and (10) respectively.

Accuracy = (TP + TN) / (TP + FP + FN + FP), (7)
Recall (R) = TP / (TP + FN), (8)
Precision (P) = TP / (TP + FP), (9)
F-measure = 2TP / (2TP + FP + FN), (10)

where:

- True Positive (TP) is the number of positive data correctly classified as positive
- True Negative (TN) is the number of negative data correctly classified as negative
- False Positive (FP) is the number of negative data misclassified as positive
- False Negative (FN) is the number of positive data misclassified as negative

The experiments used the terms extracted from 1050 web page documents and stemmed using Porter-Stemmer algorithm [21]. [22]. Term weighting methods were applied on the terms before used on the classification of 350 web page documents comprised of 7 categories, each consisted of 50 documents. In these experiments, we selected Random Forest to classify the documents.

We began the experiment by picking the dataset: the dataset with the genre-specific keywords included or excluded. Then we applied the tested term weighting method on the dataset. The next step was the feature selection process as explained in chapter 2.3. After that, we used the terms gained

from feature selection process for classification using Random Forest. Finally, we examined and evaluated the results of the classification. Figure 3 shows the steps of the conducted experiments.

The experiments were conducted twice for each method, with the 137 genre-specific keywords included in the dataset or excluded. For each method, we used different number of terms, respectively 100, 500, and 1000 terms. With these adjustments, we had 12 different test scenarios. The scenarios are shown in Table 2.

At the end of each scenario, each document label would be compared with the actual label. The comparison result would then be displayed in quantitative form, in average accuracy, precision, recall, and F-measure values. Hence, each scenario produced four values.

Table 2 Test scenarios

No.	Method	Genre keywords	Number of best terms
1	TF-IDF	Including keywords	100
2			500
3			1000
4		Excluding keywords	100
5			500
6			1000
7	TF-IGF	Including keywords	100
8			500
9			1000
10		Excluding keywords	100
11			500
12			1000

4 EXPERIMENTAL RESULT

The presentation of the experimental result will be divided into two parts: the classification experimental result with the 137 genre-specific keywords included and not. The experiment was conducted three times, with the variation in number of terms, 100, 500, and 1000 terms respectively.

4.1 Excluding the Genre-specific Keywords

In this experiment, 137 genre-specific keywords were excluded from the dataset. The performance of the TF-IDF weighting method with 100 terms and its effect on the classification result are presented in the form of confusion matrix in Table 3.

Table 3 TF-IDF confusion matrix with 100 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	6	5	0	0	0	39	0
eshop	1	0	0	1	0	47	1
faq	0	0	0	0	0	50	0
fpage	0	0	0	0	0	50	0
listing	0	0	1	1	5	42	1
php	1	0	0	1	0	47	1
spage	1	0	2	0	0	47	0

The performance of the TF-IGF weighting method with 100 terms and its effect on the classification result are presented in the form of confusion matrix in Table 4.

Table 4 TF-IGF confusion matrix with 100 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	45	1	0	0	0	3	1
eshop	0	23	1	4	3	13	6
faq	0	0	43	0	7	0	0
fpage	0	1	0	49	0	0	0
listing	1	3	2	1	18	22	3
php	2	2	1	0	7	32	6
spage	0	1	0	7	7	19	16

The performance of the TF-IDF weighting method with 500 terms and its effect on the classification result are presented in the form of confusion matrix in Table 5.

Table 5 TF-IDF confusion matrix with 500 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	24	1	0	1	1	23	0
eshop	1	6	1	2	1	38	1
faq	0	1	3	0	4	42	0
fpage	6	2	2	4	0	36	0
listing	3	2	3	1	11	26	4
php	2	0	0	4	0	43	1
spage	1	3	0	3	0	38	5

The performance of the TF-IGF weighting method with 500 terms and its effect on the classification result are presented in the form of confusion matrix in Table 6.

Table 6 TF-IGF confusion matrix with 500 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	46	0	0	0	1	3	0
eshop	0	39	0	0	2	5	4
faq	0	0	42	0	0	8	0
fpage	0	0	0	50	0	0	0
listing	1	6	1	1	18	20	3
php	1	2	0	0	1	44	2
spage	1	4	0	1	0	11	33

The performance of the TF-IDF weighting method with 1000 terms and its effect on the classification result are presented in the form of confusion matrix in Table 7.

Table 7 TF-IDF confusion matrix with 1000 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	39	3	0	3	0	5	0
eshop	0	21	5	6	3	14	1
faq	0	2	4	0	1	41	2
fpage	1	8	0	23	6	10	2
listing	2	3	4	4	12	18	7
php	3	2	1	2	1	38	3
spage	0	7	0	5	3	32	3

The performance of the TF-IGF weighting method with 1000 terms and its effect on the classification result are presented in the form of confusion matrix in Table 8.

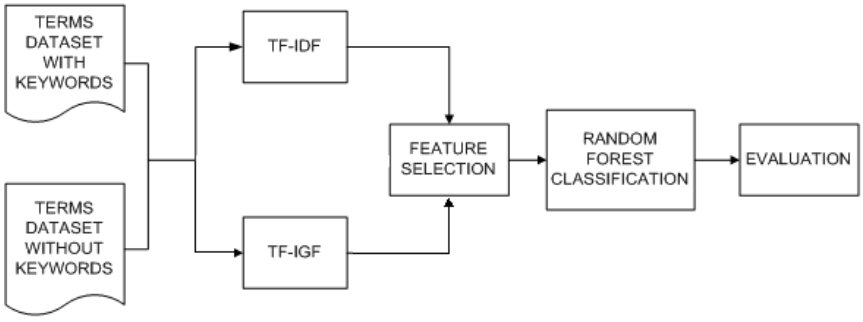

Figure 3 Experiment scheme

Table 8 TF-IGF confusion matrix with 1000 terms

x	blog	eshop	faq	fpage	listing	php	spage
blog	45	1	0	0	0	4	0
eshop	0	41	0	1	1	5	2
faq	0	0	48	0	0	2	0
fpage	0	0	0	49	0	0	1
listing	1	6	1	4	19	14	5
php	2	3	0	0	2	41	2
spage	0	3	0	2	2	13	30

The rows in confusion matrix are the actual genre label of the web pages, while the columns are the predicted genre label of the web pages as the result of classification. By excluding the genre-specific keywords and using the terms weighted with TF-IDF method, many of the web pages were misclassified, as shown in confusion matrix in Table 3, Table 5, and Table 7. On the contrary, using the terms weighted with TF-IGF increases the number of correctly classified web pages, as seen from the confusion matrix in Table 4, Table 6, and Table 8. It should also be noted that in TF-IDF confusion matrix, the increase in number of terms also significantly increase the correctly classified documents, showing that with more amount of data samples used, the better classification would be. This effect is not apparent in TF-IGF confusion matrix. Another interesting phenomenon in the matrix is the many number of web pages that were misclassified to the php genre. This might happen due to the web pages format used in php genre, which is a format widely used on many other genres. Thus, there is a possibility that this php genre overlaps with other genres.

The overall classification performance using the terms weighed with TF-IDF and TF-IGF are presented in the form of accuracy, precision, recall, and F-measure values for each scenario with 100, 500, and 1000 best terms. These four values are shown in Table 9.

Table 9 shows that the accuracy of classification using the terms weighted with TF-IGF is far better compared to TF-IDF. This increase in accuracy is linearly proportional to the number of terms used, further proving the importance of decent amount of data samples for more accurate classification result. For each of precision, recall, and F-measure, there is a significant increase in value. However, there are no specific patterns of values of precision, recall, and F-measure, related to the increase in number of terms used.

Table 9 Evaluation result of the experiment with the specific keywords excluded

Number of terms	Method	Evaluation			
		Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
100	TF-IDF	16.5714	25.89	16.57	9.11
	TF-IGF	64.5714	66.94	64.57	64.25
500	TF-IDF	27.4286	41.79	27.43	24.91
	TF-IGF	77.7143	81.85	77.71	77.52
1000	TF-IDF	40	43.04	40	37.82
	TF-IGF	78	80.17	78	77.38

4.2 Including the Genre-specific Keywords

In this experiment, 137 genre-specific keywords were included in the dataset. The performance of the TF-IDF weighting method with 100 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 10.

The performance of the TF-IGF weighting method with 100 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 11.

The performance of the TF-IDF weighting method with 500 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 12.

The performance of the TF-IGF weighting method with 500 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 13.

The performance of the TF-IDF weighting method with 1000 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 14.

The performance of the TF-IGF weighting method with 1000 terms + genre-specific keywords and its effect on the classification result are presented in the form of confusion matrix in Table 15.

Table 10 TF-IDF confusion matrix with 100 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	47	0	1	0	2	0	0
eshop	1	28	3	1	2	6	9
faq	0	0	39	0	4	1	6
fpage	2	8	4	34	2	0	0
listing	3	9	2	4	7	14	11
php	3	4	4	1	5	26	7
spage	1	3	8	4	6	6	22

Table 11 TF-IGF confusion matrix with 100 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	48	0	0	0	1	1	0
eshop	1	31	1	1	2	5	9
faq	0	0	44	0	3	0	3
fpage	0	4	0	44	1	0	1
listing	2	7	3	1	12	15	10
php	3	2	1	2	8	24	10
spage	1	4	1	5	2	6	31

Table 12 TF-IDF confusion matrix with 500 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	47	1	0	1	0	1	0
eshop	1	32	2	1	3	3	8
faq	0	0	41	0	2	1	6
fpage	2	2	4	39	2	0	1
listing	6	6	5	4	6	14	9
php	2	5	3	2	7	20	11
spage	1	9	4	4	2	7	23

Table 13 TF-IGF confusion matrix with 500 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	46	2	0	0	1	1	0
eshop	0	36	0	1	1	6	6
faq	0	0	47	0	2	1	0
fpage	0	0	0	49	1	0	0
listing	2	5	2	2	12	17	10
php	2	2	0	0	7	33	6
spage	1	5	0	1	2	7	34

Table 14 TF-IDF confusion matrix with 1000 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	45	0	0	3	0	1	1
eshop	0	33	1	2	0	5	9
faq	0	0	39	0	1	3	7
fpage	2	2	0	45	0	1	0
listing	4	3	7	2	10	13	11
php	4	2	4	4	4	24	8
spage	2	5	9	2	1	5	26

Table 15 TF-IGF confusion matrix with 1000 terms + genre-specific keywords

x	blog	eshop	faq	fpage	listing	php	spage
blog	47	1	0	0	0	1	1
eshop	0	39	0	0	2	4	5
faq	0	0	49	0	0	1	0
fpage	0	0	0	50	0	0	0
listing	1	7	1	1	18	13	9
php	3	2	0	0	4	35	6
spage	0	3	0	0	4	5	38

When including genre-specific keywords to the dataset, the confusion matrix using terms weighted with TF-IDF displayed in Table 10, Table 12, and Table 14, apparently undergoes significant increase in number of correctly classified web pages documents. Moreover, the result is most likely equal to the classification using terms weighted with the confusion matrix in Table 11, Table 13, and Table 15, which used TF-IGF. The number of web pages that were misclassified to the php genre is also decreased. However, different from the classification without using the keywords, the increase in number of terms no longer gives notable change on the amount of correctly classified web pages documents. This occurrence signifies that the existence of genre-specific keywords gives a distinct characterization that distinguishes one genre to another.

The values of accuracy, precision, recall, and F-measure for each scenario with 100, 500, and 1000 best terms are shown in Table 16. Table 16 shows that with the genre-specific keywords included in the dataset, the accuracy of classification improves significantly for TF-IDF method. This further proves that the completeness of dataset will result in much more accurate classification.

Table 16 Evaluation result of the experiment with the specific keywords included

Number of terms	Method	Evaluation			
		Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
100	TF-IDF	58	55.94	58	56.53
	TF-IGF	66.8571	65.68	66.86	65.77
500	TF-IDF	59.4286	56.32	59.43	0.5722
	TF-IGF	73.4286	72.6	73.43	72.33
1000	TF-IDF	63.4286	63.64	63.43	61.66
	TF-IGF	78.8571	78.74	78.86	78.09

Compared to the condition where the genre-specific keywords were excluded from the dataset, the value of precision, recall, and F-measure produced with TF-IGF method has higher value than the previous experiments. However, on TF-IDF, there is a drastic increase on each value of precision, recall, and F-measure after the keywords were added. Overall, TF-IGF still show a better accuracy, precision, recall, and F-measure value, either by including the genre-specific keywords or not. Hence, TF-IGF term weighting method based on genre index is able to perform better in doing classification of web pages genre, compared to TF-IDF.

TF-IDF could be combined with TF-IGF to form TF-IDF-IGF, but the combination resulted in conflict between document index and genre index. The combination would reduce the weight of terms frequently appearing in a genre, so it can eliminate the important terms that should be selected. The combination would boost the weight of terms

frequently appearing in a genre, so there would be new terms that is actually less important.

5 CONCLUSION

In this paper, we proposed a new term weighting method to optimize classification of web page genre utilizing genre categorization called TF-IGF. We applied TF-IGF term weighting method on web pages dataset based on manual categorization done semantically from previous research. Experimental results show that the term weighting based on index of genre (TF-IGF) performed better compared to term weighting based on index of document (TF-IDF). The value of accuracy, precision, recall, and F-measure in case of excluding the genre-specific keywords was 78%, 80.2%, 78%, and 77.4% respectively, and in case of including the genre-specific keywords was 78.9%, 78.7%, 78.9%, and 78.1% respectively. The four evaluation values are increased in TF-IDF and TF-IGF if the specific keywords for each genre were added to the dataset. However, TF-IGF gave more stable and better results, whether with or without the addition of the specific keywords for each genre. In the future work, the 7-genre corpus dataset used could be structured in hierarchy so the classification of the web pages would be even more specific and yield better accuracy, precision, recall, and F-measure.

6 REFERENCES

[1] Chen, G., and Choi, B., 2008. ”Web page genre classification”. Proceedings of the 2008 ACM symposium on Applied computing, page 2353-2357, ACM New York, NY, USA.

[2] Dong, L., Watters, C., Duffy, J., and Shepherd, M., 2008. ”An Examination of Genre Attributes for Web Page Classification”. Proceedings of the 41st Hawaii International Conference on System Sciences.

[3] Pritsos, D.A., and Stamatatos, E., 2013. ”Open-Set Classification for Automated Genre Identification”. ECIR 2013, LNCS 7814, pp. 207–217, Springer-Verlag Berlin Heidelberg.

[4] Eissen, S.M.Z, and Benno, S., 2004. Genre Classification of Web Pages. Advances in Artificial Intelligence Springer Berlin Heidelberg Volume 3238, 2004, pp 256-269.

[5] Vidulin, V., Lustrek, M., and Gams, M., 2007. ”Training a Genre Classifier for Automatic Classification of Web Pages”. Journal of Computing and Information Technology, Volume 4, page 305–311.

[6] Jebari, C., and Wani, M.A., 2012. ”A Multi-label and Adaptive Genre Classification of Web Pages”. IEEE 11th International Conference on Machine Learning and Applications (ICMLA).

[7] Wu, Z., Markert, K., and Sharoff, S., 2010. ”Fine-grained Genre Classification using Structural Learning Algorithms”. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 749–759.

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

[9] Lei, D., Watters, C., Duffy, J., and Shepherd, M., 2008. ”An Examination of Genre Attributes for Web Page Classification”. IEEE Proceedings of the 41st Annual Hawaii International Conference on System Sciences.

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

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

[12] Manning, C.D., Prabhakar, R., and Hinrich, S., 2009. An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.

[13] Gebre, B.G., Zampieri, M., Wittenburg, P., and Heskes, T., 2013. ”Improving Native Language Identification with TF-IDF Weighting”. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223, Atlanta, Georgia, June 13 2013.

[14] Fuji, R., and Mohammad, G.S., 2009. ”Class-indexing-based term weighting for automatic text classification”. Journal of Informetrics, vol. 3, no. 1, pp. 72-77.

[15] Gautam, J., and Kumar, E., 2013. ”An Integrated and Improved Approach to Terms Weighting in Text Classification”. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, January 2013.

[16] Tang, B., Michael, M., Shepherd, E., and Heywood, M.I., 2005. Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering. Faculty of Computer Science, Dalhousie University, Halifax, Canada, B3H 1W5.

[17] Ali, J., Khan, R., Ahmad, N., and Maqsood, I., 2012. ”Random Forests and Decision Trees”. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[18] Breiman, L., 2001. Random Forests. Machine Learning 45: 5–32.

[19] Hastie, T.J., Tibshirani, R.J., Friedman, J.H., 2008. The Elements of Statistical Learning: Data-mining, Inference and Prediction. Second Edition. New York: Springer-Verlag.

[20] Zhu, M., 2008. Kernels and Ensembles: Perspectives on Statistical Learning. The American Statistician 62: 97 – 109.
[21] Karaa, W.B.A., 2013. "A New Stemmer To Improve Information Retrieval". International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013.
[22] Willett, P., 2006. The Porter stemming algorithm: then and now. Program: electronic library and information systems, 40 (3). pp. 219-223.

APPENDIX

Table 17 Genre-specific keywords

Genre	Keyword
Blog	archive, comments, diary, blog, journal, posted by, weblog, web log, Monday, mon, Tuesday, tues, Wednesday, wed, weds, Thursday, thurs, Friday, fri, Saturday, sat, Sunday, sun, january, jan, february, feb, march, mar, april, apr, may, june, jun, july, jul, august, aug, september, sept, october, oct, november, nov, december, dec
E-shop	buy, basket, catalogue, store, shop, story, sell, trolley, cart, price, rebate, cost, purchase, order, offer, checkout, save, pay, debit card, credit card, shipping, delivery
FAQ	faq, frequently, asked, question, answer, enquir, inquir, A:, assistance, enquir/inquir, help, q&a, Q:, question
Frontpage	column, editor, front page, frontpage, headline, news, newspaper, opinion, report, stories, story
Listing	Checklist, check list, contents, hotlist, hot list, sitemap, site map, toc, table of contents, index, step, list, map
Php	about me, cv, curriculum vitae, guestbook, guest book, my page, homepage, home page, page, my site, my web-site, my webpage, my web site, my web, interests, personal web-site, personal web site, personal page, resume, research, publications, project, 's page, 's webpage, vita, 's web page
Search page	Engine, directories, crawl, search, find, see, advanced search