Chapter 6
Formulating Representative Features with Respect to Genre Classification

Yunhyong Kim and Seamus Ross

6.1 Introduction

Document classification is one of the most fundamental steps in enabling the search, selection, and ranking of digital material according to its relevance in answering a predefined search. As such it is a valuable means of knowledge discovery and an essential part of the effective and efficient management of digital documents in a repository, library, or archive. Document classification has previously been dominated by the classification of documents according to topic. Recently, however, there has been a growing interest in the classification of documents with respect to factors other than topic (e.g. classification into forms of dissemination such as scientific papers, emails, blogs, and news reports). This type of classification has been labelled in many different ways, including the phrase *genre classification*. The vast number of different contexts in which genres have emerged across classification attempts illustrate that genre is a high-level, context-dependent concept (cf. literature review [24]). Genre has been referred to as aspects of the text described by level of information or degree of elaboration, persuasion and abstraction (cf. [5]), as well as, to common document forms such as FAQ, Job Description, Editorial or Reportage (e.g. [9, 14, 16]). In some cases, genre has been used to describe the classification of a document according to whether or not it is a narrative and the target level of audience (e.g. [16]), and whether it is fact or opinion, and, in the case

Y. Kim (✉)
Humanities Advanced Technology and Information Institute (HATII),
University of Glasgow, Glasgow, UK; School of Computing, Robert Gordon University,
Aberdeen, UK
e-mail: ykim1@rgu.ac.uk

The work presented in this chapter was supported by DELOS: Network of Excellence on Digital Libraries (http://www.delos.info), funded by European Commission's IST FP6 [G038-507618], and the Digital Curation Centre (http://www.dcc.ac.uk), funded by the Joint Information Systems Committee (JISC, http://www.jisc.ac.uk) and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC, http://www.epsrc.ac.uk)[GR/T07374/01].

A. Mehler et al. (eds.), *Genres on the Web*, Text, Speech and Language Technology 42, DOI 10.1007/978-90-481-9178-9_6,

of opinion, whether it is positive or negative (e.g. [10]). On occasion it has been used to describe membership to selected journals and brochures (e.g. [1]), and, to denote similar feature cluster groups (e.g. [2, 21].

Despite the elusive nature of genre, it is undoubtedly true that being able to bind together tools trained to retrieve information within selected genre domains would be invaluable to automating the ingest, management and preservation of material in digital repositories (cf. [23]). This is especially true where metadata describing the technical characteristics, function, source and content of digital material play a core role in the efficient and effective management and re-use of the same. The manual collection of metadata is labour-intensive, costly and susceptible to variation in quality and precision across different actors; automating the process of semantic metadata extraction is, therefore, essential. Past efforts (e.g. [3, 7, 11, 12, 15, 25]) to extract metadata automatically from digital documents have relied heavily on the structure that characterises the genre class to which the document under consideration belongs. The reliance of these methods on document structure emphasises the benefits of constructing a tool that enables automated genre classification. An effective automated genre classifier would function as an overarching tool for integrating genre-specific tools and, in any case, provide a first-level classification of documents into those of a similar structure, which would facilitate the extraction of further information.

The interest in forms of documents classification other than that of topic is also growing in the area of information retrieval and reflects the limitations of relevance measurements defined on the basis of topical similarity. Topic alone does not provide insight into whether or not a retrieved document is relevant to your purpose; a document with the same topic may be created with different objectives resulting in different levels of usefulness as a source of information (e.g. compare an advertisement about a camera to a product review of the same camera). These objectives of document creation seems to be at the centre of what characterises document genre. On the other hand, these objectives define the functional requirements imposed on the document (e.g. to narrate, to argue against, to argue for, to present research results, to record) and the structures found within the document are designed to meet these functional requirements. In this chapter we do not claim a deep understanding of the nature of genre, but merely are driven by the observation that the structural classification of documents is a fundamental component in understanding a document with respect to its purpose and function.

Classical models of document classification largely depend on term frequency weighting and counting instances of specified linguistic constructs. The former does not reflect much conceptual structure and the latter results in a highly language dependent model that incorporates some local conceptual structure but largely disregards the global structure of the document and its components. In this chapter we examine the role of word distribution pattern in classifying documents. More specifically,

- we describe an approach to document representation that incorporates more document structure by considering how strings are distributed throughout the document (Section 6.2.2), and,
- give evidence that this approach is better than the bag-of-words approach by comparing it against the rainbow classifier developed by McCallum (see [20] and Section 6.6.2).

It is not the purpose of this chapter to advocate the structural classification of documents as a definition for genre classification, but to show by experimental evidence that our model may be more appropriate in dealing with high level concepts (such as genre). We are not disputing the fact that genre is a social construct (cf. the Chapter 2 by Karlgren, this volume) and that it is the social context that defines genre. We wish to merely state that, just as the phenotype of a group of genetically distinct organisms (e.g. whales and fish) may lead to the extinction or survival of the entire group (e.g. if the water should become contaminated), the structure of a document is likely mirror the social objectives related to the document creation and provide a key to gauging the usefulness of a document and extracting further information. In particular, we report evidence that some of the previously established genre schemas and collections are better distinguishable by our distribution model than previously reported results.

The importance of structure has also been discussed elsewhere (e.g. the Chapter 1, by Lindemann and Littig, this volume) but, while others have introduce structure as the measurements of *structural entities* within the document distinct from topical terms or content , we will be discussing structure as an organisation of terms (regardless of topicality) throughout the document (Section 6.2.2) akin to *burstiness of terms* discussed in [6] and again in [8].

The combined representation of content and structure that we are attempting to establish in this chapter is also intended to raise questions about a prevailing notion in earlier analyses that genre classification is a task orthogonal to topic classification (e.g. the Chapter 8 by Stein et al., this volume). While this may be true on a conceptual level, there is reason to believe that this may not be a statistically sound approach. For example, the topic of *algebraic variety*, a well-known subject area in higher mathematics, would not be expected to appear as frequently in the genre class Reportage as it would in the genre class Research Article. In fact, preliminary results from a recent experiment, classifying documents belonging to 10 genre classes into twenty newsgroup topic classes, shows that, while there are genre classes whose documents are randomly distributed across the 20 topics (e.g. Poem), there are also genres 95% of whose documents are classified into only four newsgroup topics (e.g. Minutes). Given these examples where genre is interactively intertwined with topic, it would seem beneficial to build a general classification model that encompasses both tasks. With this in mind, we would like to introduce genre classification, not as a classification task distinct from topic classification, but as a point in a continuum of classifications, emphasising both genre classification and topic classification as a special case of a general abstract classification model.

6.2 Defining Genre Classification

6.2.1 Document Representation in Conventional Text Classification

The conventional method of text classification can be contracted to a formula for the weight of a term T within a document expressed by:

$$TF \times IDF \times N \tag{6.1}$$

where TF denotes the frequency of the term in the document, IDF denotes the inverse of the number of documents in the collection containing the term, and N denotes a normalisation factor dependent on the length of the document. The calculation method of each of these terms differs according to the research or application in question. This model is based on the notion that:

- if a term appears frequently in a document, it is likely to be a characterising feature of the document;
- if a term appears across several documents, then it is not likely to be a strong feature in distinguishing any one of those documents from the others; and
- if the same term appears in equal numbers within a short document and a long one, then it is likely to be a stronger feature of the short document.

While it may be considered a gross simplification to represent all the various classification methods by this one description, it still seems true that the basic principles that drive various text classification methods are closely related to this model. In a subject classification task, the term may surface as words or N-grams (N consecutive words or characters), while in other classification tasks term may manifest itself also as functional groups of words (e.g. verb) or combinations of such words and phrases and groups. Nevertheless, the mechanism driving the classification is largely dependent on counting patterns, and weighing the number against the pattern count throughout the collection being examined. The location of patterns, the relationship between instances of the patterns, and the interplay between different types of patterns are largely by-passed and only represented implicitly through the pattern of the expression being counted.

6.2.2 Harmonic Descriptor Representation (HDR) of Documents

A document can be described as a sequence of symbols. Symbols should not be confused with the alphabet of a natural language, although they may take the form of alpha-numeric characters in some instances. In the present terminology, each symbol may form any group of these characters or a much larger set of characters (e.g. white space, %, + and ?) and could also refer to the functional category of a group of characters (e.g. the part-of-speech).

Because of its static appearance, a document is often misunderstood to be time independent, but the interpretation of each symbol is possible only as a consequence of its temporal relationship to other symbols. In this light, document classification can be considered to be a subtask of signal processing. Viewed in this way, an accurate measure of term frequency is expressed by how many times a symbol occurs with respect to time. The term weight calculated in Section 6.2.1 presents no awareness of the role of temporal progression in the semantic analysis of the document. That is, if the word "clock" were to appear in two documents 10 times, then the weight of this word would be equal with respect to both documents: the fact that the word appears only in the first half of the document with respect to one of the documents in contrast to being evenly distributed throughout the document (which may be the case with respect to the other document) would be disregarded. A proper consideration of the time dimension would suggest "clock" in the first document as a signal having twice the frequency of that of the second document, but lasting only half the length of time. Time should not be taken to be the length of the text. Although the two are closely related, the length of the text is not equivalent to the tempo of the piece of writing, beginning with an introduction and ending with a conclusion. To understand the notion of time, we will compare a document to a string of a musical instrument. An occurrence of a symbol within the document partitions the document into two parts. If the two partitions are equal in length, then the phase division is akin to a harmonic with twice the frequency of the fundamental of the string (the document with zero occurrence of the symbol). If the division is not equal, then the frequency can not be considered to be uniform throughout the document.

In the case of topic detection, a loose application of time (e.g. taking the frequency to be uniform throughout the document) may be sufficient to capture salient vocabulary, but in other types of classification, where the main interest lies in the physical or conceptual structure of the object, the lack of temporal and relational placement of symbols contributes to a considerable loss of information. To fill this gap, we propose incorporating the symbol's range and period as an effective means of characterising the symbols in the document. We will refer to this characterisation as the Harmonic Descriptor Representation (HDR) of the document (inspired by the musical analogy given above). We define range as the interval between the initial and ultimate occurrence of the symbol, and period as the time duration between two consecutive occurrences of the symbol. When the symbol occurs at regular intervals, the resulting signal in the document is akin to a harmonic of the document as a wave. Brookstein et al. [6] observed that content-bearing words would clump together and therefore result in non-harmonic behaviour. In contrast to the content-bearing words that they discuss, our research focuses on words that may be indicative of style and structure. We observe that document structure is captured by words displaying both harmonic and non-harmonic behaviour; harmonic words define the physical structure of the document, while non-harmonic words define conceptual landmarks or structure. In our description, we attempt to capture the degree of non-harmonic behaviour using three quantities derived from the range and period of each symbol:

1. The time duration before the first occurrence within the document of the symbol (FP), measured by the number of characters (including white space) before the symbol, divided by the number of characters in the entire document.
2. The time duration after the last occurrence of the symbol to the end of the document (LP), measured by the number of characters after the last symbol divided by the number of characters in the entire document.
3. The average period ratio (AP), defined as 1 if the maximum number of characters between two occurrences is zero, and, otherwise, as $T/(N \times MP)$, where:

- N is the total number of occurrences of the symbol plus one;
- MP is the maximum number of characters found between two consecutive occurrences of the symbol; and,
- T is the total number of characters in the document minus N.

The average period ratio is an average ratio of the distance between two occurrence over the maximum distance. It is intended to measure how regular the occurrences are, regardless of how far apart the actual occurrences are, as. The more harmonic the behaviour of a symbol, the closer AP will be to 1. The other two measures FP and LP, on the other, hand are intended to measure when the term is first introduced and how focused the occurrences are against the entire document. In Fig. 6.1, we display an example of six documents (D1–D6) of different lengths, portrayed as light-coloured strips where the top of the strip is the beginning of the document. Occurrences of symbols in the documents ($s1$–$s7$) have been represented as horizontal lines across the strips. The period between two consecutive occurrences have been indicated to be x. This example will be used in Figs. 6.2, 6.3, and 6.4 to demonstrate how FP, LP, and AP change under different conditions.

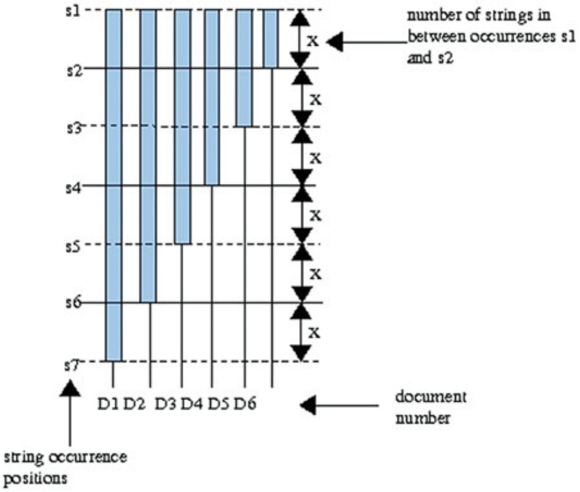
s1
s2
s3
s4
s5
s6
s7
x
x
x
x
x
x
number of strings in between occurrences s1 and s2
D1 D2 D3 D4 D5 D6
document number
string occurrence positions

Fig. 6.1 Example of symbol occurrence in six documents of different lengths

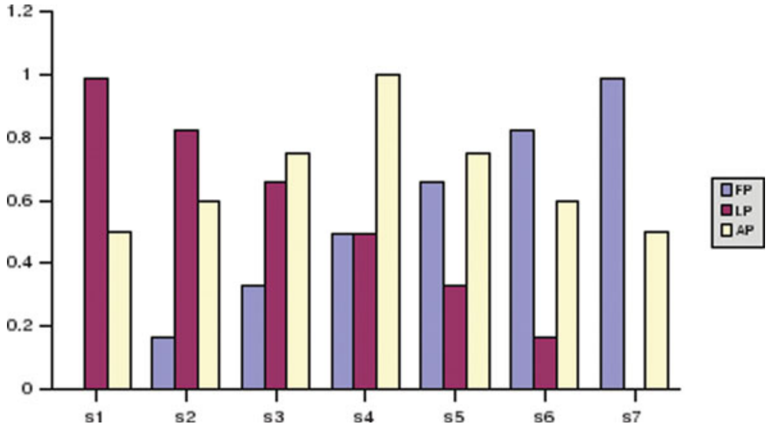

Fig. 6.2 FP, LP, and AP with respect to the position (X-axis) of a single occurrence of a symbol in $D1$

We present in Fig. 6.2, a graph illustrating how FP, LP and AP change as the position of a symbol occurring once in $D1$ (see Fig. 6.1) changes from $s1$–$s7$. In Fig. 6.3, we show how FP, LP and AP for a symbol occurring twice in $D1$ change with respect to the period between the two instances, as the second occurrence of the symbol moves away from the first occurrence. Finally, the graph in Fig. 6.4 presents how FP, LP and AP, for a symbol occurring once halfway between $s1$ and $s2$, change as the document length varies.

Given a document, each word or symbol in the document is associated to their FP, LP and AP values. By taking all the words in a collection or by using a pre-compiled list of indicative words (say, in either case, the resulting word list is

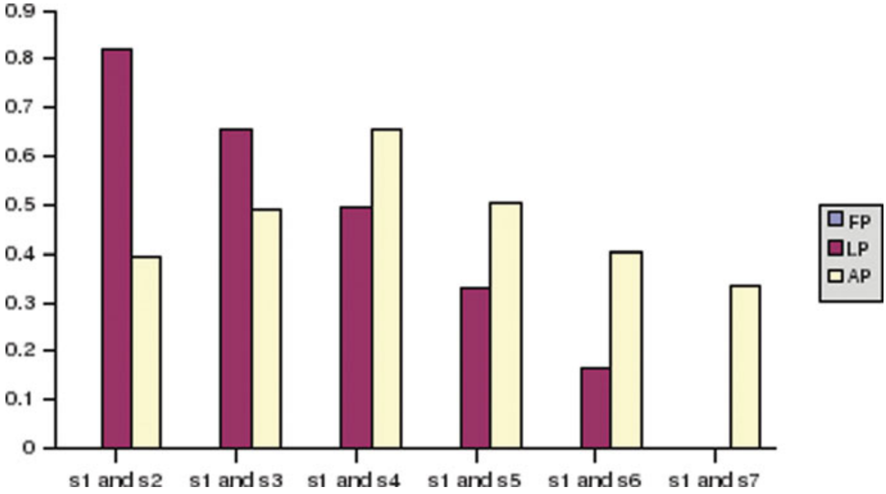

Fig. 6.3 FP, LP, and AP for a symbol occurring twice in $D1$ as the period between the two instances become larger

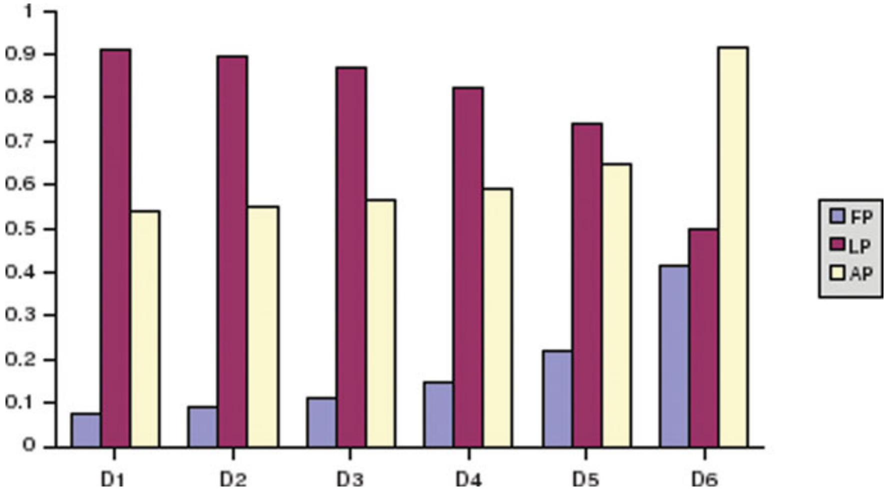

Fig. 6.4 FP, LP, and AP for a symbol occurring once in the same position relative to the beginning of different length documents

of size N), each document can be represented as a vector of dimension $3N$, where each term in the vector is the FP, LP, or AP value of each word. In our model we pre-compiled a list of words from a sample dataset (which is discarded from the test dataset after the words are collected) by aggregating a list of words that appear in 75% of all the documents in at least one genre class in the sample dataset.

The relevance of term distribution has been mentioned by others including Manning et al. (see [19]), and, more recently, by De Roeck et al. (e.g. [8]) who carried out a study of profiling datasets to determine the degree of homogeneity or heterogeneity in the distribution of frequent terms. However, there have only been few explicit implementations of the measurement for the purpose of automated classification, and most of these previous analyses have been based on a count of words in selected chunks of the texts. Term *dispersion* measured using Juilland's D coefficient (formula to be found in [22]) also depends on examining selected texts within a larger collection, for variations of standard deviation in word frequency. The model presented here, on the other hand, compares relative distances between term instances, viewing the entire document as a time dependent whole, and does not involve arbitrary choices of text chunk sizes.

6.2.3 Defining Genre

While the definition of genre may not be easily pinned down, there is a shallow agreement that genre is a concept that can be used to categorise documents by structure and function. In fact, the structural properties (e.g. the existence of a title page, chapter, section, the number of columns, use of diagrams, and font variations) evolve in ways that are designed to optimise the document's capability to fulfil its functional intention(s) (e.g. to describe, to inform and to argue, to advertise) within its target environment (e.g. the user community, publisher and creator), much the same as the structure of an organism evolves to optimise its survival function in the natural environment (cf. Kim and Ross [18]). As a consequence, genre reflects one or more of the following:

- the intention of the creator (e.g. to inform, to argue, to instruct);
- the interpretation of the user community (e.g. as a collection of facts, an expression of opinion, a piece of research);
- the prescription of a process (e.g. article for journal publication, job description for recruitment, minutes of a meeting); and
- the type of data structure (e.g. table, graph, chart, list).

The model described in Section 6.2.1, while effective in distinguishing some intentional and interpretive aspects of genre, seems insufficient to capture distinguishing features in the case of prescriptive, conceptual or physical structure. Such structure can be characterised even by low frequency terms of the class (e.g. single occurrence of “minutes” in the title of meeting minutes, or paragraph headings in a curriculum vitae), and the distributional pattern of words throughout the document (variation of density) is often bound to its class (e.g. the even distribution of wh-words in a FAQ sheet). The last observation is a generalisation of the observation by Brookstein et al. [6], who noted the clumping properties of content-bearing words and their role in text classification. In contrast to the content-bearing words that they discuss, we are interested also in words indicative of style and structure. These words can exhibit both clumping and uniform distributional properties. We present evidence that documents of each genre class display distinctive distributional characteristics and these can be more effectively captured using the HDR of documents introduced in Section 6.2.2.

A genre schema of seventy classes (KRYS I corpus) was introduced in Kim and Ross [17, 18]. The schema was constructed and populated to represent the diverse range of intentional and structural aspects of genre listed above. At the time of building the corpus, we were focusing on document genres and, therefore, did not include webpage genres. In the experiments described in this chapter, We have compensated for the deficiency by further augmenting the schema with 7 webpage genres identified within the 7-webgenre collection introduced by Santini [24]. The inclusion of the 7-webgenre collection also enables us to compare our method to other results that have been achieved on the same dataset.

6.3 Classifiers

In Section 6.6, we will compare support vector machine (SVM) classification using the harmonic descriptor representation of documents modelled using Weka machine learning software [27] against the SVM classification performed using the Bow Toolkit rainbow text classifier developed by MacCallum [20], and the classification attempts of Santini [24], to show that the performance is consistently better when using the new description. The reason we have selected SVM as the classification method is that it showed the best results for rainbow when compared with

Rocchio/TFIDF and Naive Bayes. Also it has been evidenced to be effective in other text classification tasks as demonstrated by Yang et al. [26]. The rainbow text classifier, included in the BOW toolkit [20], indexes the alpha-numeric content of the text as a bag-of-words for an analysis of significant term frequencies, while Santini's method employs a combination of linguistically motivated features. The three way comparison was motivated by a desire to make a comparison of term distribution models (e.g. HDR), term frequency models (e.g. BOW) and linguistically motivated models (e.g. [24]).

6.4 Dataset

The dataset in our experiment consists of 24 classes from KRYS I and the seven classes from the 7-webgenre collection, altogether consisting of 3,452 documents in 31 genres (see Table 6.1). The test was initially confined to 31 genres, partly, due

Table 6.1 Scope of genres

Creative	Book of Fiction(29)
	Poem(90)
Determined by user context	Email(90)
	Exam/Worksheet (90)
	Form (90)
	Handbook (90)
	Letter (91)
	Minutes (99)
	Resumé/CV (96)
	Sheet music (90)
	Speech transcript (91)
	Technical manual (90)
Determined by organisational prescription	Abstract (89)
	Academic monograph (99)
	Advertisement (90)
	Business report (100)
	Magazine article (90)
	Scientific article (90)
	Memo (90)
	Periodicals (67)
	Poster (90)
	Slides (90)
	Technical report (91)
	Thesis (100)
Webpage genres	Blog (190)
	Eshop (190)
	FAQ (190)
	Front page (190)
	List (190)
	Personal home page (190)
	Search page (190)

to some computing problems. Although clever distributed computing might have circumvented the problem observed, it was not uncommon for the support vector machine on Weka to crash due to lack of memory. This problem seemed to arise especially when many classes or number of features are introduced into the classification. Increasing the number of documents did not seem to affect the system as badly as long as the number of classes and features are moderate (e.g. experiments on a newsgroup data consisting of nearly 20,000 samples in 20 classes represented by less than 300 features did not seem to cause the same difficulty). The 24 classes from KRYS I were selected to reflect a proportion of classes from each of the ten genre groups presented in Kim and Ross [18].

A comparison of automated classification methods on a dataset that has not been tested for human agreement can give misleading information as human agreement analysis conveys to us how clean the dataset is and the nature of the genre class schema of the dataset. The experiments reported here were carried out on a collection consisting of the genres in Table 6.1 (numbers of documents in each genre, excluding those used to construct the word list in the previous section, are indicated in parentheses). The dataset for the twenty-four document genres were collected by:

1. assigning genres to collectors (in this case students) who retrieved from the Internet as many PDF files as they could find in English; and
2. having two classifiers (in this case secretaries) reclassify the PDF documents using the initial schema but without the knowledge of the initial label for each document.

None of the labellers were given a definition for the genres in the schema. This was partly to establish whether there was already a well understood genre vocabulary. The human performance was examined by taking the number of labels given by a single labeller in agreement with the other two labellers over the total number of documents on which the other two labellers agreed. The three numbers obtained in this way are 0.675, 0.73 and 0.829. Although the difference between the lowest and the highest recall is a noticeable 14%, this should be viewed with the knowledge that the highest recall is the result of student classification while the lowest recall is that of secretary classification. The human classification agreement on the KRYS I corpus has been further analysed in the research presented in Berninger et al. [4]. User studies that have been presented here and Berninger et al. [4] are speculative and far from conclusive. The results that have been presented here have been provided mainly to give context to the dataset being used in the experiments. User studies with respect to the 7 webgenre collection is found in [24].

Other human labelling analyses of genre classification from the bottom up approach (i.e. giving the users the freedom to assign and define the genres) have been carried out in Chapter 3 by Rosso and Haas (this book). Note, however, that the numbers in their work are slightly different from the numbers that have been presented here, and, in [4]: while they examine overall agreement (e.g. number of labels in agreement per document regardless of the labeller), [4] examine the agreement of selected labellers as well as overall agreement on a document.

6.5 Features

For the HDR SVM experiments reported in Section 6.6, we set aside a sample dataset consisting of ten random documents from each of the genres classes in the whole collection, and compiled all the symbols that appear in more than 75% of the documents in each genre. The symbols examined with respect to SVM HDR in the experiments reported here are simply white space delimited words in the document text,[1] inclusive of any HTML (Hyper Text Markup Language) tags. These tags are part of the vocabulary that indicates document structure and relations between entities in the HTML hybrid language, just as functional words (e.g. auxiliary verbs) might do in natural language. The compiled word list, in the current experiment, consisted of 2,477 words. Each of these words/symbols represent three features FP, LP, and AP (see Section 6.2.2) in our HDR of documents (i.e. each document is represented by a vector of dimension 7,431). The words/symbols compiled are expected to represent symbols that are prolific within at least one of the genre classes being examined (but not necessarily prolific within any one document). The list is expected to include stop words as well as HTML tags. As an illustration of the varying characteristics of vocabulary with respect to genre, we present (in Table 6.2) the number of selected word types (the range of types are indicated in the column labelled “WT”, in the table) found to be prolific (based on ten random documents from each genre) within the classes Poem, Letter and Thesis. The numbers were estimated manually by the author.

Most of the numbers in Table 6.2 are not very illuminating by itself in that the median lengths of documents belonging to Poem, Letter and Thesis are 1,718, 4,265, and 132,994, respectively (in bytes), that is, we expect the numbers to be increasing in that order for each type of word. However, we immediately notice an exception in this pattern with respect to subject pronouns, and, closer examination of the actual words show that at least one of the two subject pronouns found to be prolific in poems (i.e. “you” and “I”) is not found to be as prolific in letters (i.e. “it”) and theses (i.e. “I”, “we”, “they”, “it”). Further, the word “Dear” is only found to be prolific within letters.

To illustrate how the FP, LP and AP of the HDR description varies across documents of the same genre we present a snapshot of these values with respect to the word “whose” across 90 poems, 100 theses, 91 letters and 91 technical reports in Fig. 6.5. The segments corresponding to the documents belonging each genre are indicated at the bottom of the figure. The figure shows that FP, LP, and AP are similar for documents belonging to the same genre but diverge as we move across documents belonging to different genres.

[1] Text was extracted from the PDF using the XPDF pdftotext tool (http://www.foolabs.com/xpdf/)

Table 6.2 Number of words found in seven out of ten documents belonging to three genres (top row) with respect to word type (left column). Median length of documents in each genre are expressed in the parentheses next to the genre label as number of bytes

	Poem	Letter	Thesis
	(1,718)	(4,265)	(132,993)
Article	2	2	3
Wh-word	0	0	6
Modal	0	1	9
Have verb	0	1	3
Be verb	1	3	7
Verb	0	0	29
Noun	0	0	46
Subject pronoun	2	1	4
Object pronoun	0	0	1
Possessive pronoun	0	0	0
Possessive adjective	0	0	2
Adjective	1	1	43
Adverb	0	1	29
Quantifier	0	1	9
Demonstrative	1	2	6
Conjunction	1	3	9
Preposition	5	8	20
Punctuation	2	3	4
Other	1	1	12

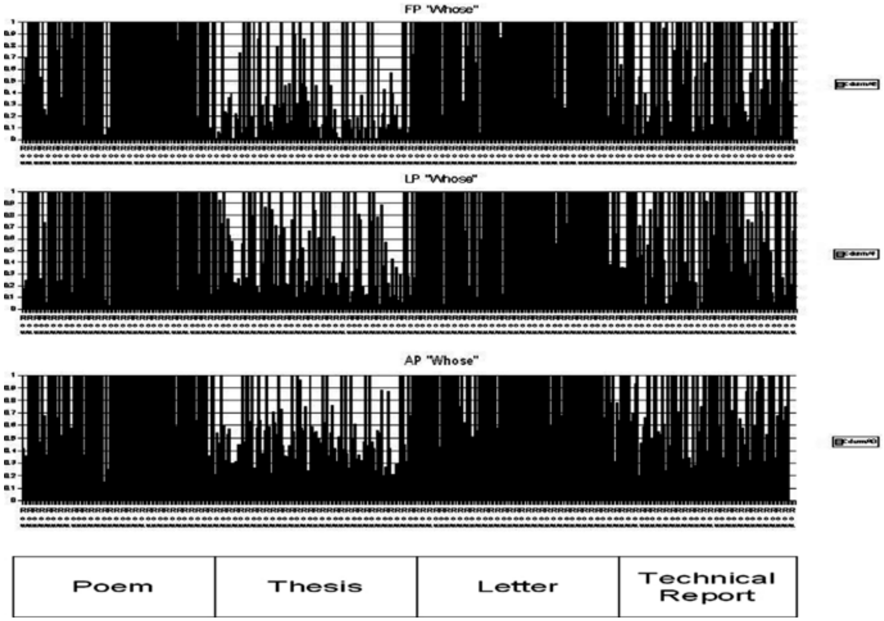

Fig. 6.5 Example of FP (*top*), LP (*middle*), and AP (*bottom*) values with respect to the word "whose" across documents belonging to four distinct genres (the documents corresponding to each of these genres are noted by segmentation indicated at the *bottom* of the figure)

6.6 Results

The performance will be evaluated using one or more of three conventional metrics: accuracy, precision and recall. To re-visit the definition for these terms, let N be the total number of documents in the test data, N_c the number of documents in the class C, $TP(C)$ the number of documents correctly predicted to be a member of class C, and $FP(C)$ the number of documents incorrectly predicted as belonging to class C. Accuracy, A, is defined to be:

$$A = \frac{\sum TP(C)}{N}, \tag{6.2}$$

precision, $P(C)$, of class C is defined to be:

$$P(C) = \frac{TP(C)}{TP(C) + FP(C)}, \tag{6.3}$$

and recall, $R(C)$, of class C is defined to be:

$$R(C) = \frac{TP(C)}{N_c}. \tag{6.4}$$

In addition we also examine the average of $P(C)$ and $R(C)$ expressed as the F-measure $F(C)$ defined as $F(C) = 2*(P(C)*R(C))/(P(C)+R(C))$. Although some debate surrounds the suitability of accuracy, precision and recall as a measurement of information retrieval tasks, for classification tasks they are still deemed to be a reasonable indicator of classifier performance.

It should also be mentioned here that all the results reported in this section are based on the average taken on ten-fold cross validation.

6.6.1 *Overall Accuracy*

The figures in Table 6.3 are the overall accuracies of the support vector machine rainbow classifier (SVM rainbow), the support vector HDR classifier (SVM HDR), and the average human agreement estimated by assuming that human agreement on the 7-webgenre collection is perfect. The classifier we are considering to be a baseline classifier in this comparison is the SVM rainbow classifier. The human agreement is included to indicate the cleanliness level of the dataset being used.

Table 6.3 Overall accuracy across all 31 genre classes

Classifier	SVM rainbow	SVM HDR	Human avg
Overall accuracy	0.73	0.80	0.84[a]

[a] Estimated assuming agreement is perfect on the 7-webgenre collection.

The numbers in Table 6.3 suggest that the performance level of the SVM rainbow classifier is already comparable to the average performance of three human labellers, and shows that the SVM HDR improves on the SVM rainbow classifier by 7%.

To test the limits on a cleaner dataset, we analysed the classification results with respect to the 7-webgenre collection. This is the overall accuracy of the classification when the recall of the documents belonging to the webpage genre classes is calculated upon the classification of the entire dataset into 31 classes. There is a slight increase of 0.002 when the webpage classes are classified on their own. The results are shown in Table 6.4: the numbers suggest that SVM HDR is a strong contender in webpage genre classification.

Table 6.4 Overall accuracy of classifiers across webpage genres (Blog, Personal Home Page, FAQ, List, Search Page, EShop, Front Page)

Classifier	SVM rainbow	Santini's result	SVM HDR
Accuracy	0.92	0.89	0.96

6.6.2 Precision and Recall

The challenge in document classification is to improve the overall accuracy of the classification without compromising the performance with respect to any one class in the schema. In this section we will show that SVM HDR meets this challenge.

In Figs. 6.6 and 6.7, we present the recall and precision of SVM rainbow and SVM HDR with respect to each of our classes. The graphs show that SVM HDR outperforms SVM rainbow with respect to most of the classes in both recall and precision. The recall of SVM rainbow with respect to Academic Monograph, Book of Fiction, Front Page (of a website), Minutes, periodicals, Technical Manual and Thesis is Marginally higher than SVM HDR and the precision of SVM rainbow with respect to Abstract, Exam/Worksheet, Home Page, Poem, and Slides is somewhat higher than that of SVM HDR. However, with respect to the majority of the classes, SVM HDR outperforms SVM rainbow.

The graphs also demonstrates that SVM rainbow's performance varies widely across different genres, while the deviation of performance is much more confined in the case of SVM HDR. The recall (resp. precision) of SVM rainbow ranges from 0.08 to 1 (resp. 0.24–0.99), while recall (resp. precision) of SVM HDR ranges from 0.42 to 1 (resp. 0.38–0.99). The difference between precision and recall with respect to each class is also notable: the maximum absolute difference between precision and recall across the genre classes for SVM HDR is observed at approximately 0.24, while the same for SVM rainbow is observed at 0.46. The small deviation of performance across classes and the comparability of precision and recall with respect to each class seems to suggest that HDR is more successful in characterising the genre classes.

The graph in Fig. 6.8 presents the F-measures of SVM rainbow and SVM HDR with respect to each class. This graph shows that the F-measures of SVM HDR

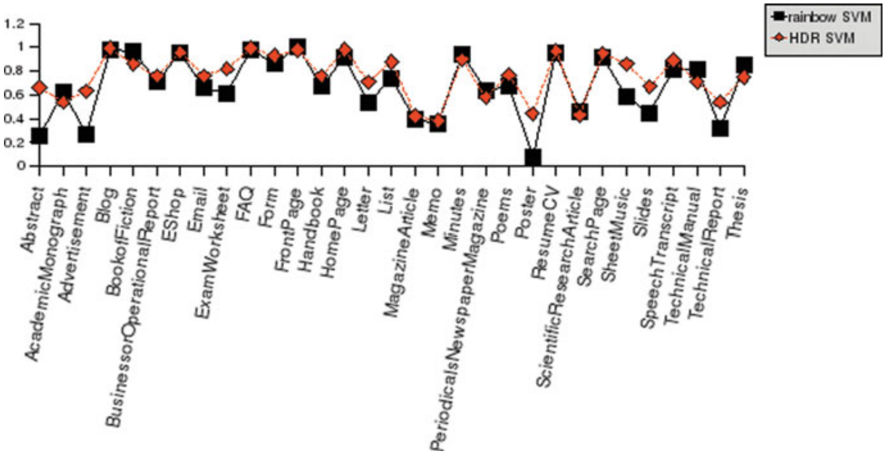

Fig. 6.6 Recall: a comparison, SVM rainbow and SVM HDR

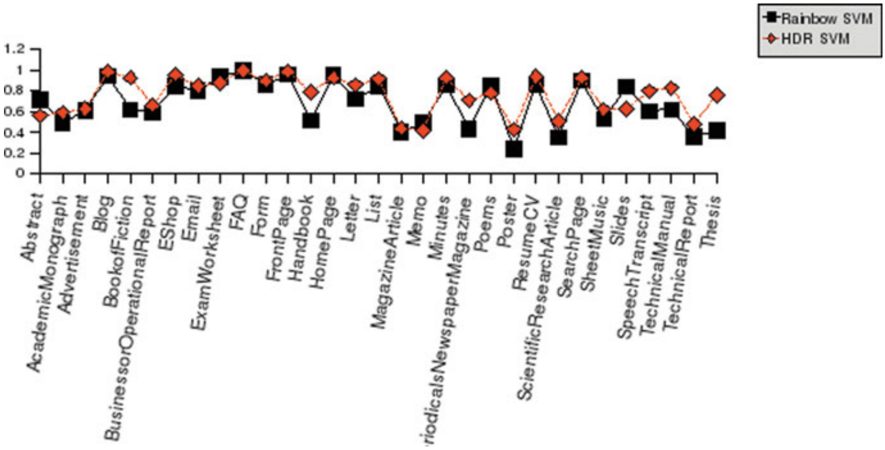

Fig. 6.7 Precision: a comparison, SVM rainbow and SVM HDR

are greater than those of SVM rainbow with respect to every class except the class Memo. With respect to Memo, the difference is 0.02 in favour of SVM rainbow. Latest experiments using HDR to analyse a newsgroup dataset of 19,597 documents in 20 topical classes (obtained from McCallum's website1), show that the same SVM HDR model is also promising in topic classification, with an overall accuracy of over 95% (detailed report of this experiment available shortly). A list of 82 words was compiled from 400 documents (20 documents from each genre) set aside from the original 19,997 documents for this experiment. We have also calculated the F-measures of SVM HDR with respect to the classes in this dataset to find them all greater than the best results (overall accuracy 93.7%) of the rainbow classifier. The details of this experiment will be published shortly.

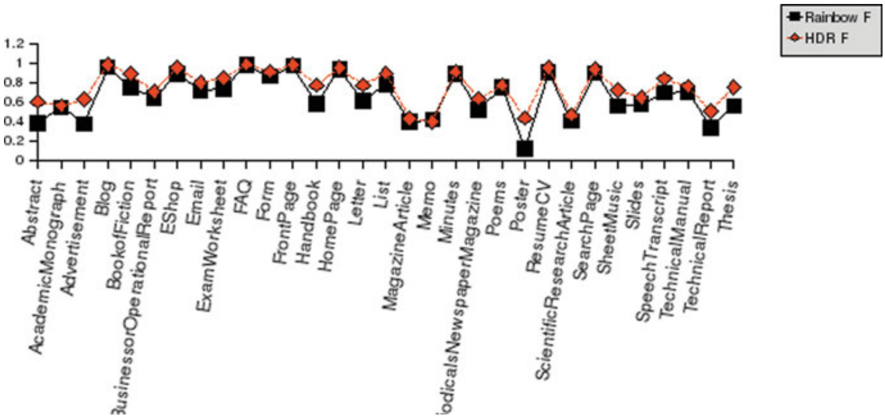

Fig. 6.8 F-measure: a comparison, SVM rainbow and SVM HDR

In the HDR of documents we have presented here, we have measured FP, LP and APR with respect to the length of the whole document. Just as performing discrete Fourier transform to obtain the harmonics of waves in signal processes involves sampling the signal, documents can also be examined at different resolutions by varying the range in which harmonic behaviour is examined (e.g. when examining the string “axbxcxdefghixjklmn”, and examining the occurrences of “x” throughout the string, it does not seem to exhibit harmonic behaviour but, if you select the first seven letters “axbxcxd”, it is perfectly harmonic). It is likely that shorter windows of examination will produce interesting comparisons.

6.7 Conclusions

The results of automated experiments described in this chapter provide evidence that the overall accuracy of the support vector machine rainbow text classifier is already comparable to that of an average human classifier in genre classification. Here we have shown that the SVM HDR, which uses the layout of words in the document, outperforms the SVM rainbow text classifier. This makes it a promising candidate for further study. In particular, a comparison of the SVM HDR classifier against classifiers other than SVM rainbow is required for fuller analysis. It would also be desirable to make direct comparisons of LP, FP and AP across genre classes.

The results with respect to the 7-webgenre collection suggest SVM HDR as a promising candidate for comparison to classifiers that rely on counts of terms or patterns. There have been reports of high accuracy levels of classification on the same dataset carried out by Kanaris and Stamatatos [13]. Although their numbers are similar to ours, it must be noted that the accuracy presented by them is from classifications of the set carried out in isolation while, the accuracy reported in this chapter is obtained from a classification of the seven webpage genres when accompanied by a classification of 24 additional document genres.

Previous text classification methods actively integrate mathematical methods in feature selection, statistical modelling and error analysis, but the concept we are trying to capture is still only described through examples in the domain. This leads to a semantic gap (especially with high-level concepts such as those represented by genre classes) not dissimilar to that encountered in image retrieval.

A more rigorous study of genre is required to reflect two considerations: first, we need to scope different communities for potentially useful genre classes that can support other applications and, second, we need to incorporate basic mathematical concepts into the actual description of the identified genres. Hence, future efforts in this field should not only study the implication of term distribution versus term frequency further by:

- examining the resolution mentioned at the end of Section 6.6.2;
- looking at, and comparing, other forms of symbols apart from words; and
- considering ways in which the two approaches might be integrated

but also include user studies of genres to identify the possible applications to guide genre classification work, and isolate base mathematical concepts that can be used to build the concepts gradually to describe higher-level concepts of genre.

References

1. Bagdanov, A., and M. Worring. 2001. Fine-grained document genre classification using first order random graphs. In *Proceedings of the 2001 Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, 79–90. Seattle, WA. USA. http://doi.ieeecomputersociety.org/10.1109/ICDAR.2001.953759
2. Barbu, E., P. Heroux, S. Adam, and E. Turpin. 2005. Clustering document images using a bag of symbols representation. In *Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)*, 1216–1220. Seoul, Korea. http://doi.ieeecomputersociety.org/10.1109/ICDAR.2001.953759
3. Bekkerman, R., A. McCallum, and G. Huang. 2004. Automatic categorization of email into folders: Benchmark experiments on Enron and Sri corpora. Technical Report IR-418, Center for Intelligent Information Retrieval, UMASS. http://www.cs.umass.edu/\homedirmccallum/papers/foldering-tr05.pdf
4. Berninger, V.F., Y. Kim, and S. Ross. 2009. Building a document genre corpus: A profile of the KRYS I corpus. In *Proceedings of Corpus Profiling Workshop with 'BCS-IRSG Workshop on Corpus Profiling'*. http://www.bcs.org/server.php?show=conWebDoc.26115
5. Biber, D. 1995. *Dimensions of register variation: a cross-linguistic comparison*. New York, NY: Cambridge University Press.
6. Bookstein, A., S.T. Klein, and T. Raita. 1998. Clumping properties of content-bearing words. *Journal of the American Society of Information Science* 49(2):102–114.
7. Dc-dot: UKOLN Dublin Core Metadata Editor (webpage last updated Aug 2000). http://www.ukoln.ac.uk/metadata/dcdot/
8. De Roeck, A., A. Sarkar, and P. Garthwaite. 2004. Frequent term distribution measures for dataset profiling. Technical Report 2004/2006, Faculty of Mathematics and Computing, Open University. Milton Keynes. http://computing-reports.open.ac.uk/index.php/
9. Dong, L., C. Watters, J. Duffy, and M. Shepherd. 2008. An examination of genre attributes for web page classification. In *Proceedings 41st Hawaiian International Conference*

on System Sciences. IEEE Computer Society Press. Waikoloa, Big Island, HI, USA. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4438836

10. Finn, A., and N. Kushmerick. 2006. Learning to classify documents according to genre. *Journal of American Society for Information Science and Technology* 57(11):1506–1518.
11. Giuffrida, G., E. Shek, and J. Yang. 2000. Knowledge-based metadata extraction from PostScript Files. In *Proceedings 5th ACM International Conference on Digital Libraries*, 77–84. San Antonio, TX, USA. http://portal.acm.org/citation.cfm?id=336597.336639
12. Han, H., L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E.A. Fox. 2003. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 77–84. Houston, TX, USA. http://portal.acm.org/citation.cfm?id=827146
13. Kanaris, I., and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings 19th IEEE International Conference on Tools with Artificial Intelligence*. Patras, GR. http://portal.acm.org/citation.cfm?id=1337285
14. Karlgren, J., and D. Cutting. 1994. Recognizing text genres with simple metric using discriminant analysis. In *Proceedings 15th Conference on Computational Linguistics* 2:1071–1075. Kyoto, Japan.
15. Ke, S.W., C. Bowerman, and M. Oakes. 2006. PERC: A personal email classifier. In *ECIR 2006* (London, UK), eds. M. Lalmas et al., LNCS 3936, 460–463, Heidelberg: Springer-Verlag, http://www.springerlink.com/content/r27700t736786455/fulltext.pdf
16. Kessler, G., B. Nunberg, and H. Schuetze. 1997. Automatic detection of text genre. In *Proceedings 35th Annual Meeting ACL*, 32–38. Madrid, Spain.
17. Kim, Y., and S. Ross. 2007a. Detecting family resemblance: Automated genre classification. *CODATA Data Science Journal* 6:S172–S183. ISSN: 1683–1470. http://www.jstage.jst.go.jp/article/dsj/6/0/S172/_pdf
18. Kim, Y., and S. Ross. 2007b. Searching for ground truth: A stepping stone in automated genre classification. *In Digital libraries: R&D* (Tirrenia, Italy), eds. C. Thanos, F. Borri, and L. Candela, LNCS 4877, 248–261, Heidelberg: Springer-Verlag, http://www.springerlink.com/content/lt760613m2731723
19. Manning, C., and H. Schutze. 1999. *Foundations of statistical language processing*. Cambridge, MA: MIT Press.
20. McCallum, A. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/\verb1∼11mccallum/bow
21. Rauber, A., and A. Müller-Kögler. 2001. Integrating automatic genre analysis into digital libraries. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. Roanoke, VA, USA. http://doi.acm.org/10.1145/379437.379439
22. Rayson, P., A. Wilson, and G. Leech. 2002. Grammatical word class variation within the British National Corpus sampler. In *New frontiers of corpus research: Papers from the 21st International Conference on English Language Research on Computerized Corpora*, Sydney 2000, eds. P. Peters, P. Collins, and A. Smith, 295–306. Amsterdam: Rodopi.
23. Ross, S., and M. Hedstrom. 2005. Preservation research and sustainable digital libraries. *International Journal of Digital Libraries* 5(4):317–325.
24. Santini, M. 2007. Automatic identification of genre in web pages. PhD Thesis, University of Brighton, Brighton. http://www.itri.brighton.ac.uk/\homedirMarina.Santini/MSantini_PhD_Thesis.zip
25. Thoma, G. 2001. Automating the production of bibliographic records. Technical report, Lister Hill National Center for Biomedical Communication, US National Library of Medicine. http://archive.nlm.nih.gov/pubs/thoma/mars2001.php
26. Yang, Y., J. Zhang, and B. Kisiel. 2003. A scalability analysis of classifiers in text categorization. In *Proceedings 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 96–103. Toronto, ON, CA. http://doi.acm.org/10.1145/860435.860455
27. Witten, H.I., and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann.