Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization

Yong-Bae Lee
Dept. of Computer Education
Jeonju National University of Education
yblee@jnue.ac.kr

Sung Hyon Myaeng
School of Engineering
Information & Communication University
myaeng@icu.ac.kr

Abstract

Genre characterizes text differently than the usual subject or prepositional content that has been the focus of most information retrieval and classification research. We developed a new method for automatic genre classification that is based on statistically selected features obtained from both subject-classified and genre-classified training data. The main idea of the genre classification method is to calculate the weight of a feature for a genre class by using its frequency statistics for subject classifications. Having observed that the deviation formula and discrimination formula using document frequency ratios work as expected, we went on to study the roles of various types of features such as content-bearing words, function words, morphemes, and punctuations marks. In the first part of this paper, we present some of our findings in the roles of the feature types for genre classification, with a brief discussion of the genre-based classification. The genre classes we used are those often found in Web documents: accident reportages, newspaper editorials, personal homepages, product reviews, product specifications, research articles, and Q&A's. The second part of the paper addresses the issue of how text genres help classifying documents based on the subject content of documents. This is a corollary to our original hypothesis that subject classification would help identifying the genre class of a document automatically. Our experimental work shows that while subject classes clearly help improving the genre-based classification, it is not clear whether using the genre class information for documents in the same way helps subject-based classification. However, we found that training a subject classifier with a set of documents belonging to a particular genre class improves subject-based classification.

1. Introduction

Genre characterizes text differently than the usual subject or prepositional content that has been the focus of most information retrieval and classification research. We view text genre or the style of text as characterizing 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. As text-based applications have become more diverse and the amount of information has increased, different aspects of text, such as genre, can prove useful for various purposes. In this article, we address two issues: automatic detection of the genre class of text using subject-based classification information and the roles of genre in subject-based classification.

Genre classes are clearly different from subject classes that most classification research has dealt with. Even though a set of documents may belong to the same class because they share the common topic, they often times serve different purposes, falling into diverse genre classes. As such, classifying documents based on genre would result in a totally different outcome than that from ordinary subject-based classification. From the traditional information retrieval point of view, a retrieval query about a certain topic such as "Samsung" would retrieve many documents related to the company when submitted to 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. Genre provides a new dimension for text retrieval and classification, in addition to topicality, and help users become more selective in their information seeking process and obtain higher quality information.

Automatic genre classification has been studied in the recent past[1, 2, 3]. 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*. [11] 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.

For genre identification, we take the stance more related to traditional information retrieval and text categorization approaches than to deep natural language processing. Our text analysis is at the level sufficient to obtain statistics of term and symbol distributions relying on *tf* and *idf*. Features generated from the text analysis are used for genre-based classification of documents. Our approach is unique in that we developed our own deviation-based statistical feature selection method utilizing subject-based classification information of the training documents.

Having developed a genre classifier, we began to investigate on the issue of how text genres help classifying documents based on the subject content of documents. In other words, the question is whether the knowledge about a genre class of a document would help determining its subject class. This is a corollary to our original hypothesis that subject classification would help identifying the genre class of a document automatically. Some experimental results are provided, but the work included in this paper is quite exploratory in nature.

2. The Genre Classifier

2.1. Overall method for feature extraction

On the surface, the genre classifier we developed is no different from the traditional learning-based classifier. The learner extracts features representing genre classes from training documents whose genre classes are known already, and a classification algorithm determines to which class a new document should belong using the learned representations of the genre classes. The major difference lies in the method by which features are extracted and their weights are calculated. In comparison with previous genre classification approaches, furthermore, the difference lies in the types of features we use as well as the extraction method itself.

The feature extraction method was derived from our observation that the frequency of a feature (e.g. a word) may be high in a set of document belonging to a genre class, not because it represents the particular genre class, but because it represents a subject class to which many training documents happen to belong. This phenomenon is likely to happen especially when the training documents are not collected randomly from the entire document space.

As such our feature selection method uses the statistics from two different class sets, genre classes and subject classes, in the training data. The weight of a feature for a genre is computed based on three factors [8]:

- how many training documents belonging to the genre contain the feature (test 1)
- how evenly the feature is distributed among the subject classes that divide the genre class (test 2), and

- how discriminating the feature is among different genre classes (test 3).

The main idea for the first two factors is to find features that are found in as many genre documents as possible and distributed as evenly as possible among all the subject classes that divide the training documents in the genre class. Our assumption is that a good genre-revealing feature would show up across different subject classes while appearing in many documents in the genre. In other words, even if a feature appears in many documents belonging to a particular genre class, we don't wan it to be specific to a particular subject area that happens to be discussed heavily in the training documents.

Furthermore, the third factor ensures good features are as specific to a genre class as possible by downgrading the features that happen to occur in several genre classes. This is important when the training documents are collected from a single subject area, disabling the first two tests, or when a feature occurs in many training documents. The latter case is the main reason for using *idf* in addition to *tf* for traditional information retrieval.

2.2. Computation

For each feature in the training document set, we incorporate two different types of weight: intra-genre weight and inter-genre weight [8]. The first is to use a feature's document frequency ratios for genre and subject classes, i.e. in how many documents in a genre class or in a subject class within the genre class the feature appears. More specifically, $DFR_m\ (t_k)$ for a feature 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 feature appearing in the subject class i is defined to be:

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

The weight of a feature in a genre class can now be computed as:

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

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 feature is distributed among various groups of documents corresponding to subject classes, or variance of feature 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 feature is. The sum of the document frequencies for all the subject classes is equal to the document frequency for the genre class. The first part of the formula (1) reflects the test 1 and the second test 2 explained above.

Alternatively we can compute the feature weight using feature frequencies as follows using term frequency statistic, term frequency ratios, rather than document frequency:

$$TFR_m(t_k) = \frac{tf_{m,k}}{tf_m}$$

- $tf_{m,k}$: the frequency of term k in a genre class m
- tf_m : the total frequency of all terms in a genre class m

$$TFR_m(t_k^i) = \frac{tf_{m,k}^i}{tf_m^i}$$

- $tf_{m,k}^i$: the frequency of term k that appears in a subject class i within a genre class m
- tf_m^i : the total frequency of all terms in a subject class i within a genre class m

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

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

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

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, as defined above. The first part of the formula is a normalizing factor.

Accommodating the discrimination factor (test 3) above, the final feature 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 feature weight in the i-th genre class. According to this formula, a feature whose weights in all the genre classes are similar to each other would get a low weight.

2.3. Classification algorithm

Given the feature weights above a threshold for each of the genre classes, calculated in the training stage, we can form a feature vector representing each genre class G_m. Each genre class may have a single feature vector for all the different types of features or different vectors for different feature classes such as nouns, pronouns, suffixes, punctuation marks, etc. In this sub-section, we consider a single vector for a class. Multiple vectors can be used to build multiple classifiers and integrate the classification results from them as described later in the experiment section.

The genre classification decision for a new document D is made by computing the similarity between a document vector and each of the genre class vectors and selecting the most similar one:

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

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

The method for computing the feature weight for a genre class can be used for other kinds of classifiers. In our study, the naïve Bayesian classifier approach (see, for example [9]) 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 feature 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 feature extraction 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 feature k representing the genre class m, so that we get:

$$P(d \mid g_i) = P(g_i) \prod_{k=1}^{T} W_{i,k} \quad (6)$$

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 feature 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

This section reports on our testing of the initial hypothesis that knowing the subject class (secondary class) of training documents helps classifying documents based on genre class (primary class), where the secondary class partitions the documents belonging to a primary class. The feature extraction method and the associated classification algorithm were compared against the case without using the secondary class information. In addition, we ran experiments to see to what extent several different types of features such as nouns, pronouns, proper nouns, verb endings, exclamations, and special characters are useful for genre-based classification.

3.1. Testing ground

The documents we used for training and testing were collected from the Web[2] in seven genre classes: reportage, editorial, technical paper, critical review, personal homepage, Q&A, and product specification. The collected documents were classified into subject categories using the hierarchy to which they were assigned in the Web sites and the results were examined manually to correct possible errors.

The total numbers of Korean and English documents collected are 7,828 and 7,615 respectively. The Korean documents were

[1] While the naïve Bayesian classifier approach is known to be inferior to other approaches like support vector machines and k-nearest neighbors [12], we chose it because of its simplicity and extensibility to Web documents with links as in [10].

[2] 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.

Genre classes	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, …	2,362	1,490
Homepage	students, teachers, professors, celebrity, …	906	1,067
Q&A	laws, customers, cuisine, medicine, computer, …	960	1,020
Spec	computer, cosmetics, sports, video, …	870	1,174
Total		7,828	7,615

Figure 1. Document counts in each genre and subject Classes

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 document types determined by the Web sites. 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 subject classes they can belong to. 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 [12] 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. Overall effectiveness

The first experiment was to see how effective the proposed genre classification method is, in comparison with a direct application of the Naïve Bayesian approach to genre-based classification. In other words, the combination of the proposed feature selection method (formula (3)) and the similarity-based classification (formula (4)) was compared against the method using formula (5) and (6). The formula (5) is pure the Naïve Bayesian method using *tf* whereas the formula (6) is a variation using the proposed feature extraction method in place of the usual interpretation of the probability based on *tf*.

The result was promising in that the score for our proposed method was 0.87 for English documents, which is 4.8% improvement over the *tf*-based Naïve Bayesian approach (0.83), and 0.90 for Korean documents, which is 10.6 % improvement over the the *tf*-based Naïve Bayesian approach (0.75). An interesting result is that when the proposed feature extraction method was applied to the Naïve Bayesian approach, the performance dropped. Similarly, when the *tf* approach was applied to the similarity-based approach, the performance was lower.

Further experiments show the contribution of individual steps of the proposed method. In particular, we compared the following cases:

Case 1: the use of *tf* or *df* values only in the genre-specified training documents only,

Case 2: the use of *df* ratios to compute the deviation between two distributions of features: 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 *df* ratios) together with the use of inter-genre discrimination power of features (using the formula (1) or (2) together with (3))

Table 1 shows the summary of the results indicating the differences among the three cases, using *tf* (ordinary method) or *df* only, using deviation values, and using the combination of deviation values and the discrimination values with *df*. They were applied to the Naïve Bayesian approach and the similarity approach using the English and Korean collections. The numbers in parentheses are for the Korean collection.

Table 1. Overall comparisons among different approaches

Approach / Case	Similarity-Based	Naïve Bayesian Approach
Case 1	0.75 (0.72)	0.83 (0.75)
Case 2	0.81 (0.87)	0.80 (0.74)
Case 3	0.87 (0.90)	0.79 (0.70)

Case 1: with ordinary tf (Naïve Bayesian) or df (Similarity-Based) 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

We were encouraged by the promising result that the use of the deviation formula and the discrimination formula used in Case 2 and Case 3, respectively, actually improves the performance, confirming our initial hypothesis. In other words, the deviation formula helps selecting features that are more genre-related than subject-related, and the discrimination formula also helps increasing the weight of a feature that appears in a smaller number of genre classes.

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.

3.3. Effectiveness of various feature types

It is well known that nouns are more important than any other part-of-speech words in information retrieval, especially with Korean text. As such, only nouns were considered in the evaluations described above, too. Other research in genre identification also used nouns alone or some special symbols like punctuation marks [6, 10]. In our work, we evaluated the importance of pronouns and proper nouns separately from general nouns and other parts-of-speech such as verb endings and exclamations because they may reveal the genre of a document even though they are not content-bearing words. Special symbols included as a feature group are: '(', ')', '[', ']', '<', '>'. '!', '?', '.', ';', ':', '=', '#', '+', '-', '~', '*', '^', '@', '&', '$', '%', a comma, and a single and double quotes.

Table 2 shows the micro-average precision/recall values achieved by individual feature types. While general nouns as a group are most effective features, other types do make some contributions to genre classification when applied individually.

Table 2. Effectiveness of Various Feature Types

Feature Types	Precision/Recall
General Nouns	0.899
Pronouns	0.360
Exclamations	0.414
Verb Endings	0.448
Person Names	0.180
Special Symbols	0.107

We investigated further to see the synergistic effect of pairs of feature types by taking the linear combination of the two feature vectors for a genre class, which correspond to the pair. Table 3 shows that the gains by linearly combining two similarity values for a document with respect to a genre class are not much. It seems that the information gained by other feature types are already reflected by the nouns. The best is when the nouns and person nouns are combined.

Table 3. Effectiveness of Combinations of Feature Types

	Pronouns (0.360)	Exclamations (0.414)	Verb endings (0.448)	Person Names (0.180)	Special Symbols (0.107)
Nouns (0.899)	0.900	0.902	0.901	0.905	0.900
Pronouns (0.360)		0.439	0.448	0.413	0.407
Exclamations (0.414)			0.476	0.485	0.450
Verb endings (0.448)				0.507	0.448
Person Names (0.180)					0.180

Since the best possible value we obtained by combining all the feature types was 0.906, a marginal improvement, we concluded that the individual feature types are not of great value

when the noun type achieves a high precision/recall already.

4. Roles of Genre in Subject-Based Categorization

In the experiments described above we confirmed the hypothesis that subject-based classification (secondary class) information for training documents helps classifying documents based on the genre class (primary class). It is natural to ask a similar question whether genre-based class information can be useful for subject-based classification. If the answer for this question is positive, it will be a good support for a more general hypothesis that information about a secondary class for training documents helps classifying documents based on a primary class.

Another use of genre-based classification information is to build a classifier (i.e. extract features) for a class of documents corresponding to each genre class. The genre-specific subject-based classifier can be applied to a new document whose genre class is known to be the same. For instance, a set of training documents belonging to the 'review' genre is used to extract features for individual subject classes to form a classifier. This classifier is then used to determine the subject class of an incoming document whose genre class has been determined to be 'review'.

In the next sub-section, we provide a preliminary report on our experiments for the following two hypotheses:

- Hypothesis 1: Genre classification information for training documents helps subject-based classification when it is used with the deviation formula proposed in this research.
- Hypothesis 2: Training a classifier with a set of documents belonging to a particular genre class improves subject-based classification.

4.1. New testing ground

With this goal in mind, we ran experiments by which we classify documents based on subjects using the genre class information for the training documents. For this purpose, we collected a total of 2,660 documents in Korean with seven subject classes (education, culture, video/audio, sports/leisure, automobile, computer, cosmetics), which fall into six different genre classes (research article, review, editorial, reportage, product specification, Q&A). Figure 2 shows the statistics of the new collection.

The documents are entirely different from those used in the previous experiments because those used in the genre-based classification belong to so diverse subject classes that the number of documents in each subject class is usually too small to run experiments. As can be seen in Fig. 2, the genre classes used are a subset of those used previously to ensure that the genre classes determined automatically are actually useful.

The shaded parts and underlined parts form two sub-collections (sub-collection 1 and sub-collection 2) , respectively, by which we tested whether using documents belonging to a specific genre class for training would help subject-based classification. A total of 1,462 documents (sub-collection 1) belonging to the 'review' genre were used in an experiment first. A subsequent experiment used 1,162 documents (sub-collection 2) belonging to the 'spec' genre, which cover only five subject categories excluding 'education' and 'culture'.

Subject (primary class)	Genre (secondary class)	#Docs
Education	Article (12), Review (106), Editorial (70)	188
Culture	Article (20), Review (207), Editorial (53)	280
Video/ Audio	Review (130), Spec (106)	236
Sports/ Leisure	Review (242), Editorial (95), Spec (122)	459
Auto-mobile	Review (103), Reportage (93), Spec (103)	299
Computer	Article (64), Review (407), Editorial (50), Spec (126), Q&A (161)	808
Cosmetics	Review (266), Spec (124)	390

Figure 2. Document Distributions among the Primary and Secondary Classes for the New Collection

4.2. Results and observations

The two separate experiments conducted on the sub-collection 1 and sub-collection 2 are summarized in Table 4. Three different classification methods were used as shown in the first column in the Table: the proposed similarity-based method using the deviation formula and the inter-class discrimination value, the Bayesian formula using the proposed feature extraction, and the pure Naïve Bayesian method

using *tf*. The left half of the table is divided into two columns: one using all the genre classes for Hypothesis 1 and the other using only one genre ('review' in this case) for Hypothesis 2. The right half of the Table is also divided into two in a similar way. The only difference is that the 'spec' genre is used for Hypothesis 2. Since the documents belonging to the 'spec' genre fall into only 5 subject categories (no documents belong to 'education' or 'culture'). It should be noted that for the 'review' only case and 'spec' only case, it wasn't possible to use the deviation formula, and the discrimination power criterion was only used.

Table 4. Results for Testing Hypothesis 1 and 2
Micro-average precision/recall

	Seven Subject Classes		Five Subject Classes	
	All Genre Classes	'Review' Only	All Genre Classes	'Spec' Only
Proposed Method	0.60 (0.62)*	0.51	0.67 (0.66)	0.78
Bayesian with Proposed Feature Extraction	0.52 (0.52)	0.55	0.55 (0.55)	0.71
Naïve Bayesian	0.56	0.63	0.59	0.72

* The numbers in the parentheses are for those cases where no deviation formula was used.

Based on the results, it is difficult to prove or disprove Hypothesis 1. For the 'seven class case' shown in the first column, the performance figures without the deviation formula (those within the parentheses) are better or equal to those with the formula (i.e. using the secondary class information). Even for the 'five class case' shown in the third column, the situation is almost the same. Using the deviation formula with the secondary class (genre in this case) information is not helpful. One positive aspect, however, is that the proposed method outperforms the original Naïve Bayesian approach.

For Hypothesis 2, we used two different genre classes, 'review' and 'spec'. When the documents belonging to the 'review' class were used for training in the experiments, the performance figure is even lower than those obtained with all the genre classes. This result, on the surface, seems to indicate that using the genre class information is actually harmful. However, we found that the 'review' documents show an erratic distribution characteristic with respect to the Web sites from which the documents were collected; most documents were collected from a small number of sites, and the documents from each site tend to have a common page frame generating a set of common words across those documents.

We suspect that those words occurring in most of the 'review' training documents misguided the feature extraction procedure. Since the proposed method uses *df* values, a feature (e.g. a word 'city' in the web site title 'computer city') appearing in most of the training documents in the subject class (say, 'computer'), which in fact has little to do with the subject area, would receive a relatively high weight for the subject class ('computer') at hand. An incoming document with many occurrences of that feature ('city') would be mistakenly classified into the subject class ('computer') even though it is supposed to be classified into a different class (say, 'culture'). This conjecture is supported by the performance obtained from the Naïve Bayesian approach with the 'review' documents alone, which is higher than the 'all genre classes' case and that of the proposed method. Since *tf* values are used for this method, the fact that the feature occurs in many training documents may not have a negative effect because as long as the frequency is not high. In fact, the features included in document frames occur only a few times, at most.

Subject Class	Web Sites & the Number of Documents
Education (53)	EN(49), GC(4)
Culture (104)	EV(94), SS(4), WC(6)
Video (64)	EN(13), EV(31), SS(17), WC(3)
Sports (121)	EN(59), EV(52), GC(29), WC(1)
Automobile (52)	EN(23), EV(13), SS(13), WC(3)
Computer (203)	GC(30), EN(131), EV(37), WC(5)
Cosmetics (133)	EN(47), EV(31), GC(30), SS(24), WC(1)

Figure 3. Distribution of the Documents among the Web Sites (7 subject case)

In the experiment with the documents belonging to the 'spec' genre class, the result was much in line with what we expected. The performance figures across the three methods were much higher than others, indicating that confining training documents within a genre class helps extracting good features. Unlike the case with the 'review' genre, the documents

came from a variety of Web sites with no obvious distributional bias and idiosyncrasy specific to a particular Web site. Figure 3 and 4 show the characteristics of the documents in the two genre classes. The abbreviations represent various Web sites, and the numbers are for the documents collected.

There are only five distinct Web sites involved, and the distributions of documents among them are also skewed in some cases (e.g. education and computer). In particular, the documents obtained from the EV site happen to contain many noise words.

Subject Class	Web Sites & the Number of Documents
Video (53)	DT(7), HA(16), NA(7), MA(2), SP(12), NN(1), MK(2), ET(6)
Sports (61)	GA(6), DE(6), DA(14), DW(3), DP(8), NA(2), MA(2), SP(6), MK(2), SS(7), ET(5)
Automobile (52)	N1(5), N2(23), N3(3), N4(6), DP(3), SP(6), DW(1), MK(1), NN(2)
Computer (63)	LO(15), HY(24), SH(6), SP(7), MK(3), NN(7), DA(1)
Cosmetics (62)	WH(14), NS(9), ET(6), N5(5), MA(7), N6(4), SP(6), MK(1), NN(7), DA(1), LE(2)

Figure 4. Distribution of the Documents among the Web Sites (5 subject case)

In this sub-collection, as many as 26 Web sites are involved, preventing a Web site from dominating the statistics. The distributions of the documents among the sites and among the subject classes are also less skewed.

In order to convince ourselves, we analyzed the 'review genre case' in more detail, attempting to identify the source of errors and prove our conjecture that the bad performance is originated from the nature of the sub-collection. Table 5 shows a confusion matrix where the letters represent the seven subject categories and the number in the parentheses with a letter in the first column is the total number of documents for the subject category.

We found that the errors (99 and 50 documents misclassified into F) made for the class B and C were due to the documents collected from the same web site. Among the 104 testing documents for the class B, more than 90 were from the particular Web site, including some common words that happen to be good features for the class F. On the other hand the four documents classified correctly for the class B are from different sites. The 57 documents that should have been classified into D but actually classified into G are from the Web site that contains the 64 documents correctly classified into G. It turns out that when a site contains a large number of documents of one subject area, some peripheral features included in a frame are given a high weight. If those features happen to be important to another class, errors would occur as described above. This problem may be severe especially when many documents are from a single Web site.

Table 5. Confusion Matrix among the Seven Subject Classes with the Sub-Collection 1

Guess Actual	A	B	C	D	E	F	G	P/R
A(53)	47	0	0	0	0	0	6	0.98/0.89
B(104)	0	4	0	1	0	99	0	0.44/0.04
C(65)	0	0	15	0	0	50	0	1.00/0.23
D(121)	1	1	0	52	1	9	57	0.75/0.43
E(52)	0	0	0	1	29	18	4	0.94/0.56
F(203)	0	0	0	0	1	163	39	0.42/0.80
G(133)	0	4	0	15	0	50	64	0.38/0.48

While the genre classification information did not eliminate the problem mentioned above, the phenomenon is a strong argument for using information for a secondary class for classification based on the primary class. In fact, we suspect that the Web sites on which documents are found can serve as another dimension as is genre.

5. Conclusion

In this paper we first described our methodology for genre classification using feature statistics [8]. The proposed method uses the *df* ratio, *df* ratio deviation formula, and discrimination formula that are combined to select genre-revealing features from the training document set. The deviation formula makes use of both genre-classified documents and subject-classified documents to eliminate features that more subject-related than genre-related. We also summarized the experimental results that prove the efficacy of the method using both English and Korean collections constructed for genre-based classification using subject classes. In addition, we tested the importance of various feature types and their combinations for genre-based classification.

The main focus of this paper, on top of the proposed genre-based classification method, was to investigate on the hypotheses that genre classification information for training documents helps subject-based classification when it is used with the deviation formula proposed in this research, and that training a classifier with a set of documents belonging to a particular genre class improves subject-based classification. Based on the experiments we conducted, it is not possible to prove or disprove the first hypothesis, although the effectiveness level in comparison with the Naïve Bayesian approach is promising enough to warrant further investigation. However, it seems quite safe to conclude that the second hypothesis is reasonable, meaning that the knowledge on the genre of a given document would help classifying it based on its subject more correctly, given that a classifier has been built for documents belonging to the specific genre.

In the course of analyzing the experimental results, we gained an insight that documents obtained from a Web site may have their own characteristics that can be used for classification purposes. We view this information as still another dimension in addition to the subject content and the genre of documents, which needs to be studied further in more detail.

6. Acknowledgments

Our thanks go to Enquest Technology, Inc. for allowing us to use the English and Korean genre collections. This research was supported by KOSEF-funded Software Research Center.

7. 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] Yong-Bae Lee and Sung Hyon Myaeng, "Text Genre Classification with Genre-Revealing and Subject-Revealing Features," Proceedings of the 25th ACM SIGIR Conference, pages 145-150, Tampere, Finland.

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

[10] 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, Athens, Greece, 2000.

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

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