AUTOMATED GENRE CLASSIFICATION IN THE MANAGEMENT OF DIGITAL DOCUMENTS

Yunhyong Kim and Seamus Ross
Digital Curation Centre (DCC) & Humanities Advanced Technology Information Institute (HATII)
University of Glasgow
11 University Gardens
Glasgow
UK
email: {y.kim, s.ross}@hatii.arts.gla.ac.uk
URL: http://www.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. The characterisation of digital objects in terms of genre also associates the object to the objectives that led to its creation, which indicates its relevance to new objectives in information search. 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.

INTRODUCTION

The volume of digital resources inundating our everyday lives is growing at an enormously rapid pace. This information is emerging from unpredictable sources, in different formats and channels, sometimes involving little regulation and control. The storage, management, dissemination and use of this information has consequently become increasingly complex during recent years. Metadata, embodying the technical requirements, administrative function, and content description of an object, provide quick access to the core characteristics of an object, and therefore lead to efficient and effective management 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, webpages and emails). These efforts often rely on structural elements found to be common among documents belonging to the genre. The structural properties of a document (i.e. where and how elements appear within the document) are selected to satisfy functional requirements imposed on it which, in turn, are derived from objectives that existed at the time of its creation. These objectives characterise the genre class of the document (e.g. to describe research to the postgraduate committee). The structural properties that characterise the genre evolve to accommodate the effective performance of the associated function within the target community or process. Thus, knowing the document's genre (amongst those in a genre schema associated to a community or process) is likely to help predict the region and style in which other metadata may appear in the document. This observation has led us to undertake the construction of 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 binds genre-dependent tools, but also to support the selection, acquisition and search of material in terms of the objectives of document creation. As these objectives convey the relevance of the document with respect to its use within new circumstances, automated genre classification could play a valuable role in appraisal activities.

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 &

[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/

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 in to 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, all of these views still seem to comprise a set of functional requirements that describe:

- *the perspectivic category,* associated to the perspective that the creator wished to retain,
- *the structural category,* i.e. the vehicle of expression used for the document (e.g. whether it is expressed as a graph or flowing text; the symbolic or natural language chosen to express the content),
- *the relational categor*y of a document as *part of a process* such as publication, recruitment, event, or use (i.e. how it is related to other objects and people).

Determining the perspectivic category of the document involves substantial semantic analysis of the text while the relational category of the document involves incorporation of domain knowledge. The detection of document data structure type, on the other hand, is solely dependent on document content. This seems to suggests that the detection of data structure type would be the most immediately manageable goal. However, popular approaches to building genre classification schemas tend to reflect the perspectivic or the relational category of a document rather than the data structure type, and there is no established data structure schema for documents. Unlike data structures in computer programming (trees, graphs, stacks etc.), the data structure of a document is mostly implicit. For example, the genre class Curriculum Vitae, immediately suggests the inclusion of selected sub-topics (e.g. “educational background” and “transferable skills”) and a list of document uses or functions (e.g. job and funding application) but, the structural type of documents belonging to the class, though heavily prescribed by convention, is not explicitly characterised. Nevertheless, there are structural aspects of the documents belonging to the class which is tacitly understood: these can be approximated by a tree data structures (i.e. a the selected sub-topics constituting the nodes in the tree with a variable number of children). The nodes of the same depth are usually either completely ordered or completely invariant under permutation. In addition, entities at different depths in the document are rarely co-referent. A scientific article, on the other hand, involves a more complex range of sub-topic vocabulary and processes (e.g. journal, conference paper, preprint, research description and report). Structurally, it seems more akin to a graph, i.e. multiple relations exist between entities at different depths.

Automated recognition of data structural type may require 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). However, automated tools for such annotation are domain dependent and error prone, i.e. would result in error propagation. Even well-tested part-of-speech (POS) taggers and parsers are domain dependent and can not be trusted to perform well across untested genres and subject areas. For example, *He* is astronomy is most often a reference to the chemical element Helium but, is tagged by the CANDC POS tagger (Clark and Curran 2004) as a personal pronoun (Kim and Webber 2006). 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 modeled 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 to establish a relationship between classification tasks and word statistics. We will investigate this with respect to thirty-one genre classes (Table 2.1) comprising twenty-four classes constructed as general document genres and seven classes constructed as webpage genres.

There are other studies which have focused on word frequency analysis for the purpose of genre classification (e.g. Stamatatos, Fakotakis and 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 which have been frequently used in subject classification of documents. We, on the other hand, want to examine words which appear in a large proportion of documents in a genre without necessarily having a high frequency within each document or the entire corpus.

It should also be pointed out that there have already been studies which 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 of why one would still invest energy on models which examine words only. The most important reason for doing this is that models which already integrate involved linguistic information in feature selection are heavily language dependent and likely to require significant internal change and training to accommodate other languages. We will also show that there are instances where the word frequency model out-performs the sophisticated linguistic model, and also suggest refinements of the model which may effectively approximate the 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.

CORPORA

In this section we introduce two corpora from which we have obtained our experimental data. It is a well recognised fact that there is lack of consolidated data for the study of automated genre classification; there is no standardised genre schema and the number of contexts where genre classification arises as a useful tool require widely different approaches to genres. At this stage of establishing consolidated data, it seems important to scope for as many genres in different contexts as possible to determine which genres lead to the most useful outcome and application. With this motivation in mind, KRYS I has been constructed to encompass a schema of seventy genres. The corpus, however, when it was built, was not constructed to reflect webpage genre classes. To compensate, we have augmented our experimental data with documents from the Santini Web corpus.

Table 2.1. Scope of genres under examination (numbers in parenthesis indicate number of documents in each class).

parent genre	**classes in the parent genre**	
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)
Webpage	Blog (190) FAQ (190) Front Page (190) Search Page (190)	Home Page (190) List (190) E-Shop (190)

Other Functional Document	Slides (90)	Curriculum Vitae (96)
	Speech Transcript (91)	Advertisement (90)
	Poems (90)	Exam/Worksheet (90)

- **KRYS I:**

This corpus consists of documents belonging to one of seventy genres (Table 2.1). The corpus was constructed through a document retrieval exercise where university students were assigned genres, and, for each genre, asked to retrieve from the Internet as many examples they could find (but not more than one hundred) of that genre represented in PDF and written in English. They were not given any descriptions of the genres apart from the genre label. Instead, they were asked to describe their reasons for including the particular example in the set. For some genres, the students were unable to identify and acquire one hundred examples. The resulting corpus now includes 6478 items. The collected documents were reclassified by two people from a secretarial background. 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 total of 5305 documents stored with three labels.

- **SANTINI Web:**

This corpus consists of 2400 web pages classified as belonging to one of seven webpage categories or a pool of unclassified documents. There are 200 documents in each of the classified seven categories and one thousand documents in the unclassified pool. 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] http://www.nltg.brighton.ac.uk/home/Marina.Santini/

HUMAN AGREEMENT

We have conducted an analysis of human agreement in genre classification displayed by the three human labellers of the KRYS I corpus. 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	2745*
student & secretary II	2852*
secretary I & II	2422*
all three labellers	2008*

*****out of 5305

There are 1523 documents (out of 3452) which were labelled by at least two labellers as belonging to the same class and 795 documents which were labelled by all labellers as belonging to the same class. However, in the current investigation, we are not so much interested in the overall agreement of human labelling as the difference in the performance of independent labellers, when the data on which the other two labellers have given the same label is considered to be ground truth. To this end, we have taken the three datasets resulting from the agreement of the three possible pairs of labellers and examined the *accuracy* of the remaining labeller's classification. We have used two metrics to examine this: the numbers in the left most column of Table 3.2 represent the average agreement level intervals and those in the top most row represent the standard deviation intervals of the accuracy with respect to the three datasets. Each box in the table lists the genres corresponding to the indicated accuracy and deviation (using numeric keys indicated in Table 3.3) . The partition of the genres resulting from the table suggests a notion of genre classes which are less context dependent (classes in darker boxes, e.g. Handbook, Minutes, CV, Sheet Music) and those that are highly context dependent (e.g. (e.g. Academic Book and Abstract). That is, the genre classes with high average agreement and small discrepancy are expected to be more *context free*.

Table3.2 Partition of 24 genre classes with respect to agreement and standard deviation of agreement.

	0.05-	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**	**0.5+**
0.95+	17, 18, 22										
0.9	3										

	0.05-	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**	**0.5+**
0.85	21	81	9								
0.8			10, 12, 14								
0.75	7										
0.7					11		80				
0.65			16								
0.6									15		
0.55				13				2			
0.5	6									4	
0.5-	20		5, 8				19				1

Table 3.3: Keys for table 3.2.

1. Academic Monograph
2. Book of Fiction
3. Handbook
4. Abstract
5. Scientific Article
6. Magazine Article
7. Poems
8. Periodicals
9. Email
10. Letter
11. Memo
12. Thesis
13. Business Report
14. Technical Report
15. Technical Manual
16. Form
17. Minutes
18. Sheet Music
19. Poster
20. Advertisement
21. Exam/Worksheet
22. Resume/CV
23. Slides
24. Speech Transcript

EXPERIMENTAL DATA AND EVALUATION METHOD

- **Experimental data**

We have constructed two datasets consisting of the documents in Krys I[*] and Santini Web which belong to one of the thirty-one genres listed in Table 1.1. 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 all the other documents in the thirty-one genre classes (Table 2.1). Dataset I is only involved in the construction of the Prolific Words List in Genres (ProWLinG) described in Section 5.

[*] After steps were taken to remove most of the documents of error type I, II, and III.

We did not deem the study of all the genres in the corpus as being useful at this stage, because the current study is aimed at establishing a relationship between genre classes and word frequency statistics. Increasing the number of genres, while an important step in the complete analysis, 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.

- **Evaluation method**

The experimental results in this paper, for the performance of both human labellers and the automated system will be evaluated using one or more of three conventional metrics: accuracy, precision and recall. To make precise what we mean by 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

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

precision *P(C)* of class *C* is defined to be

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

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

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

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.

In Section 6 we will also discuss error analysis using confusion matrices (which makes explicit which classes have been assigned to the documents of each class) of the performance of the automated classifiers.

AUTOMATED EXPERIMENTS

The automated experiments reported here use three classifiers. The classifiers are defined by one of three statistical methods and the stylistic word frequency features which define the document representation. The three statistical methods are Naïve Bayes (NB) [c.f. Minsky, 1961], Support Vector Machine (SVM) [cf. Burges, 1998] and Random Forests (RF) [Breiman 2001], all available as part of the Weka Machine learning Toolkit (Witten & Frank, 2005). As there have been a lack of controlled experiments to determine the best method for genre classification, we have chosen three very different methods: Naïve Bayes uses the basic principles of Bayes theorem, Support Vector Machine, proven to be effective in previous document classification tasks [Yang 2003], is dependent on classification by hyperplanes, and Random Forest constructs several decision trees that cast a vote, where each tree is built on a random selection of features, and instances are classified as a result of the vote.

The stylistic word frequency representation is intended to capture words commonly used within all the genres under consideration. The notion is similar to other significant term analyses in that the intention is to capture words which appear frequently in each genre as words relevant to the classes under consideration. However, the nature of the relevance in the current model is the pattern of all frequencies together as a relative frequency distribution of words, rather than the frequency of each term independently as a measure of term significance within selected documents or genres. Another distinguishing feature is the use of two disjoint training datasets: one for the creation of a word list to be used in creating frequency distributions and another for the probabilistic modeling of distribution patterns as a feature of genre classes. This prevents the model from over-fitting either training dataset. The models were tested using the standard 10-fold cross validation.

- **Prolific words list in Genres (ProWLinG)**

A word list is constructed by taking words which 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 were 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% 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 which 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 which 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 which 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 which are prolific within any number of genres. 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 2476. The sample set size was varied to test an increased size, and the document count was tested at lower percentage; 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 V 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 using the larger set of classes.

- **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 with the frequency of the most popular word in the file so that each entry is a *relative frequency* with respect the most frequent word in the file. The word frequency representation via a pre-constructed list of words which are prolific within genres is inspired by the following:

- having a separate dataset for word list construction lessens the ill effects of over-fitting the training data
- terms in a document express genre style in different ways: for example,
    1. some terms function as a structural cue merely by their presence (e.g. “minutes” in the title of meeting minutes) [**presence**],
    2. some terms are indicative of style by their count within the document (e.g. verbs “examine”, “investigate” and “show” is expected to have a higher frequency than “love”, “hate” and “goad”in scientific articles)[**count**],
    3. some terms appear very frequently in most genres (e.g. determiner and pronouns) but signal stylistic structure through its ratio to other words of similar character (e.g. see Figure 5.1 and 5.2) [**ratio**],
    4. some terms are stylistically important by their distribution throughout the document (e.g. regularly spaced words such as “what” or “how” in a FAQ sheet) [**density**]
- the above four notions of a term can be captured from three directions: as a discrete string, as part of a functional group (whether linguistic or otherwise) and as part of a conceptual group
- in particular, the term *stop word* is not an absolute concept, but a part of a continuum describing words with a large count and uniform density in the document and may also play an important role in stylistic characterisation (see 3 above).

Absolute frequencies of words are strongly influenced by the length of the document. Although this can be controlled to some extent by truncating the document, this would influence the distributional characteristics of a document such as the ratio mentioned in item 2 above, and also affect statistics which depend on the position within the document being examined such as items 1 and 4. By pre-determining a limited list of genre-related words within a reduced 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 on the one dimensional interpretation of 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.

RESULTS

- **Overall accuracy**

The figures in Table 6.1 are the overall accuracies of the ProWLinG features modeled on three statistical methods tested by 10-cross validation on the entire Dataset II. The top two rows specify the classifier being tested. The overall accuracy in Table 6.1 does not measure up to the best accuracies in genre classification which have been reported in the literature. However, previous studies have mostly examined genre schema of approximately ten genres. Classifications across as many as thirty-one classes have not been sufficiently tested. In fact, the absolute frequency ProWLinG Random Forest displays an overall accuracy of 0.927 when tested only on the Santini Web Data.

The focus of this paper is on the relationship between ProWLing and genre classes, rather than the overall optimised performance rate. That is, even though the overall accuracy for the Random Forest ProWLing model is 0.739, the number is not representative of all the classes in the schema. To examine the differences, we analyse the precision and recall with respect to individual classes in the next section.

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

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

- **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 which have shown F-measure less than 0.5 on either the basis of 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.

Table 6.2. Precision and recall: 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

The selection of best precision and recall include six of the seven webpage genre classes represented in Santini's dataset. The three of the four classes (Curriculum Vitae, Minutes, Handbook, Sheet Music) with respect to which human classifiers have shown the highest levels of agreement were also well-recognised by the ProWLinG Random Forests classifier. 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 training or context.

Table 6.3. Precision and recall: 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

Note that the relative frequency representation does not increase performance with respect to all the classes in Tables 5 and 6. The difference in performance is so slight that it does not seem reasonable to make conclusive remarks. However, there do seem to be more increases, and, where there is a decrease in one of precision or recall, this seems to be usually compensated by an increase in the other. A brief look at the incorrect classifications made by the system, shows that most of the errors produced by ProWLinG Random Forest conform to types of errors that might be produced also by human labellers. The most notable confusion cluster groups are reported in Table 6.4.

At first, the inclusion of Advertisements in Cluster group 1 was surprising, but, upon further thought, it seems reasonable, since 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 will also tend to be fairly short, i.e. the probability of finding words from ProWLinG in the documents of these genres will be lower than documents of other genres. There are less understandable errors such as advertisements being labelled as belonging to the class Poem or Sheet Music. It may be that ProWLinG and the document representation proposed does not contain sufficient number of words and functional/conceptual engineering to distinguish between these shorter documents.

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

CONCLUSIONS

The results in this paper provide evidence that genre classification tasks can be characterised by different levels of context dependence. It has also shown that the relative ProWLinG Random Forest model which expresses documents as a vector whose terms are relative frequencies with respect to the most frequent word from the ProWLinG in the document 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 which have been determined as less context dependent on the basis of human labelling. And, in a few cases outperforms average human performance.

To improve the model to perform high precision classification, it may be necessary to incorporate linguistic analysis of style as been demonstrated by other research (e.g. Santini 2007). However, it is our belief there are several types of frequency statistics to be examined before the model is made heavily language dependent. For instance, although the simplistic relative frequency presented in this document may not be sufficient to emulate classification performance of an expert human labeller, the ProWLinG can be modified and partitioned to represent different linguistic functions or high level concepts, and, also, perhaps, augmented to target words representing specific concepts, so that the presence, count, ratio and distribution of words within each functional or conceptual group may be sufficient to realise expert classification in many cases without sophisticated linguistic engineering. In such a model each word will be represented by a number of relative frequencies expressing all the functional groups the task may require.

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, has been tested with Support Vector Machine[4] and has shown a twelve percent 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 made. 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 and Ross 2006, Kim and Ross 2007a, and Kim and Ross 2007b use

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

a minimal amount of syntactic structure specific to the language of the document. The immediate applicability of these models across many languages and communities suggest that the investigation of suggested models and further variations on these models would be worthwhile.

ACKNOWLEDGMENTS

This research is collaborative. DELOS: Network of Excellence on Digital Libraries (G038-507618) [5]funded under the European Commission's IST 6th Framework Programme provides a key framework and support as does the UK's Digital Curation Centre. The DCC[6] is supported by a grant from the Joint Information Systems Committee (JISC)[7] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)[8]. The EPSRC supports (GR/T07374/01) the DCCs research programme. We would like to thank colleagues at HATII, University of Glasgow[9] who facilitated document retrieval and classification of the KRYS I corpus with web support.

Note on URL references: last accessed 25 April 2008.

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*, 79-83. http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/7569/20622/00953759.pdf?arnumber=953759

Barbu E., Heroux P., Adam S., and Turpin, E. (2005), Clustering document images using a bag of symbols representation. *In International Conference on Document Analysis and Recognition*, pages 1216–1220. http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/10526/33307/01575736.pdf?arnumber=1575736

[5] http://www.delos.info
[6] http://www.dcc.ac.uk
[7] http://www.jisc.ac.uk
[8] http://www.epsrc.ac.uk
[9] http://www.hatii.arts.gla.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.
http://whitepapers.silicon.com/0,39024759,60306687p,00.htm

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

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

Boese, E. S. (2005), Stereotyping the web: genre classification of web documents. *Master's thesis*, Colorado State University.
http://www.cs.colostate.edu/~boese/Research/index.html

Breiman, L. (2001), Random forests. *Machine Learning*, 45:5–32.

Burges, C. J. C. (1998), A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, Vol 2, 121-167.
http://citeseer.ist.psu.edu/burges98tutorial.html

Clark, S. and Curran, J. (2004), Parsing the WSJ using CCG and log=linear models. *In proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

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*, pages 77–84.
http://citeseer.ist.psu.edu/giuffrida00knowledgebased.html

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*, pages 37–48.
http://portal.acm.org/citation.cfm?id=827146

Karlgren, J. and Cutting, D. Recognizing text genres with simple metric using discriminant analysis. (1994), *In Proceedings of the 15th Conference on Computational Linguistics*, volume 2, pages 1071–1075.
http://portal.acm.org/citation.cfm?id=991324&dl=GUIDE,

Ke, S. W. and Bowerman, C. (2006), Perc: A personal email classifier. *In Proceedings of the 28th European Conference on Information Retrieval*, pages 460–463.
http://www.springerlink.com/content/r27700t736786455/

Kessler, G., Nunberg, B., and Schuetze, H. (1997), Automatic detection of text genre. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38.
http://www.aclweb.org/anthology-new/P/P97/P97-1005.pdf

Kim, Y. and Ross, S. (2006), Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, *In Proceedings of the European Conference on advanced technology and research in Digital Libraries*, volume 4172 of Lecture Notes in Computer Science, pages 63–74. Springer.
http://www.springerlink.com/content/2048x670g9863085/

Kim Y. and Webber, B. (2006), Implicit references to citations: a study of astronomy papers. *Presentation 20th International CODATA conference.*
http://eprints.erpanet.org/127

Kim, Y. and Ross, S. (2007a), Detecting family resemblance: Automated genre classification. to appear, *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. *In Proceedings 41st Hawaiian International Conference on System Sciences.*
http://www.ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=4438696&arnumber=4438835&count=502&index=138

Minsky, M. (1961), Steps toward Artificial Intelligence. *In Proceedings of the IRE* 49 (1), 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*, pages 1–10, Roanoke, VA.
http://portal.acm.org/citation.cfm?id=379437.379439&coll=&dl=&type=series&idx=SERIES492&part=series&WantType=Proceedings&title=DL

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.
http://www.itri.brighton.ac.uk/~Marina.Santini/

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*, 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, Morgan Kaufmann, San Francisco.

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 in Information Retrieval*, ISBN 1-58113-646-3, 96-103.