Variation of Word Frequencies across Genre Classification Tasks

Yunhyong Kim and Seamus Ross
Digital Curation Centre (DCC)
&
Humanities Advanced Technology Information Institute (HATII)
University of Glasgow
email: {y.kim, s.ross}@hatii.arts.gla.ac.uk

Abstract

This paper examines automated genre classification of text documents and its role in enabling the effective management of digital documents by digital libraries and other repositories. Genre classification, which narrows down the possible structure of a document, is a valuable step in realising the general automatic extraction of semantic metadata essential to the efficient management and use of digital objects. In the present report, we present an analysis of word frequencies in different genre classes in an effort to understand the distinction between independent classification tasks. In particular, we examine automated experiments on thirty-one genre classes to determine the relationship between the word frequency metrics and the degree of its significance in carrying out classification in varying environments.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Management, Design, Performance, Experimentation

Keywords

genre classification; metadata; digital library management; information extraction.

1 Introduction

The volume of digital resources as a common form of information exchange in our everyday life is growing at an exponential rate. As a consequence, the task of storing, managing and utilising this information has become increasingly demanding during recent years. Metadata embodying the core technical requirements, administrative function and content description of an object play a crucial role in the efficient and effective management and use of materials in digital repositories (cf. Ross and Hedstrom, 2005). The manual collection of such information is costly and labour-intensive and a collaborative effort to automate the extraction of such information has become an immediate concern.[1]

There have been several efforts (e.g. Giuffrida, Shek & Yang, 2000; Han *et al*., 2003; Thoma, 2001; dc-dot metadata editor;[2] Bekkerman, McCallum & Huang, 2004; Ke & Bowerman, 2006) to extract relevant metadata from selected genres (e.g. scientific articles, web pages and emails). These efforts often rely on structural elements found to be common among documents belonging to the genre. The structural properties that characterise the genre class, which evolve to accommodate the effective performance of its function in the target community or process, help to predict the region and style in which other metadata may appear. Inspired by this fact, we have undertaken to construct a prototype tool for automated genre classification as a first step to further metadata extraction. The prototype is expected not only to aid metadata extraction as an overarching tool that

[1] For example, the Cedar Project at the University of Leeds: http://www.leeds.ac.uk/cedars/guideto/collmanagement/guidetocolman.pdf

[2] dc-dot, UKOLN Dublin Core Metadata Editor, http://www.ukoln.ac.uk/metadata/dcdot/

binds genre-dependent tools, but also to support the selection, acquisition and search of material in terms of style, depth of content and functional intent.

A diverse range of notions are discussed under the single umbrella of genre classification, including Biber's text typology into five dimensions (Biber, 1995), the examination of popularly recognised document and web page genres (Karlgren & Cutting, 1994; Boese, 2005; Santini 2007), and the consideration of genre categoric aspects of text such as objectivity, intended level of audience, positive or negative opinion and whether it is a narrative (Kessler, Nünberg & Schütze, 1997; Finn & Kushmerick., 2006). Some have investigated the categorisation of documents into a selected number of journals and brochures (Bagdanov & Worring, 2001), while others (Rauber & Müller-Kögler, 2001, Barbu *et al*., 2005) have clustered documents into similar feature groups without assigning genre labels. Despite the variety of characterisations under examination, the previously introduced notions of document genre can be summarised as compromising one or more of the following:

- the *semantic category* of a document conveying the creator's intention, or the interpretation of a user community;
- the *data structure type* of a document as a vehicle for expressing information;
- the *functional categor*y of a document as *part of a process* such as publication, recruitment, event or use.

These aspects can be integrated under a single notion of entities and relations between entities. On a coarse level of granularity, documents themselves are entities defined by their relations to other extra-document entities (e.g. processes, agents, users, and other documents). On a finer level of granularity, documents are defined by intra-document entities such as paragraphs, sections, sentences, text blocks, images, tables, citations, links, phrases, words, terms, symbols, pixels and characters, and their relations to each other and extra-document entities. The semantic category of a document is approximated by a conceptual equivalence model of entity relations largely consisting of, but not confined to, intra-document relations (relations between intra-document entities), and the functional category of the document is approximated by entity relations largely consisting of, but not confined to, extra-document relations (between intra-document entities and extra-document entities). The data structure type of a document relates exclusively to a selected group of intra-document entities and intra-document relations. The intra-document entities and relations that define the data structure type are selected to best serve the functional category of the document given the semantic category of the document; thus the data structural type reflects and is closely linked to both the semantic category and the functional category of a document. In this light, successful data structural type detection would lead to a wealth of information about the other two categorial identities of the document.

Automated recognition of document data structural type requires a variety of document-understanding techniques to annotate relations between two or more intra-document entities (e.g. parsing and the detection of co-referent terms). The use of automated tools for such annotation is domain dependent and error prone, i.e. would result in error propagation. Even well-tested part-of-speech taggers and parsers are domain dependent and can not be trusted to perform well across untested genres. Hence, initially, we have opted to limit ourselves to identifying the entities and relations on a reasonably crude level of sophistication. The process of adding additional layers of grammatical analysis, phrasal stylistics and co-reference resolution will be left to the next stage of the exercise. Examples of low-level entities include analyses of the words and their frequencies in the document, and analyses of white and dark pixels and their frequencies in the document. In previous papers (e.g. Kim & Ross, 2007), we have tried to compare the role played by these low-level features modelled using three statistical methods (Naïve Bayes, Support Vector Machine and Random Forest) to establish a relationship between genre classes and feature strengths on selected genres. In these papers, the simple use of word frequency emerged as a strong feature in genre classification. In the current paper we would like to present a more comprehensive analysis to examine the variation of word frequencies across genres and the performance of classifiers incorporating this feature in order to establish a relationship between classification tasks and word statistics. We will investigate this with respect to thirty-one genre classes (Table 1.1) comprising twenty-four classes constructed as general document genres and seven classes constructed as web page genres.

Table 1.1. Scope of genres under examination

Genre group	**Genre**		
Book	Academic Monograph	Book of Fiction	Handbook
Article	Abstract	Magazine Article	Scientific Article
Short Composition	Poem		
Serial	Periodicals (Newspaper, Magazine)		
Correspondence	Email	Letter	Memo
Treatise	Business/Operational Report Technical Manual	Thesis	Technical Report
Information Structure	Form		
Evidential Document	Minutes		
Visually Dominant Document	Poster	Sheet Music	
Webpage	Blog Front Page Search Page	E-Shop Home Page	FAQ List
Other Functional Document	Advertisement Speech Transcript	Exam/Worksheet Slides	Curriculum Vitae

There are other studies that have focused on word frequency analysis for the purpose of genre classification (e.g. Stamatatos, Fakotakis & Kokkinakis, 2000). These, however, concentrate on common words in the English language to model stop word statistics, or employ standard significant word detection such as those that have been frequently used in subject classification of documents. We, on the other hand, want to examine words that appear in a large proportion of documents in a genre without necessarily having a high frequency within each document or within the entire corpus.

There have also been studies that incorporate high-level linguistic analysis to model genre characterising facets (e.g. Santini, 2007) exhibited by documents with some success. This leads to the question why one would still invest time on models using word frequency only. A prominent reason for doing this is that models that already integrate involved linguistic information are heavily language dependent, and likely to require significant internal change to accommodate other languages. We will discuss the potential of a refined frequency model, which may effectively approximate higher-level concepts without being heavily dependent on the exact syntax of the language.

The study here is not an effort to present an optimised automated genre classification tool. The objective is to show that a simple word frequency model is moderately effective across a wide variety of genres (even without the incorporation of further syntactic analysis), that the level of efficacy is heavily dependent on the scope of genre classes under examination, and to suggest reasons for the failure where the method fails.

2 Corpora

It is a well-recognised fact that there is a lack of consolidated data for the study of automated genre classification; there is no standardised genre schema, and those contexts where genre classification arises as a useful tool require very different approaches to genres. At this stage of establishing consolidated data, it seems important to scope for genres in as many different contexts as possible in order to determine genres leading to useful applications. With this in mind, KRYS I has been constructed to encompass a schema of seventy genres of varied types. The corpus, at the time, was not constructed to reflect web page genre classes. To compensate, we have augmented our experimental data with documents from the Santini Web corpus.

KRYS I:

The corpus was assembled through a document retrieval exercise, where university students were assigned one of seventy genres and, for each genre, asked to retrieve from the Internet documents that they believed to be an example of that genre, represented in PDF, and written in English. They were not given any description of the genres apart from the genre label. They were asked to describe their reasons for including the particular example in the set. Initially we aimed to collect one hundred examples for each genre, but examples of some genres were hard to find and students were not able to retrieve one hundred examples. The resulting corpus now includes 6,478 items. Collected documents were reclassified by two secretaries. The secretaries were not allowed to confer and the documents, without their original label, were presented in a random order from the database to each labeller. The secretaries were not given descriptions of genres. They were expected to use their own training in record-keeping to classify the documents. Not all the documents collected in the retrieval exercise have been re-classified by both secretaries. There are a total of 5,305 documents stored with three labels.

SANTINI Web:

This corpus consists of 1,400 web pages labelled as belonging to one of seven web page categories. There are 200 documents in each of the classified seven categories. The seven categories include Blog, FAQ, Front Page, Search Page, Home Page, List and E-Shop. These datasets are available from Santini's home page[3] and discussed in (Santini, 2007).

3 Human agreement

The figures in Table 3.1 show the number of documents on which different groups of labellers have agreed.

Table 3.1. Human agreement analysis

Labeller group	Agreed
student & secretary I	2,745*
student & secretary II	2,852*
secretary I & II	2,422*
all three labellers	2,008*

*out of 5,305

A large number of documents were erroneously submitted by students to the KRYS I corpus. For instance, there were:

- ●documents that are not examples of the genre but whose topic relates to the genre (e.g. instead of actual emails, research articles about email were found labelled as email) [Error type I];
- ●empty templates included as examples of the genre (e.g. instead of selecting 'actual' receipts, empty receipt forms were found labelled as receipts) [Error type II];
- ●entire magazines, conference proceedings or journals included as research articles, and vice versa [Error type III].

These items have not been removed from the corpus but have been marked for later analysis. By considering only the documents that were given the same label by at least two of the labellers (Secretary Student Agreement Data - SSAD), we expect to eliminate almost all of these errors. There are 1,523 documents in SSAD that have been labelled as belonging to genres of Table 1.1. Of these, 795 documents have been given the same label by all the labellers. In the current investigation, we are not so much interested in the overall accuracy of human performance as the difference in agreement across independent labellers and what this signifies for the genre schema and the genre classes in question. To this end, we have taken the data labelled as belonging to the same class in Table 1.1 by one of the secretaries and the students, and compared this with the labels of the other secretary, giving us two secretary performances on SSAD. We have divided the twenty-four classes into twenty groups (Table 3.2) according to two metrics: on the topmost row of

[3] http://www.nltg.brighton.ac.uk/home/Marina.Santini/

Table3.2 we have indicated the discrepancy between the two agreement levels, and down the left column we have indicated the average agreement level by intervals.

Table 3.2. Partition of documents according to human labelling agreement

Avg. Agreement	0-0.1	0.1-0.2	0.2 - 0.3	0.3+
0.90+	Minutes Handbook CV Sheet Music	-	-	-
0.80-0.90	Exam Worksheet	Speech Transcript	Email	-
0.70-0.80	Poem Form	-	Thesis Letter Technical Report	Book of Fiction
0.50-0.70	Periodicals	-	Memo	Slides
0-0.50	Advertisement Academic Monograph Magazine Articl	-	Business Report Scientific Article	Abstract Technical Manual Poster

The partition displayed in Table 3.2 introduces a notion of genre classes that are less context dependent (classes in darker boxes) and those that are highly context dependent. That is, the genre classes with high average agreement and small discrepancy (e.g. Handbook, Minutes, CV, Sheet Music) are expected to be less dependent on labeller background and domain definitions. The classes less bound to context, unsurprisingly, seem to be those that consist of documents distinguished by the common use of genre label in the title of the document (e.g. “meeting minutes” and “curriculum vitae”), controlled vocabulary (e.g. the phrase “positions held” in CVs) and special symbols (musical notation in Sheet Music).

4 Experimental data and evaluation method

4.1. Experimental data

Our experimental datasets consist of documents belonging to one of thirty-one genres listed in Table 1.1, from Krys I* and Santini Web. The first dataset (Dataset I) consists of ten random documents from each of the thirty-one genre classes, and the other dataset (Dataset II) consists of the remaining documents from the same classes. Dataset I is used to generate the Prolific Words List in Genres (ProWLinG) described in Section 5. The number of items in Dataset II, by class, is indicated below:

Dataset II

Article [Abstract (89), Magazine Article (90), Scientific Research Article (90)]
Book [Academic Monograph (99), Book of Fiction (29), Handbook (90)]
Correspondence [Email (90), Letter (91), Memo (90)]
Evidential Document [Minutes (99)]
Information Structure [Form (90)]
Serial [Periodicals (Magazine and Newspaper) (67)]
Treatise [Business Report (100), Technical Report (90), Technical Manual (90), Thesis (100)]
Visually Dominant Document [Sheet Music (90), Poster (90)]
Web page [Blog (190), FAQ (190), Front Page (190), Search Page (190), Home Page (190), List (190), E-Shop (190)]
Other Functional Document [Slides (90), Speech Transcript (91), Poems (90), Curriculum Vitae (96), Advertisement (90), Exam/Worksheet (90)]

* The twenty-four genre classes from KRYS I have been chosen to minimise error types I, II and III. The agreement data in SSAD did not contain enough documents in certain genres (e.g. only nine documents belonging to Academic Monograph) for a reliable 10-fold cross-validation experiment.

The other genre classes in KRYS I have not been examined in this paper: the current study is aimed at establishing a relationship between genre classes and word frequency statistics. Increasing the number of genres without an increase of data in each class will introduce confusion and computation time, without lending more credibility to the statistics with respect to individual classes.

4.2. Evaluation method

The experimental results in this paper, for both human labellers and the automated system, are evaluated using three conventional measures: accuracy, precision and recall. 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 $A = \{\sum_c TP(C)\}/N$, precision, $P(C)$, of class C is defined to be $P(C) = TP(C)/\{TP(C) + FP(C)\}$, and recall, $R(C)$, of class C is defined to be $R(C) = TP(C)/N_c$. Although 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. In section 6, we have also provided some confusion analysis.

5 Automated experiments

The automated experiments use three classifiers. The classifiers are defined by one of three statistical methods and the ProWLinG word frequency features, which define the document representation. The three statistical methods are Naïve Bayes (NB) [cf. Minsky, 1961], Support Vector Machine (SVM) [cf. Burges, 1998] and Random Forest (RF) [Breiman, 2001], all available as part of the Weka Machine learning Toolkit (Witten & Frank, 2005). A lack of controlled experiments to determine the best method for genre classification makes the selection of statistical methods problematic. We have chosen three very different statistical methods for wide coverage: Naïve Bayes, reliant on basic principles of Bayes theorem; Support Vector Machine, dependent on separating data by hyperplane and effective in other document classification tasks [Yang, 2003], and Random Forest, based on a poll of votes cast by several decision trees built on random selection of features.

Prolific Words List in Genres (ProWLinG)

A word list is constructed by taking words that commonly appear across the thirty-one genres plus the unclassified documents of the SPIRIT data in the Santini Web corpus. Dataset I, consisting of ten documents from all thirty-one genres plus fifty documents from the pool of unclassified SPIRIT data of the Santini Web corpus, was set aside. The algorithm checks the files in each genre class and compiles all the words within the genre; it then counts the number of files in which each word is found. The final word list is constructed by taking the union of all the words found in 75 per cent or more of all the files in each genre. At this stage we do not consider the frequency of words within each file. This method collects words that have a high document count in one or more genre classes. A larger number of documents are sampled from SPIRIT because it is expected to contain documents of many different genres: by including a larger set, we hope that collecting words that appear in a high percentage of the included documents implies a high probability that the words would appear in a good number of the documents in any one genre included in the unclassified class.

There was only one word that was prolific in all of the thirty-two genre classes. The low number of common prolific words across genres is partly due to errors during the conversion of PDF documents to text. This is another reason for using words that are prolific within any number of genres. In this way the system is less likely to fail when the text extraction fails for any specific document version. The total number of words in ProWLinG is 2,476. The size of Dataset I was tested at different levels, and the document count was tested at lower percentages; the final settings were chosen because they showed better performance consistently across datasets. We have also tried varying the number of classes represented in Dataset I to see if the classification improves when the classes are focused to include only the classes considered in any one classification. However, we found that the classification was almost always better when the number of classes in the word list propagation stage is greater than the actual genres being considered.

The ProWLinG is intended to capture words commonly used in documents of all genres. The notion is similar to other significant term analyses in that the intention is to capture words that appear frequently in each genre as words relevant to the classes under consideration. However, unlike previous word lists, we are not concerned with the frequency of the word in the document at the gathering stage and no inverse class frequency is considered. Another distinguishing feature is the use of two disjoint training datasets: one for the creation of a word list to be used as a basis for frequency distributions, and another for the probabilistic modelling of classes as represented by word frequencies. This prevents the model from over-fitting either training dataset. The models were tested using the standard 10-fold cross-validation.

Document representation

Each document in the dataset is represented as a vector, where each entry corresponds to the frequency of a word in ProWLinG. We have experimented with two different ways to express the frequencies. In one representation the *absolute frequency* is indicated, while, in the other, we divide the frequency by the frequency of the most popular word in the file so that each entry is a *relative frequency* with respect to the most frequent word in the file.

The word frequency representation via a pre-constructed list of words from a separate dataset:

- lessens the ill-effects of over-fitting the training data;
- has the potential to express genre style:
  1. indicated by terms which function as structural cues by their presence (e.g. “minutes” in the title of meeting minutes) [**presence**];
  2. indicated by selected term count within the document (e.g. verbs “examine”, “investigate” and “show” are expected to have a higher frequency than “love”, “hate” and “goad” in scientific articles)[**count**];
  3. indicated by ratio between distinct terms of the same functional category (e.g. distribution of determiner and pronouns – see Figure 5.1 and 5.2) [**ratio**];
  4. indicated by term distribution throughout the document (e.g. regularly spaced words such as “what” or “how” in a FAQ sheet) [**density**];
- can be used to express the above four notions with respect to discrete strings, functional groups of terms (whether linguistic or otherwise) and conceptual groups of terms.

Figure 5.2. Pronoun ratio across three genres

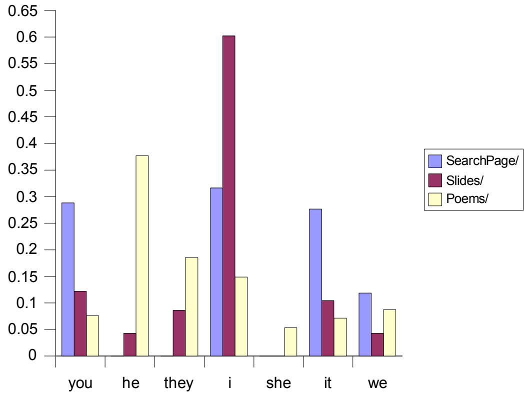
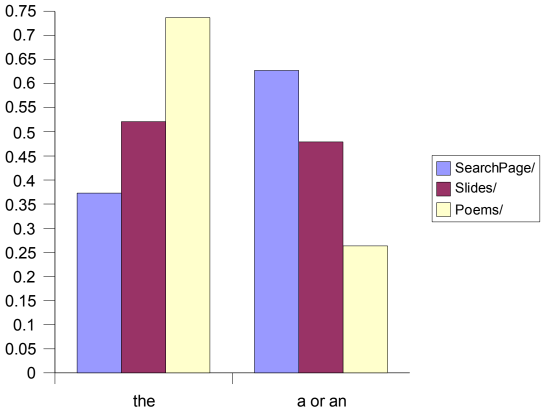


Figure 5.1. Article ratio across three genres



Absolute frequencies of words are strongly influenced by the length of the document. This can be controlled to some extent by truncating the document, but truncation would influence the distributional characteristics of a document such as the ratio mentioned in item 2 above, and also affect statistics that depend on the position within the document being examined such as items 1 and 4. By predetermining a selection of genre-related words within a set of complete documents and examining the *relative frequency* of these words, we hope to address this problem. Here, we are focusing on aspects 2 and 3 with respect to words as discrete strings within the document. Later we hope to incorporate the full spectra of the four aspects (1, 2, 3 and 4) on the level of strings, functional category and conceptual congruence.

6 Results

6.1. Overall accuracy

The overall accuracies of ProWLinG frequencies modelled using three statistical methods on the entire Dataset II are displayed in Table 6.1 (10-fold validation). The top two rows of the table specify the classifier being tested. The overall accuracy in Table 6.1 does not compare to the best accuracies in genre classification that have been reported elsewhere. However, other studies have mostly examined limited genre schema of approximately ten genres. Classifications across as many as thirty-one classes have not been tested. In fact, the absolute frequency ProWLinG Random Forest displays an overall accuracy of 0.927 when tested only on the Santini Web data (cf. Santini, 2007).

Table 6.1. Overall accuracy on twenty-four genres using three statistical methods and two ProWLinG frequency metrics

Statistics	**NB**	**NB**	**SVM**	**SVM**	**RF**	**RF**
Frequency	**absolute**	**relative**	**absolute**	**relative**	**absolute**	**relative**
Accuracy	0.504	0.642	0.503	0.659	0.739	0.749

In any case, the focus of this paper is on the relationship between ProWLing and distinct genre classes. That is, even though the overall accuracy for the Random Forest ProWLing model is 0.739, the figure is not representative of all the classes in the schema. In the next section, we will discuss results for individual classes.

6.2. Precision and Recall

The figures in Table 6.2 show the precision and recall of the ProWLing Random Forest classifier with respect to genre classes in Dataset II that have shown F-measure[4] less than 0.5 on the basis of both relative or absolute frequency. The numbers in Table 6.3, similarly, show the precision and recall of the classifier with respect to classes showing F-measure greater than or equal to 0.7.

The selection of best precision and recall includes six of the seven web page genre classes represented in Santini's dataset and three of the four classes (Curriculum Vitae, Minutes, Handbook, Sheet Music) with respect to which human classifiers have shown the highest levels of agreement. The performance with respect to Sheet Music was considerably behind the human agreement level. The classifier's recall with respect to Sheet Music was fair at 0.8 but the precision was low due to a moderate amount of confusion between Sheet Music and Poem. The automated classifier's performance on Book of Fiction and Technical Manual actually surpasses human performance. The worst performances are displayed with respect to classes identified in Section 4 as being highly dependent on labeller background and/or domain context.

Note that the relative frequency representation does not increase performance with respect to all the classes in Tables 6.2 and 6.3. However, the difference in performance is so slight that it does not seem reasonable to make conclusive remarks.

The most frequent misclassifications have been indicated as confusion cluster groups (Table 6.4). At first, the inclusion of Advertisements in Cluster group 1 might seem surprising, but further thought suggests this as being reasonable: documents of the classes in the group are all descriptions of an activity, research or product with the intention of promoting the content. The documents are also fairly short, i.e. the probability of finding words from ProWLinG in the documents of these genres will be lower than in documents of other genres. It is likely that ProWLinG does not contain

[4] F-measure = 2 x (precision x recall)/(precision + recall)

sufficient number of words and/or functional/conceptual engineering to distinguish between these shorter documents.

Table 6.2. Precision and recall with respect to classes with F-measure less than 0.5

genre	recall		precision	
	absolute	relative	absolute	relative
Abstract	0.584	0.596	0.426	0.469
Slides	0.4	0.444	0.507	0.556
Technical Report	0.352	0.473	0.4	0.494
Memo	0.3	0.244	0.338	0.373
Scientific Article	0.478	0.522	0.434	0.465
Poster	0.3	0.278	0.415	0.321
Magazine Article	0.244	0.367	0.415	0.452
Academic Monograph	0.434	0.374	0.5	0.514

Table 6.3. Precision and recall with respect to classes with F-measure greater than or equal to 0.7

genre	recall		precision	
	absolute	relative	absolute	relative
Search Page	0.911	0.929	0.906	0.931
Form	0.933	0.967	0.824	0.757
FAQ	0.989	0.989	0.979	0.984
Blog	0.989	0.974	0.969	0.974
CV	0.969	0.969	0.802	0.861
Book of Fiction	0.862	0.897	0.926	0.929
Handbook	0.922	0.933	0.761	0.792
Minutes	0.899	0.899	0.918	0.856
Home Page	0.911	0.905	0.878	0.891
Front Page	0.974	0.968	1	1
E-Shop	0.905	0.895	0.864	0.859
Speech Transcript	0.835	0.846	0.697	0.726
Technical Manual	0.7	0.711	0.741	0.79
Thesis	0.78	0.78	0.757	0.813
Exam Worksheet	0.678	0.7	0.753	0.788
List	0.774	0.805	0.817	0.797

Table 6.4. Selected cluster groups formed on the basis of confusion

Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Genres	Advertisement Abstract Poster Slides	Academic Monograph Scientific Article Technical Report Thesis	Poems Sheet Music	E-Shop Home Page List	Email Letter Memo

7 Conclusions

The results in this paper provide evidence that genre classification tasks are characterised by different levels of context dependence and that the relative ProWLinG Random Forest model, which expresses documents as relative frequencies of words with respect to the most frequent word in the ProWLinG, shows a performance comparable to an average untrained human labeller. In particular, the ProWLinG Random Forest model performs well with respect to the genre classes that have been determined as less context dependent.

To improve the model to perform high-precision classification, it may be necessary to incorporate linguistic analysis, as has been demonstrated by other research (e.g. Santini, 2007). However, it is our belief that there are several types of frequency statistics to be examined before the model is made heavily language dependent. For instance, the ProWLinG can be modified by partitioning it to target words representing different linguistic functions or high-level concepts, and incorporating presence and density of words (not only ratio and count) within each functional or conceptual group. This may be sufficient to approximate expert classification in many cases without involving sophisticated linguistic engineering. In such a model each word will be represented by several relative frequencies.

The multi-level frequency model described above has not been tested yet, but the ProWLinG has been tested with a partition into sixteen linguistic categories, where words in each category are considered using the relative frequency within each category, using Support Vector Machine,[5] and has shown a 12% improvement on the previous representation (outperforming the best ProWLinG Random Forest performance in this paper). Further experiments will be required before firm conclusions can be reached.

One prominent reason for recommending the multi-level word frequency model is that it is easily adaptable across different languages. That is, both this model and the image feature models introduced in Kim & Ross, 2006; Kim & Ross, 2007a; and Kim & Ross, 2007b use a minimal amount of syntactic structure specific to the language of the document. The immediate applicability of these models across many languages and communities makes it a desirable subject for further investigation.

8 Acknowledgments

This research is collaborative. DELOS: Network of Excellence on Digital Libraries (G038-507618),[6] funded under the European Commission’s IST Sixth Framework Programme, provides a key framework and support, as does the UK's Digital Curation Centre. The DCC[7] is supported by a grant from the Joint Information Systems Committee (JISC)[8] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)[9] [GR/T07374/01]. We would like to thank colleagues at HATII, University of Glasgow, who helped to build the KRYS I corpus with web support.

9 References

Bagdanov, A. and Worring, M. (2001) Fine-grained document genre classification using first order random graphs. In Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR2001).

Barbu, E., Heroux, P., Adam, S. and Turpin, E. (2005) Clustering document images using a bag of symbols representation. In Proceedings of the International Conference on Document Analysis and Recognition, pp. 1216–1220.

[5] Random Forest was too computationally intense to examine thoroughly at the time of writing this paper.

[6] http://www.delos.info

[7] http://www.dcc.ac.uk

[8] http://www.jisc.ac.uk

[9] http://www.epsrc.ac.uk

Bekkerman, R., McCallum, A. and Huang, G. (2004) Automatic categorization of email into folders. Benchmark experiments on enron and sri corpora. Technical Report IR-418, Centre for Intelligent Information Retrieval, UMASS.

Biber, D. (1993) Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), pp. 243-257; DOI: 10.1093/llc/8.4.243.

Biber, D. (1995) Dimensions of Register Variation: a Cross-Linguistic Comparison. New York: Cambridge University Press, 1995.

Boese, E.S. (2005) Stereotyping the web: genre classification of web documents. Master's thesis, Colorado State University.

Breiman, L. (2001) Random Forests. Machine Learning, 45, pp. 5–32.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, Vol. 2, pp. 121-167.

Giuffrida, G., Shek, E. and Yang, J. (2000) Knowledge-based metadata extraction from postscript file. In Proceedings of the 5th ACM International Conference on Digital Libraries, pp. 77–84.

Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z. and Fox, E.A. (2003) Automatic document metadata extraction using support vector machines. In Proceedings of the 3rd ACM/IEEECS Conference on Digital Libraries, pp. 37–48.

Karlgren, J. and Cutting, D. (1994) Recognizing text genres with simple metric using discriminant analysis. In Proceedings of the 15th Conference on Computational Linguistics, Vol. 2, pp. 1071–1075.

Ke, S.W. and Bowerman, C. (2006) Perc: A personal email classifier. In Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006), pp. 460–463.

Kessler, G., Nunberg, B. and Schuetze, H. (1997) Automatic detection of text genre. In Proceedings of the 35th Annual Meeting ACL, pp. 32–38.

Kim, Y. and Ross, S. (2006) Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, Proceedings of the European Conference on advanced technology and research in Digital Libraries, Vol. 4172 of Lecture Notes in Computer Science, pp. 63–74, Springer.

Kim, Y. and Ross, S. (2007a) Detecting family resemblance: Automated genre classification. *Data Science Journal*, Vol. 6, S172-S183, ISSN 1683-1470. http://www.jstage.jst.go.jp/article/dsj/6/0/s172/_pdf

Kim, Y. and Ross, S. (2007b) Examining variations of prominent features in Genre Classification. To appear in Proceedings of the 41st Hawaiian International Conference on System Sciences. Preprint at http://eprints.erpanet.org/130

Minsky, M. (1961) Steps toward Artificial Intelligence. Proceedings of the IRE 49(1), pp. 8-30.

Rauber, A. and Müller-Kögler, A. (2001) Integrating automatic genre analysis into digital libraries. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pp. 1–10, Roanoke, VA.

Ross, S. and Hedstrom, M. (2005) Preservation research and sustainable digital libraries. *International Journal of Digital Libraries*. DOI: 10.1007/s00799-004-0099-3.

Santini, M. (2007) Automatic identification of genre in web pages. Thesis submitted for the degree of Doctor of Philosophy, University of Brighton, Brighton, UK.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000) Text genre detection using common word frequencies. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbruecken, Germany.

Thoma, G. (2001) Automating the production of bibliographic records. Technical report, Lister Hill National Center for Biomedical Communication, US National Library of Medicine.

Witten, H.I. and Frank, E. (2005) Data mining: Practical machine learning tools and techniques. 2nd Edition, San Francisco: Morgan Kaufmann.

Yang, Y., Zhang, J. and Kisiel, B. (2003) A scalability analysis of classifiers in text categorization. In Proceedings of the 26th Annual International ACM SIGIR Conference on research and development information retrieval, ISBN 1-58113-646-3, pp. 96-103.