Text Genre Classification with Genre-Revealing and Subject-Revealing Features

Yong-Bae Lee & Sung Hyon Myaeng
Dept. of Computer Science, Chungnam National University
Daejon,Korea
+82-42-821-5446
{yblee, shmyaeng}@cs.cnu.ac.kr

ABSTRACT

Subject or prepositional content has been the focus of most classification research. Genre or style, on the other hand, is a different and important property of text, and automatic text genre classification is becoming important for classification and retrieval purposes as well as for some natural language processing research. In this paper, we present a method for automatic genre classification that is based on statistically selected features obtained from both subject-classified and genre-classified training data. The experimental results show that the proposed method outperforms a direct application of a statistical learner often used for subject classification. We also observe that the deviation formula and discrimination formula using document frequency ratios also work as expected. We conjecture that this dual feature set approach can be generalized to improve the performance of subject classification as well.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: System and Software – *Current awareness systems, efficiency and effectiveness.*

General Terms: Performance, Experimentation, Verification.

Keywords: text genre, classification, statistical method, tf, idf, subject categories.

1. INTRODUCTION

Text classification research has been mostly focused on subject or prepositional content of text. As text-based applications have become more diverse and the amount of information has increased tremendously, however, different aspects of text can be useful for various purposes including classification. We focus on text genre or the style of text that often characterizes the purpose for which the text has been written. Examples for genre are: research article, novel, poem, news article, editorial, homepage, advertisement, manual, court decision etc.

Genre classes are clearly different from subject classes that most classification research has dealt with. From an information retrieval point of view, a retrieval query about a certain topic such as "IBM" would get many documents related to the company IBM from an Internet search engine, but they may be of different genre such as a company homepage, product specification, product advertisement, or critical review of a certain product. As such, classifying documents based on genre would result in a totally different outcome than that from ordinary subject-based classification.

Automatic genre classification has been studied in the recent past. Karlgren and Cutting [6] explored the use of structural cues and rather simple cues such as counts of third person pronouns in text with discriminant analysis. In subsequent work[4, 5], she investigated the relationship between the genre of retrieved vs un-retrieved documents and relevant vs non-relevant documents. Used features are simple statistics, such as sentence length and word length, and syntactic complexity such as average depth of a parse tree.

Identifying text genre would be beneficial to many text-based applications. For instance, if the genre of every document is known a priori, information retrieval results could be better presented to the user, depending on the preference the user has. As pointed out by Kessler *et al.* [7], the performance of many natural language processing tools, such as part-of-speech tagging, parsing, and word sense disambiguation, could be enhanced since some language usages embedded in grammatical constructions and word senses are related to the genre of text. In Web applications, genre detection would help wrappers that attempt to extract specified information from semi-structured.

Kessler *et al*. [7] identified cues in four categories: structural cues (e.g. counts of POS tags), lexical cues (e.g. words used in expressing dates), character-level cues (e.g. punctuation marks), and derivative cues (e.g. average sentence length as a ration and standard deviation in sentence length as a variation). They decided not to use the structural cues because of the high computational cost. Their computational methods were logistic regression and neural networks (a simple perceptron and multi-layer perceptron) that combine 55 cues.

More recently, Stamatatos *et al*. [10] reported on their work for genre detection using word frequencies and punctuation marks. Instead of using sophisticated linguistic cues, they attempted to develop a method that works for unrestricted text in any domain and language with minimal computational cost in extracting cues.

In this paper, we take the stance more related to traditional information retrieval and text classification approaches than to natural language processing. Instead of focusing on sophisticated linguistic cues and their values as to how they are related to individual genre classes, we attempt to improve upon a direct application of a well-studied text classification method to genre classification by using the *tf* and *idf* statistics. Our hypothesis is that the use of the features selected and the weights calculated for one

SIGIR'02, August 11-15, 2002, Tampere, Finland.

type of classification (e.g. genre) can be complemented with the use of features and weights obtained for other type (e.g. subject). By using both genre class and subject class training data for a statistical classifier, we can achieve a better performance than simply using the same classifier with genre class training data alone.

2. THE GENRE CLASSIFIER

2.1 Method

The main thrust of the genre classifier we developed is the use of the statistics from two different class sets, genre classes and subject classes, in the training data. The goodness value of a term for a genre depends on three factors:

- how many documents belonging to the genre contain the term
- how evenly the term is distributed among the subject classes that divide the genre class, and
- how discriminating the term is among different genre classes.

The main idea for the first two factors is to find terms that are found in as many genre documents as possible and distributed as evenly as possible among all subject classes. Our assumption is that a good genre-revealing term would show up across different subject classes while appearing in many documents in the genre. In other words, even if a term appears in many documents belonging to a particular genre, we don't wan it to be specific to a particular subject area that happens to be discussed heavily in the subset. Therefore it would be helpful to eliminate the terms that are too specific to a certain subject area by looking at their distributions. Furthermore, the third factor ensures terms that are good indicators for many genre classes are downgraded since they are not good discriminators.

The table in Appendix illustrates the point made above. Among the genre classes, we chose the 'personal homepage' genre that has been identified as a truly digital genre as opposed to those borrowed from the paper world [3]. In our preliminary analysis, 533 documents belonging to the 'personal homepage' genre class are sub-divided into four subject categories. The numbers in parentheses represent the number of documents in each genre or subject class. The term 'home' appearing in many of the 288 genre class documents also has high document frequencies for all the subject classes. It should be considered a good genre-revealing term. On the other hand, the term 'play' having a high document frequency (154) for the genre class but appearing only in the 'celerity' subject class is not a good representative for the genre.

2.2 Computation

For each term in the training document set, we compute its goodness value in two ways. The first is to use a term's document frequency ratios for genre and subject classes, i.e. in how many documents belonging to a genre class or to a subject class within the genre class the term appears. More specifically, $DFR_m(t_k)$ for a term t_k 's document frequency ratio for genre m, is:

$$DFR_m(t_k) = \frac{df_{m,k}}{df_m}$$

Likewise, $DFR_m(t_k^i)$ for the term appearing in the subject class i is defined to be:

$$DFR_m(t_k^i) = \frac{df_{m,k}^i}{df_m^i}$$

The goodness value for a term in a genre class can now be computed as:

$$V_{m,k} = DFR_m(t_k) * (1 - \sigma) \quad (1)$$

where deviation σ is defined to be:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n_c} (DFR_m(t_k) - DFR_m(t_k^i))^2}{n_c}}$$

Here n_c is the number of subject categories. The formula in the square root computes the degree to which the term is distributed among various groups of documents corresponding to subject classes, or variance of term distributions among the subject classes. $DFR_m(t_k)$ serves as the mean document frequency ratio for all the subject classes, and $DFR_m(t_k^i)$ is the document frequency ratio for t_k in the subject class i within the genre class m. The smaller deviation in the subject classes, the better the term is. The sum of the document frequencies for all the subject classes is equal to the document frequency for the genre class.

Alternatively we can compute the goodness values using term frequencies as follows:

$$V_{m,k} = \frac{tf_{m,k}}{\max_{1 \le i \le n} [tf_{m,i}]} * (1 - \sigma) \quad (2)$$

where the first part is the normalized term frequency, and σ is newly defined as:

$$\sigma = \sqrt{\frac{1}{n_c}} \times \sqrt{\sum_{i=1}^{n_c} \left(\frac{tf_{m,k}^i}{n_c \cdot tf_m} - \frac{tf_{m,k}^i}{tf_m^i} \right)^2}$$

where the $tf_{m,k}$ and $tf_{m,k}^i$ represent the total frequency of t_k in the genre class m and in the subject class i within the genre class, respectively. The first part of the formula is a normalizing factor.

Accommodating the discrimination factor above, the final term weight for t_k in the genre m becomes:

$$W_{m,k} = \sqrt{\frac{\sum_{i=1}^{n_g} (V_{m,k} - V_{i,k})^2}{n_g}} \quad (3)$$

where n_g is the number of genre classes and $V_{i,k}$ is the goodness value for the term in the i-th genre class only.

Given the term weights above a threshold for each of the genre classes, calculated in the training stage, we can form a vector for each genre class G_m. The genre decision for a new document D is computed as follows:

$$\max_m [sim(G_m, D)] \quad (4)$$

where the cosine similarity is used for $sim(G_m, D)$.

This method for computing the term weight for a genre class can be used for other kinds of classifiers. In our study, the naïve Bayesian classifier approach (see, for example [8]) was considered[1]. By applying the approach directly to genre classification, we can compute the probability of a document belonging to a genre class g_i as follows:

$$P(d \mid g_i) = P(g_i)\prod_{k=1}^{T} P(t_k \mid g_i) \quad (5)$$

where $P(t_k \mid g_i)$ is the probability of a term appearing in the genre class g_i, which can be computed from a training document set.

Given this framework, we were interested in understanding the role of our term selection method, i.e., the method of using both genre and subject classes in the training document set. It is quite natural to consider the weight $W_{m,k}$ in (3) as an estimate for the probability of a term k representing the genre class m, so that we get:

$$P(t_k \mid g_i) = P(g_i) \prod W_{i,k}$$

Here $P(g_i)$ can be estimated with the ratio of the number of documents in the genre i to the total number of documents in the training set. This is compared against the case where the term frequency observed in the documents belonging to the genre class i are used to estimate the probability $P(t_k \mid g_i)$.

3. EXPERIMENTS

3.1 Testing Ground

We collected Web documents to test the genre classification method proposed in this paper. While Brown Corpus has been used for some previous genre-related research, it doesn't have subject classes assigned to individual documents, which are critical to our study. Also, our primary interest lies in Web applications.

For practical purposes, we considered seven genre classes: reportage, editorial, technical paper, critical review, personal homepage, Q&A, and product specification. These genre classes would help Web search engine users distinguish what they want from the rest of the documents retrieved, all of which contain subject-bearing query terms.

The total numbers of Korean and English documents collected are 7,828 and 7,615, respectively. The Korean documents were collected by a linguist and a computer scientist and the English documents by three people who majored in English literature and by one linguist. More than twenty portal sites were used in the collection building process to eliminate possible bias toward particular document types. Each document was examined by at least two people for inclusion in the collection as well as in the designated genre and subject classes. A half of the collected documents in each sub-collection, Korean and English, was used for training and the other half for testing. Fig. 1 shows the numbers of Korean and English documents in each genre class, together with possible category classes they can belong to.

[1] While the naïve Bayesian classifier approach is known to be inferior to other approaches like support vector machines and k-nearest neighbors [11], we chose it because of its simplicity and extensibility to Web documents with links as in [9].

At the beginning, we were most interested in the general performance level of the proposed genre classification method. It wasn't clear how effective it would be to use the idea of selecting terms based on the notion of document frequency and term frequency ratios, not other more linguistically oriented features, from genre-specified training documents. We had two approaches to be tested: one based on document frequency ratios and the other based on term frequency ratios.

Genre classes and subject classes within each genre		Korean	English
Reportage	robbery, fraud, violence, suicide, murder, fire, …	929	815
Editorial	economy, education, sports, international, politics,…	750	849
Research articles	engineering, arts, basic science, biomedical, …	1,051	1,200
Reviews	education, finance, foods, culture, sports, cosmetics, students, teachers, professors, celebrity, …	2,362	1,490
Homepage	laws, customers, cuisine, medicine, computer, …,	906	1,067
Q&A	computer, cosmetics, sports, video	960	1,020
Spec		870	1,174
Total		7,828	7,615

Fig. 1. Document counts in each genre and subject classes

Immediately after finding out the performance was reasonable, we launched other experiments to see the value of using both genre-revealing and subject-revealing terms for genre classification. In particular, we wanted to compare the following cases:

Case 1: the use of term frequency values in the genre-specified training documents only,

Case 2: the use of document frequency (or term frequency) ratios to compute the deviation between two distributions of terms: one from the genre-specified and the other from the subject-specified training documents (using the formula (1) or (2))

Case 3: Case 2 (using both document and term frequency ratios) together with the use of inter-genre discrimination power of terms (using the formula (1) or (2) together with (3))

The Naïve Bayesian approach as in (5) and the similarity approach as in (4) were both used as the genre determination method for testing documents. In the experiments, we used both English and Korean document collections to confirm whatever trend we see in one language is also the case in the other.

Effectiveness was measured with micro-average recall/precision scores [11] when different cases were compared. Since no duplicate classes are assigned to each document, precision and recall values are the same in our experiments.

3.2 Results and Discussions

3.2.1 Overall Comparisons

Table 1 shows the summary of the results indicating the differences among the three cases, using *tf* (ordinary method) only, using

deviation values, and using the combination of deviation values and the discrimination values but with *df* or *tf* ratios. They were applied to the Naïve Bayesian approach and the similarity approach using the English collection. The numbers are micro-average precision/recall values.

Table 1. Overall comparisons among different approaches

cases \ approaches		Similarity-Based	Naïve Bayesian Approach
Case 1	df	0.75	-
	tf	0.63	0.83 (0.75)
Case 2	df	0.81 (0.87)	0.80 (0.74)
	tf	0.79	-
Case 3	df	0.87 (0.90)	0.79 (0.70)
	tf	0.80	-

Case 1: with ordinary tf values
Case 2: with deviation values using df ratios or tf ratios
Case 3: with deviation values using df or tf ratios and discrimination values

First of all, the result shows that in the similarity-based approach, the use of *df* ratios gives significantly better performance than *tf* ratios across all the cases. This is in line with our initial hypothesis that *df* ratios would be a better factor than *tf* ratios in this context. The other promising result is that the use of the deviation formula and the discrimination formula used in Case 2 and Case 3, respectively, actually improves the performance, again confirming our initial hypothesis. In other words, the deviation formula helps selecting terms that are more genre-related than subject-related, and the discrimination formula also helps increasing the weight of a term that appears in a smaller number of genre classes.

It is interesting to see that the overall performance orders in the cases are quite different when the naïve Bayesian approach is used. The best result was obtained when the ordinary *tf* values were only used. The use of the *df* ratios and the new formulas gave slightly worse results. We suspect that in order to get an optimal result, it is important to stick to the literal interpretation of the Naïve Bayesian approach, especially in computing the probability of a term occurring in a genre class, instead of estimating with *df* ratios and the deviation and discrimination formulas.

We were able to repeat many of the performance comparison results using the Korean collection. The micro-average precision/recall values for the three cases under the naïve Bayesian approach were 0.75, 0.74, and 0.70, respectively, and the same for Cases 2 and 3 under the similarity-based approach were 0.87 and 0.90. Again the superiority of using *tf* values was observed when the Naïve Bayesian approach was used. In the similarity-based approach, we only confirmed the superiority of the Case 3 against the Case 2. The performance difference between the new method and the best Naïve Bayesian approach was greater in the Korean collection than in the English collection.

3.2.2 Inter-genre variations

Table 2 is an example of a detailed result for the proposed method (Case 3) using *df* ratios with performance numbers for individual genre classes in the English collection.

Table 2. Documents assigned to the seven genre classes

Actual \ Guess	A	B	C	D	E	F	G
A (425)	351	67	3	0	1	3	0
B (408)	26	381	1	0	0	0	0
C (745)	35	11	655	1	6	27	10
D (600)	1	3	1	586	7	2	0
E (534)	25	4	32	10	435	26	2
F (510)	16	11	61	4	45	372	1
G (587)	0	1	46	2	2	9	527
P/R (%)	77/ 83	80/ 93	82/ 88	97/ 98	88/ 81	85/ 73	98/ 90

A: Editorial B: Reportage C: Review D: Research Paper
E: Homepage F: Q&A G: Product Spec

Looking at the first row, we can see that among the 425 documents in the 'editorial' genre (A), 351 documents are correctly assigned and 67 are incorrectly assigned to the 'reportage' genre (B). The first column, on the other hand, indicates that among all the documents assigned to the 'editorial' genre (A) by the classifier, 351 are correct where as 26 were incorrectly assigned to the 'reportage' genre.

By browsing through the numbers in the cells, we can easily identify the genre classes that confuse the classifier. For instance, confusion between the 'reportage' and 'editorial' genre classes is the greatest. On the other hand, there was no confusion at all between the 'editorial' and 'product spec' genre classes. The confusion case is understandable because documents in the two genre classes are both from newspapers, perhaps sharing some subject-bearing terms. This indicates that our use of terms alone, as opposed to other more linguistically oriented features in sentences, has a limitation in some genre classification.

Table 3. Documents assigned to six genre classes

Actual \ Guess	A+B	C	D	E	F	G
A+B (833)	812	12	0	2	7	0
C (745)	25	673	0	6	31	10
D (600)	3	1	583	11	2	0
E (534)	12	39	8	441	32	2
F (510)	14	67	4	48	377	0
G (587)	1	48	0	1	11	526
P/R (%)	94/ 97	80/ 90	98/ 97	87/ 83	82/ 74	98/ 90

A+ B: Newspaper C: Review D: Research Paper
E: Homepage F: Q&A G: Product Spec

Table 3 is the case where two of the genre classes that confused the classifier most, 'reportage' and 'editorial', are merged into 'newspaper'. Overall, it clearly gives a better precision/recall values (micro-average being 90% vs 87%) than the seven-genre case. The

precision/recall values for the newly formed genre class are much higher than either in the seven-genre case

3.2.3 Tuning the threshold

In order to obtain the best possible results for each case, so that the comparisons are made among them, we had to adjust the cutoff levels for *tf* and *df* ratio values in selecting the terms to represent each genre class that are used in subsequent computation. That is, we modulated the number of genre-revealing terms to be considered by adjusting the thresholds for *tf* and *df* ratios manually. The number of terms selected for the seven genre classes are in table 4.

Table 4. Selected term numbers for each genre class

Genre	A	B	C	D	E	F	G
# terms	89	86	131	130	166	117	128

The adjustments were based on the observations of correlations between the threshold and the performance values. This can be done when a result as in Table 2 is obtained. The general strategy is to increase the threshold (i.e. to select a smaller number of terms) when precision for the particular genre class is low, and to lower the threshold (i.e. to select a larger number of terms) when recall for the particular genre class is low. Using a larger number of terms for the genre class means that more documents would be classified into the genre, resulting in higher recall for the genre. On the other hand, a smaller number of terms for the genre class would reduce the number of documents assigned to it, increasing precision.

4. CONCLUSION

In this paper we presented a new methodology for genre classification using word statistics. In particular, we introduced the *df* ratio, *df* ratio deviation formula, and discrimination formula that are combined to select genre-revealing terms from the training document set. The deviation formula makes use of both genre-classified documents and subject-classified documents to eliminate terms that more subject-related than genre-related.

We constructed our own Web document collections in English and Korean, and tested effectiveness of the proposed method against other possibilities using some of the formulas as well as *tf* and the *tf* ratios within the same similarity-based framework. It turned out that the *df* ratio always gave a better result than *tf* itself or *tf* ratios, although the ordinary *tf* values gave the best result when the Naïve Bayesian framework was employed. The proposed method outperformed the best case of the Naïve Bayesian results.

Analyzing the results in detail, we realized that some genre classes are significantly confusing to the classifier because they tend to share some subject-specific terms, although we attempted to eliminate subject-related terms using the deviation formula. We probably need to explore other ways to obtain the same effect. In addition, we need to consider other more linguistically oriented features and expressions in our future study so that the mutually confusing genre classes are separated more easily.

Another avenue we need to explore further is the method for determining the number of genre-revealing terms as well as the effects of different numbers in more detail. The number of terms we used for each genre is much larger than the number (around 30) found in other study [10]. The heuristic method we used to decide on the term numbers for individual genre classes need to be modeled into an automatic method.

Finally, further studies are needed to compare the proposed method against other frameworks that can be applied to genre classification. We are in the process of devising a method for using both *tf* and *idf* and applying the current method to other classification theories such as k-nearest neighbor and support vector machine.

5. ACKNOWLEDGMENTS

Our thanks go to Enquest Technology, Inc. for allowing us to use the English and Korean genre collections.

REFERENCES

[1] Ivan Bretan, John Dewe, Anders Hallberg, Niklas Wolkert, Jussi Karlgren, "Web-Specific Genre Visualization", Proc. of the 30th Hawaii International Conference on System Science, Jan 1997.

[2] Johan Dewe, Jussi Karlgren, Ivan Bretan, "Assembling a Balanced Corpus from the Internet", 11th Nordic Conference of Computational Linguistics, pages 100-107, Copenhagen, 1998.

[3] Andrew Dillon, Barbara A. Gushrowski, "Genre and the Web: Is the Personal Home Page the First Uniquely Digital Genre?", JASIS, 51(2):202-205, 2000.

[4] Jussi Karlgren, "Stylistic Variation in an Information Retrieval Experiment", Proc. of the 2nd International Conference on New Methods in Language Processing-NeMLaP, 1996.

[5] Jussi Karlgren, Ivan Brettan, Johan Dewe, Anders Hallberg, Niklas Wolkert, "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres", 8th DELOS Workshop on User Interfaces in Digital Libraries, pages 85-92, 1998.

[6] Jussi Karlgren, Douglass Cutting, "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis", Proc. of COLING94, Kyoto, 1994.

[7] Brett Kessler, Geoffrey Nunberg, Hinrich Schűtze, "Automatic Detection of Text Genre", ACL'97, pages 32-38, July 1997.

[8] D. Lewis and M. Ringuette, "Compariosn of two learning algorithms for text categorization," Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[9] H. J. Oh, S. H. Myaeng, and M. Lee, "A practical hypertext categorization method using links and incrementally available class information", Proc. of the 23rd ACM SIGIR Conference, pages 264-271, Athenes, Greece, 2000.

[10] E.Stamatatos, N.Fakotakis, G. Kokkinakis, "Text Genre Detection Using Common Word Frequencies", Proc. of the 18th International Conference on COLING2000, 2000.

[11] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proc. Of the 22nd ACM SIGIR Conference, 1999.

Appendix

Genre		Subject									
Personal Homepage (533)		Student (149)		Teacher/ Professor (144)		Company Employee (128)		Celebrity (112)	
University	397	school	106	university	133	university	103	the	100
the	354	university	102	professor	101	school	83	play	99
school	329	html	94	the	98	home	76	film	89
home	288	the	84	teach	94	the	72	act	77
page	262	page	77	education	94	education	72	actor	70
html	245	student	75	science	93	association	70	television	65
education	239	computer	73	research	90	office	62	role	65
college	216	charset	73	state	87	law	62	home	64
text	203	text	72	page	82	com	62	star	60
state	203	home	70	school	80	experience	58	school	60
include	203	time	66	home	78	computer	58	new	60
science	202	web	64	department	77	include	57	theatre	54
time	201	college	58	student	74	state	56	time	53
charset	201	name	57	publication	74	practice	56	page	53
student	193	education	57	college	74	personal	56	eye	53
computer	187	friend	55	study	67	html	55	production	52
new	184	iso	53	graduate	67	mail	54	include	51
graduate	184	graduate	52	faculty	66	service	52	love	50
study	173	experience	51	national	64	college	52	hair	49
experience	173	edu	51	technology	63	bar	52	angel	49
this	172	science	50	new	63	attorney	52	height	47
mail	172	this	49	edu	63	page	50	award	46
web	167	major	49	professional	62	business	48	stage	45
research	167	high	49	american	62	legal	47	series	45
information	165	music	48	association	61	system	46	movie	44
american	163	live	48	institute	61	science	46	commercial	44
play	154	people	47	assistant	61	lawyer	46	train	43
com	152	email	47	mail	60	information	46	life	43
site	151	site	46	information	60	court	45	career	43
office	151	play	46	associate	59	text	44	brown	43
name	150	don	46	society	57	site	44	born	43
national	148	born	46	international	57	professional	44	html	42
association	139	life	45	course	55	county	43	com	42
system	136	study	44	text	54	charset	43	university	41