Feature Type Analysis in Automated Genre Classification

First Author
Address
FirstEmail@emailserver

Second Author
Address
SecondEmail@emailserver

ABSTRACT

In this paper, we compare classifiers based on language model, image, and stylistic features for automated genre classification. The majority of previous studies in genre classification have created models based on an amalgamated representation of a document using a multitude of features. In these models, the inseparable roles of different features make it difficult to determine a means of improving the classifier when it exhibits poor performance in detecting selected genres. By independently modeling and comparing classifiers based on features belonging to three types, describing visual, stylistic, and topical properties, we demonstrate that different genres have distinctive feature strengths.

Categories and Subject Descriptors

I.2.7 **[Natural Language Processing]**: Text Analysis.

I.5.2 **[Design Methodology]:** Classifier design and evaluation.

General terms

Design, Theory, Management, Performance, Experimentation

1 INTRODUCTION

This paper examines the interaction between three types of document features and the detection of different document genres. The research is part of an initiative to automate the ingest, management, and preservation of material in a digital repository, where metadata describing the technical characteristics, function, source, and content of data play a crucial role in the efficient and effective management and re-use of material (cf. [19]). The manual collection of such 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.

There have been efforts (e.g. [10], [11], [23], [8], [3], [13]) to extract relevant metadata from digital documents. These often rely heavily on the structure which characterises the genre class (e.g. scientific article, email, and newspapers) to which the document under consideration belongs. The reliance of these methods on known genre structure emphasises the benefits of constructing a tool for automated genre classification. An effective automatic genre classifier would function as an overarching tool. It would provide a first-level classification of documents into those of a similar structure which would facilitate the extraction of further information.

Genre is a highly mutable context-dependent concept. Its conceptual variability is apparent in the literature: Biber's analysis ([4]) of document genres employed five dimensions (information, narration, elaboration, persuasion, abstraction) to characterise text, while others ([12], [5]) examined common genre classes such as FAQ, Job Description, Editorial or Reportage. There have been attempts ([14], [9]) to detect particular facets (narratives, fact versus opinion, intended level of audience, and, positivity or negativity of opinion) and an attempt ([1]) to distinguish selected journals and brochures from one another using visual layout. Others ([14], [2]) have clustered documents into similar feature groups without delving into genre facets or classes, and some have championed a multi-genre schema for webpage classification ([21], [22]). Santini has reviewed different approaches to genre classification ([20]).

A prevailing notion in earlier analyses is that genre classification is orthogonal to topic or subject classification. This notion defines genre classification as a task independent from subject classification. While there may be a conceptual level at which this is true, within the probabilistic framework on which language processing is highly reliant, there is reason to believe that this is not generally the case. For example, consider the topic of *cohomology,* a well-known subject area in higher mathematics; this topic would not be expected to appear as frequently in the genre class Reportage as it would in the genre class Research Article. This suggests that, at least on a probabilistic level, genre often moves in close proximity to subject.

This paper contends that genre classification, as described in previous studies, is actually a combination of several independent tasks masquerading as a single task. For example, the distinction between a Thesis and Scientific Paper is largely structural, while Meeting Minutes and Legal Proceedings are distinguished by topical and stylistic term frequency. On the other hand, the distinction between a Table of Financial Statistics and a Financial Report lies in the visual representation and stylistic term frequency. Using the same features to model concurrently these different types of classification would be equivalent to estimating a single distribution for items which belong to distinct populations. If you examine previous literature (e.g. Table 5 in [12], Table 3 in [14]), classification errors range anywhere from seventeen percent to seventy-six percent ([12]), and six percent to eighty percent ([14]). Observing such big differences in error rate might indicate that a re-evaluation of the task, to determine if the task is actually a combination of many tasks disguised by the single term genre classification, is necessary.

The present paper presents tests on two corpora of genre-labelled PDF documents to examine the correlation between genre classes and three feature types, to demonstrate that the best feature types for detecting any one genre class are not necessarily the best for

detecting other genre classes. The feature types we will examine are visual layout features, language modeling features and stylistic word frequency. It is not the intention of this paper to introduce a classifier optimised to perform genre classification (in contrast to [16]). The paper is intended to put forward evidence that establishing a correlation between feature types and genre classes is a reasonable step towards constructing a robust genre classification system.

We have initially limited our study to PDF documents because of its popularity across library, archival, commercial and private sectors. This popularity implies that a classification tool developed for this format is likely to have widespread application. Although the study is of PDF documents, the methods described here do not use features dependent on elements available only in PDF documents. The process is dependent on the PDF only in so far as it depends on PDF tools to convert the documents into image and text.

2 DEFINING GENRE

While the definition of genre may not be easily pinned down, there is general 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 full-fill 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. [15])**.**

There seems to be a general agreement that 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, as an expression of opinion, as 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 p(e.g. table, graph, chart, list).

In this context, it is intuitively clear that selected features will be dependent on one of five aspects: visual layout, stylistic terms, topical terms, semantic patterns, and contextual features which reflect the process for which the document was created (cf. [16]). The proposed objective is to study these *feature types* in relation to genre classes to create a multi-faceted classifier system which performs the detection of *visual genres (*e.g. data structure type)*, stylistic genres* (e.g. prescribed style) and *topical genres* (e.g. business versus legal briefing paper) independently. In this paper we first examine the visual layout, stylistic terms and topical terms in relation to genre classification. Subsequently we will examine semantic patterns as manifestations of topological structures formed by these three feature types.

A second dimension of complexity in classification tasks arises when the number classes represented increases. In genre classification, the number of classes is potentially infinite. This paper expands on the scope of previous studies in addressing the complexity arising from the inclusion of seventy genres (described in **Section 3**). It presents an analysis of human agreement over these genres. This not only creates a baseline performance against which to gauge the automated classifier but also helps to estimate the level of controversy expected within each genre. The latter information will define another means of measuring the quality of a classifier, expressed by the stability of performance (without retraining) on genres attaining a consistent level of agreement between human labellers.

3 GENRE SCHEMA

In this paper we investigate seventy genres which have been organised into ten groups (**Table 1**). The schema was constructed from an examination of 570 PDF documents gathered from the internet using random search words. For the selection of each item, the algorithm newly selected a random word from SCOWL (Spell Checker Oriented Word List - available from sourceforge.net), retrieved a list of PDFs containing the search word, and then archived a random PDF from the first hundred PDFs of the returned list. The schema is meant to capture a wide range of genres commonly available: it is not necessarily aimed to conform to a fixed institutional view of document classes. The aim was to create a logical structure rather a classical structure. In response, certain distinctions may seem at first inconsistent or ambiguous: for instance, Legal Proceedings versus Legal Order, or Technical Manual versus Manual. However, when you view the entire path (e.g. *Evidential Document - Legal Proceedings* versus *Other Functional Document – Legal Appeal, Proposal or Order)* as genres, the distinctions become clearer.

Table 1. Genre schema

Genre group	Genre
Book	Academic Monograph Book of Fiction Poetry Book Handbook Other Book
Article	Abstract Scientific Research Article Other Research Article Magazine Article News Report
Short Composition	Fictional Piece Poem Dramatic Script Essay Short Biographical Sketch Review
Serial	Periodicals (Newspaper, Magazine) Journal

	Conference Proceedings
	Newsletter
Correspondence	
	Email
	Letter
	Memo
	Telegram
Treatise	
	Thesis
	Business or Operational Report
	Technical Report
	Miscellaneous Report
	Technical Manual
Information Structure	
	List
	Catalogue
	Raw Data
	Table Calendar
	Menu
	Form
	Programme
	Questionnaire
	FAQ
Evidential Document	
	Minutes
	Legal Proceedings
	Financial Record
	Receipt
	Slips
	Contract
Visually Dominant Document	
	Artwork
	Card
	Chart
	Graph
	Diagram
	Sheet Music
	Poster
	Comics
Other Functional Document	
	Guideline
	Regulations
	Manual
	Grant or Project Proposal
	Legal Appeal, Proposal, or Order
	Job, Course, or Project Description
	Product or Application Description
	Advertisement
	Announcement
	Appeal or Propaganda
	Exam or Worksheet
	Factsheet
	Forum Discussion
	Interview
	Notice
	Resume or CV
	Slides
	Speech Transcript

4 HUMAN DATA LABELLING

In addition to the data set described in the previous section, we have constructed another corpus of documents tagged by human labellers as belonging to one of the seventy genres in **Table 1**. The objective is to eventually populate each genre with enough examples for the corpus to be deemed statistically useful.

Document retrieval

Students were assigned genres from **Table 1,** and, for each genre, asked to retrieve from the Internet one hundred examples 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[1]. For some genres, the students were unable to identify and acquire one hundred examples. The resulting corpus (**KRYS I)** included 5508 items.

Reclassification

To cross validate the genre classification within the **KRYS I** corpus, two secretaries were employed to reclassify the PDFs within the dataset using the same schema. Secretaries were selected for this task because they were expected to have substantial experience with diverse classes of documents. They worked independently and documents were presented to them without their original label, in a random order. The secretaries were provided with genre names, but were not supplied with descriptions of the genres.

Table 2 presents a description of the corpus after reclassification.

Table 2. Manually gathered data

Total	With three labels	With two labels	Damaged
5508	5373	103	32

We refined the **KRYS I** corpus used in this paper by narrowing it to reflect labelling agreement between different labellers. The figures in **Table 3** show the number of documents on which different groups of labellers have agreed.

Table 3. Human agreement analysis.

Labeller group	Agreed
student & secretary I	2745*
student & secretary II	2974*
secretary I & II	2422*
all three labellers	2008*

*out of 5373

There were several types of errors that arose during the document gathering stage of corpus development. For example, some students initially misunderstood the task and corrupted the data by

1The rationale given by the students for selecting particular items for inclusion in the KRYS I Corpus is the subject of a separate paper. .

- including items which 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*],
- including empty templates as examples of the genre (e.g. instead of selecting 'actual' receipts, empty receipt forms were found labelled as receipts) [*Error type II*],
- and, including entire magazines, conference proceedings or journals as research articles, and vice versa [*Error type III*].

The numbers of items in the database, excluding estimated numbers of *Error type I, II* and *III* are presented in **Table 4**. Numbers are estimated from a scan of several random samples of the database.

Table 4. Composition of PDF in database

Genre	Estimated number of reliable items
Book	
Academic Monograph	16
Book of Fiction	12
Poetry Book	20
Handbook	100
Other Book	30
Article	
Abstract	83
Scientific Research Article	88
Other Research Article	58
Magazine Article	77
News Report	28
Short Composition	
Fictional Piece	20
Poems	46
Dramatic Script	54
Essay	88
Short Biographical Sketch	100
Review	85
Serial	
Periodicals (Newspaper, Magazine)	20
Journals	42
Conference Proceedings	93
Newsletter	46
Correspondence	
Email	19
Letter	99
Memo	84
Telegram	17
Treatise	
Thesis	99
Business or Operational Report	80
Technical Report	90
Miscellaneous Report	95
Technical Manual	80
Information Structure	
List	50
Catalogue	95
Raw Data	50
Table Calendar	100
Menu	85
Form	100
Programme	35
Questionnaire	100
FAQ	100
Evidential Document	
Contract	36
Financial Record	95
Minutes	90
Legal Proceedings	32
Receipt	30
Slips	8
Visually Dominant Document	
Artwork	33
Card	53
Chart	95
Comics	11
Diagram	31
Graph	45
Poster	100
Sheet Music	40
Other Functional Document	
Advertisement	32
Announcement	50
Appeal or Propaganda	25
Exam or Worksheet	28
Fact sheet	100
Forum Discussion	12
Grant or Project Proposal	50
Guideline	52
Interview	100
Job, Course or Project Description	71
Manual	100
Notice	29
Product or Application Description	99
Regulations	100
Resume or CV	106
Slides	95
Speech Transcript	101

4.1 Human Performance

Human performance in genre classification lays a foundation for interpreting the results of machine-based genre classification. As a way of examining this problem, we here examine the performance of the secretaries against the student labels taken as ground truth. The results are shown in **Tables 5** and **6**. These performances have been considered on a group level rather than at genre class level to give an accessible overview. The results, as with those of the automated labelling experiment to be presented in **Section 6**, 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.

According to the definition of recall defined above, the first column of figures in **Tables 5** and **6** is the percentage of items labelled as belonging to the genre in the left most column, by both the secretary and the student (the estimated student errors are not put into consideration at this stage) out of items labelled as such by the students alone. The second column of figures is an estimate of what the recall would be likely to be if the items of *Error type I, II* and *III* have been correctly estimated and removed.

Table 5. Performance of Secretary 1 on student label

Genre group	**Recall (%)**	**Estimated true recall (%)**
Article	38.2	47.9
Book	64.1	93.3
Correspondence	54.8	73.5
Evidential Document	54.1	82.5
Information Structure	58.5	63.8
Serial	58.8	79.6
Short Composition	58.4	61.1
Treatise	34.9	39.2
Visually Dominant Document	30.5	51.7
Other Functional Document	65	70.1

Table 6. Performance of Secretary 2 on student label

Genre group	**Recall (%)**	**Estimated true recall (%)**
Article	25.4	32.6
Book	46.9	71.9
Correspondence	51.3	69.9
Evidential Document	54.1	71.5
Information Structure	73.1	80
Serial	67.9	92.5
Short Composition	72.4	78.6
Treatise	52.1	58.8
Visually Dominant Document	25.2	44.4
Other Functional Document	70.8	77.2

Although the genre groups Correspondence and Visually Dominant Document exhibit low agreement between the secretaries and the student, the performance of both secretaries are comparable, while the groups Book and Serial have produced a notable disparity in performance between both students and the secretaries and between the secretaries themselves.

The figures in **Table 7** highlight the performance of the secretaries (assuming the validity of original student labels without correction) on selected genres showing comparable high percentage of recall and, whether low or high, comparable precision. The only discrepancy in performance relates to the genre Handbook, where Secretary 1 achieved 27.3% precision in contrast to the 78.2% attained by Secretary 2..

Table 7. Secretary performance: genre level

Genre	**Secretary 1**		**Secretary 2**	
	recall (%)	**precision (%)**	**recall (%)**	**precision (%)**
Handbook (100)	96	27.3	97	78.2
Minutes (100)	96	89.6	96	92.3
Form (100)	74.7	32	82	47.4
Poems (50)	72	72	74	75.5
Dramatic Script (55)	78.2	91.5	96.4	89.8
Sheet Music (41)	97.6	97.6	92.7	100
Job Course or Project Description (71)	80.3	68.7	88.7	60.6
Resume/CV (106)	94.3	90.1	97.2	92

The most balanced performances measured in terms of precision and recall seem to be exhibited on the classification of the genres of Minutes, Poems, Sheet Music and Resume. This seems

to reflect the fact that these genres target many communities, hence tend to be more context-free, while different types of articles and books (see performance in **Table 5** and **6**) target specialised communities, and, therefore, are heavily context dependent. The aspects of human performance merit further consideration and we intend to do this using the **KRYS I** corpus and its successor **KRYS II**.

5 DATA AND CLASSIFERS

In this section we introduce the datasets and classifiers that were used in the automated genre labelling experiments described in **Section 6**.

5.1 Data

There are two datasets used in the experiments of this section:

Dataset I

This dataset consists of the documents randomly collected from the internet (see Section 3) using a PDF grabber working in conjunction with a search word from SCOWL. The dataset was labelled by one of the authors.

Dataset II

This dataset consists of the subset of **KRYS I** corpus, described in Section 4, on which all labellers have agreed, comprising 2008 documents.

5.2 Classifiers

We used three different classifiers in the experiments described in **Section 6:** Image, style and Rainbow text classifier.

Image classifier

The first page of the document was converted into a low resolution grey-scale image and sectioned into a sixty-two by sixty-two grid. Each region on the grid was examined for non-white pixels. All regions with non-white pixels were labelled 1 while those which are completely white were labelled 0 to create a low resolution bit map. The choice of sixty-two to define the size of the grid, and other parameters, were set to mirror the coarsest level of granularity at which subjects were able to recognise particular documents as members of specific genre classes. Examples of the image representation are given in **Figure 1** (we were unable to adjust the degradation resulting from the process of converting it and resizing it to fit this document). The resulting vector was then probabilistically modeled via Naïve Bayes and the Random Forest Decision method ([6]) using the Weka Machine Learning Toolkit([24]).

The reasons for using this type of classifier reflects the recognition that certain genres have more or less white space in the first page (e.g. the title page of the book). And, for some genres, the page is strictly formatted (e.g. slides for a conference presentation). Some genres use visual tactics to catch the attention of the reader (e.g. the reverse colouring on a magazine cover). Another benefit of examining documents using image processing methods is that it does not depend on extracting text, can be language independent, and supports document analysis even when the content of the document is not accessible to the public.

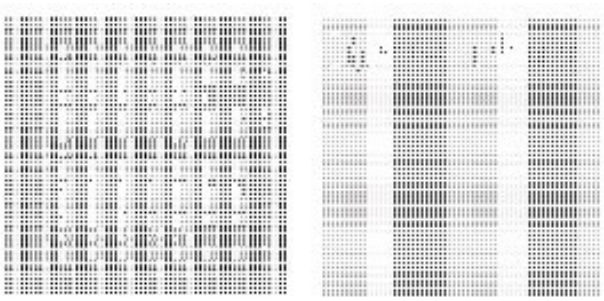

Figure 1. Image representation of documents: Research Article (left) and Periodicals (right).

Style classifier

Two models were constructed, one using Naïve Bayes, and another using the Random Forest method ([6]) provided with the Weka Machine learning toolkit ([24]). This classifier is intended to capture frequency of words common to all genres as well as words significant to only some genres. The traditional method of text classification depended on the selection of significant terms by employing a variation on the *tf*idf* (term frequency times the inverse document frequency) and then examining the absence or presence of the significant terms. The thesis of this paper is that even patterns which appear in a wide variety of genres may be a significant metric when the frequency is also taken into consideration. From a dataset previously collected from the Internet using the same method as that which was employed for the collection of Dataset I (but of third the size), the union of all words found in half or more of the files, in each genre, was retrieved and compiled into a list. For each document a vector is constructed using the frequency of each word in the compiled list. This representation is used in the classification of the sixteen genres (**Section 6.2**) which is different from those used in the binary classification (**Section 6.1**). The latter also incorporated the number of words, font sizes and variations.

Table 8. Average frequency of words in three genres.

Genre Word	Memo	Form	Slide
have	24	0	0
for	56	17	6
with	31	4	0
do	8	3	0
the	477	57	21
review	4	0	0
information	0	5	0

A typical example of the weight in measuring frequency is embodied in the fact that forms or slides are less likely to contain as many definite or indefinite articles as flowing text. Sample

average word frequencies (the average taken over ten documents) found in three selected genres are presented in **Table 8.**

Rainbow text classifier

This is a text classifier included in the BOW toolkit developed by Andrew McCallum ([17]). This toolkit indexes the alpha-numeric content of the text for an analysis of significant term frequencies. It supports several statistical methods for evaluation. We have used the Naïve Bayes method to enable comparison with the other classifiers which also use Naïve Bayes. The Random Forest method was not available with this module. This classifier is popular for subject classification; we will show its effectiveness in *topical* genre classification.

6 EXPERIMENTS AND RESULTS

Using the datasets and classifiers described in Section 5, we conducted two types of experiments: the binary classification of documents into those of a selected genre and those of other genres, and the classification of documents into sixteen genres.

6.1 Binary Classification

In this section we present the performance of image, style and rainbow classifiers modeled using Naïve Bayes in the context of sample binary classifications. The experiments were conducted on a subset of Dataset I, using 10-fold cross validation. The subset consisted of all documents classified into one of nineteen genres: Academic Book, Book of Fiction, Other Book, Scientific Research Article, Other Research Article, Magazine Article, Periodicals, Newsletter, Thesis, Business Report, Technical Report, List, Forms, Minutes, Guideline, Job Description, Product Description, Factsheet, and Slides.

The three classifiers were compared on three binary classifications (**Tables 9, 10 and 11**): a classification of the documents into Periodicals and Other, Scientific Research Article and Other, and, Thesis and Other.

Table 9. Binary classification: Periodicals and Other. Image (top), style (middle) and language model (bottom)

10 fold Cross validation with the image classifier, overall accuracy: 88.6%		
Genre	**Precision(%)**	**Recall(%)**
Periodicals (16 items)	29.8	87.5
Other Genre (291 items)	99.2	88.7

10 fold Cross validation with the style classifier, overall accuracy: 88.52%		
Genre	**Precision(%)**	**Recall(%)**
Periodicals (16 items)	14.8	25
Other Genre (291 items)	92	93.8

10 fold Cross validation with the rainbow text classifier, overall accuracy: 94.79%		
Genre	**Precision(%)**	**Recall(%)**
Periodicals (16 items)	ND*	0
Other Genre (291 items)	96.9	100
* Not defined since no items were returned as Periodicals.		

An interpretation of the overall accuracies presented in **Table 9** suggests the Rainbow classifier to be the best performer, but closer examination shows that this is only because the Other category contains a far greater number of items, and the classifier is labelling all items as belonging to the Other category. The result here seems to indicate that the image classifier is much more successful in detecting Periodicals.

The poor recall rate of the Rainbow classifier is noticeable also in the case of Scientific Research Articles (**Table 10)**. The precision on the other hand is excellent. In comparison, the image and style classifiers exhibit good recall rate, but poor precision. This suggests that stylistic terms and visual layout are shared by many scientific articles (as well as other genres) while some significant terms are only shared by scientific articles in the dataset (cf. See example of *cohomology* in **Section 1**).

Table 10: Binary classification: Scientific Research Articles and Other. Image (top), style (middle) and language model (bottom)

10 fold Cross validation with the image classifier, overall accuracy: 73.94 %		
Genre	**Precision(%)**	**Recall(%)**
Scientific Research Articles (25 items)	21.11	80
Other Genre (280 items)	97.6	73.4

10 fold Cross validation with the style classifier, overall accuracy: 91.80 %		
Genre	**Precision(%)**	**Recall(%)**
Scientific Research Articles (25 items)	50	76
Other Genre (280 items)	97.8	93.2

10 fold Cross validation with the rainbow text classifier, overall accuracy: 94.68 %		
Genre	**Precision(%)**	**Recall(%)**
Scientific Research Articles (25 items)	100	15
Other Genre (280 items)	94.6	100

Table 11. Binary Classification: Thesis and Other. Image(top), style (middle) and language model (bottom)

10 fold Cross validation with the image classifier, overall accuracy: 82.74 %		
Genre	**Precision(%)**	**Recall(%)**
Thesis (10 items)	13.6	80
Other Genre (280 items)	99.2	90.3

10 fold Cross validation with the style classifier, overall accuracy: 75.40 %		
Genre	**Precision(%)**	**Recall(%)**
Thesis (10 items)	7	60
Other Genre (280 items)	98.2	75.9

10 fold Cross validation with the rainbow text classifier, overall accuracy: 93.87 %		
Genre	**Precision(%)**	**Recall(%)**
Thesis (10 items)	40	17.4
Other Genre (280 items)	98	95.67

The results of **Table 11** do not lead us to propose any of the classifiers as a promising solution for the detection of members of the genre Thesis. The discrepancy between precision and recall suggests that the given feature types on their own are insufficient as a basis for building a reliable detector.

The image and style classifiers were further compared on two additional binary classifications (**Tables 12** and **13**): a classification of the documents into Business Report and Other, and Forms and Other. The results indicate that the style classifier performs slightly better on both classifications, however, the difference is inconclusive and further examination is required before any firm conclusions can be made.

Table 12. Binary classification: Business Report and Other. Image(top) and style (bottom)

10 fold Cross validation with the image classifier, overall accuracy: 60.72 %		
Genre	**Precision(%)**	**Recall(%)**
Business Rep (10 items)	5.6	63.6
Other Genre (280 items)	97.8	60.1

10 fold Cross validation with the style classifier, overall accuracy: 72.79 %		
Genre	**Precision(%)**	**Recall(%)**
Business Report (10 items)	9.1	72.7
Other Genre (280 items)	98.6	72.8

Table 13. Binary classification: Forms and Other. Image(top) style (bottom)

10 fold Cross validation with the image classifier, overall accuracy: 76.55 %		
Genre	**Precision(%)**	**Recall(%)**
Forms (10 items)	7.2	38.5
Other Genre (280 items)	96.6	78.2

10 fold Cross validation with the style classifier, overall accuracy: 71.48 %		
Genres	**Precision(%)**	**Recall(%)**
Forms (10 items)	10.6	76.9
Other Genre (280 items)	98.6	71.2

Table 14. Detecting five types of genres using the rainbow text classifier

10 fold Cross validation with the rainbow text classifier, overall accuracy on 5 classes: 82.4%		
Genre	**Precision(%)**	**Recall(%)**
Academic Book (5 items)	42.9	60
Business Report (11 items)	90	81.8

Fictional Book (14 items)	100	100
Minutes (13 items)	86.7	92.9
Thesis (10 items)	75	60

The performance of the Rainbow text classifier in making a five-class decision is presented in **Table 14.** The results suggest significant term analysis as an effective support for topical genre classification. The classifier failed to distinguish Academic Books from Thesis where topics overlap, while it succeeded in achieving fair precision and recall in relation to Business Report, Fictional Book and Minutes where topics are expected to overlap less frequently. The significant term analysis has proven to be effective for the detection of fictional work in other studies as well (cf. [12], [14]).

6.2 Classification of Sixteen Genres

In this section we report a comparison of image and style classifiers modeled using the Random Forest method in classifying documents representing sixteen possible genres. The 10-fold cross validation results of this experiments are evaluated on the documents of sixteen genres from **KRYS I** are given in **Tables 15** and **16.**

This experiment is meant to supplement the experiments in the binary classification, by presenting a classification of documents from a different dataset into a larger number of genres for feature type analysis, to demonstrate that the observations in **Section 6.1** are not confined to Dataset I. The classifiers were also tested on Naïve Bayes but the results were much poorer when compared to the Random Forest Method (overall accuracies approximately 22%).

Table 14. Image classifier: overall accuracy 38.37%

Genre (no. of items)	Precision (%)	Recall (%)
Article		
Magazine Article (20)	17	5
Scientific Research Article (7)	0	0
Other Research Article (18)	50	67
Book		
Book of Fiction (3)	18	25
Information Structure		
Form (60)	40	50
List (19)	0	0
Serial		
Periodicals		
(Newspaper, Magazine) (6)	33	14
Newsletter (20)	13	6
Treatise		
Technical report (46)	19	11
Business/Operational Rpt (9)	0	0
Thesis (59)	56	84
Evidential Document		
Minutes (91)	47	77
Other Functional Document		
Slides (23)	94	73
Product/Application Desc. (56)	14	10
Guideline (28)	0	0
Factsheet (69)	25	33

Table 15. Stylistic classifier: overall accuracy 69.96%

Genre (no. of items)	Precision (%)	Recall (%)
Article		
Magazine Article (20)	82	47
Scientific Research Article (7)	0	0
Other Research Article (18)	56	39
Book		
Book of Fiction (3)	0	0
Information Structure		
Form (60)	69	88
List (10)	57	47
Serial		
Periodicals		
(Newspaper, Magazine)(6)	0	0
Newsletter (20)	100	18
Treatise		
Technical Report (46)	74	73
Business/Operational Rpt (9)	67	25
Thesis (59)	72	86
Evidential Document		
Minutes (91)	99	99
Other Functional Document		
Slides (23)	40	27
Product or Application Desc. (56)	62	80
Guideline (28)	35	25
Factsheet (69)	67	82

The style classifier (cf. **Table 16**) shows a high level of accuracy on the new data, given the fact that the classifier was not developed on this data, and that the features employed by this classifier are much simpler than previous attempts by other researchers using multidimensional stylistic features (e.g. existence or frequency of clauses adjectives and adverbs along with lengths of sentences and words). This suggests the frequency of common words as a powerful feature in detecting genres. The prediction of Minutes is particularly noticeable at 99% precision and recall. The best performance of the image classifier is on the classification of the genre Slides.

In parallel to the binary classification, the results of **Table 15** and **16** continues to present the image classifier as as a better classifier in detecting periodicals (cf. **Table 9**), and to suggest stylistic features as a more prominent factor for detecting members of the genres Form and Business Report than the image features(cf. **Tables 12** and **13).** Both image and stylistic classifiers maintain a reasonable recall rate for the genre Thesis (cf. **Table 11**). Further tests on a larger dataset (which we will have upon refinement of the KRYS Corpus[2]), on other statistical methods, and an inclusion of the Rainbow text classifier for comparison on the new data (not included here for lack of space) will enable us to make stronger conclusions.

2 A robust version of the corpus will be released in Autumn of 2007.

7 CONCLUSIONS

The results in this paper can be summarised by two main observations:

- Human beings have difficulty doing classification in a context-free environment, and genre classification is more context dependent than some other classification tasks such as subject classification (i.e. it is easier to recognise that an article is about cohomology even if you don't know what cohomology is, than to recognise that a document is a scientific research article if you are not familiar with the academic research environment.)
- Different genres are characterised by different feature strengths: not all features are equally relevant to all genres and an intelligent analysis of visual, stylistic and topical properties is necessary.

The aspects of each feature type examined in this paper consisted of a very basic set. Future work should include the examination of other features of the same three type, and features representing more sophisticated semantic patterns and procedural context, for correlation with genre classification tasks.

To utilise classifiers to work in a management system for the improvement of information networks in the long term, we need a more precise understanding of the relationship between task and features, and the co-dependencies (or the absence there of) between different classification tasks.

8 ACKNOWLEDGMENTS

[acknowledgments have been omited to anonymise the paper]

9 REFERENCES

[1] Bagdanov, A. and Worring, M. Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR2001)* 2001, 79-90.

[2] Barbu, E., Heroux, P., Adam, S., and Turpin, E. Clustering document images using a bag of symbols representation. In *Proceedings International Conference on Document Analysis and Recognition* 2005, 1216–1220.

[3] Bekkerman, R., McCallum, A., and Huang, G. *Automatic categorization of email into folders: benchmark experiments on enron and sri corpora.* Technical Report IR-418, Centre for Intelligent Information Retrieval, UMASS, 2004, http://www.cs.umass.edu/~mccallum/papers/foldering-tr05.pdf

[4] Biber, D. *Dimensions of Register Variation:a Cross-Linguistic Comparison*. Cambridge University Press, New York, 1995.

[5] Boese, E. S. *Stereotyping the web: genre classification of web documents.* Master's thesis, Colorado State University, 2005.

[6] Breiman, L. Random forests. *Machine Learning*, 45:5–32, 2001.

[7] Chao, C., Liaw, A., and Breiman, L. Using random forest to learn imbalanced data.
http://www.stat.berkeley.edu/breiman/RandomForests/, 2004.

[8] Dc-dot, ukoln dublin core metadata editor.
http://www.ukoln.ac.uk/metadata/dcdot/.

[9] Finn, A., and Kushmerick, N. Learning to classify documents according to genre. *Journal of American Society for Information Science and Technology*, 57(11):1506–1518, 2006.

[10] Giuffrida, G., Shek, E., and Yang, J. Knowledge-based metadata extraction from postscript file. In *Proceedings 5th ACM Intl. Conf. Digital Libraries*, pages 77–84, 2000.

[11] Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. Automatic document metadata extraction using support vector machines. In *Proceedings 3rd ACM/IEEE-CS Conf. Digital libraries*, pages 37–48, 2003.

[12] Karlgren, J., and Cutting, D. Recognizing text genres with simple metric using discriminant analysis. In *Proceedings 15th Conf. Comp. Ling.*, volume 2, pages 1071–1075, 1994.

[13] Ke, S. W., and Bowerman, C. Perc: A personal email classifier. In *Proceedings 28th European Conf. Information Retrieval (ECIR 2006)*, pages 460–463, 2006.

[14] Kessler, G., Nunberg, B., and Schuetze, H. Automatic detection of text genre. In *Proceedings 35th Ann. Meeting ACL* 1997, 32–38.

[15] Kim, Y., and Ross, S. Detecting family resemblance: Automated genre classification. *CODATA Data Science Journal*, *Volume 6, ISSN:1683-1470* (2007), S172-S183.

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

[17] McCallum, A. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.*
http://www.cs.cmu.edu/~mccallum/bow. 1996.

[18] Rauber, A. and Müller-Kögler, A. Integrating automatic genre analysis into digital libraries. In *Proceedings ACM/IEEE Joint Conf. Digital Libraries, Roanoke, VA* (2001), 1–10, http://doi.acm.org/10.1145/379437.379439

[19] Ross, S., and Hedstrom, M. Preservation research and sustainable digital libraries. *Intnl. Journal of Digital Libraries*, v 5.4, 2005, 317-325, DOI: 10.1007/s00799-004-0099-3, http://eprints.erpanet.org/archive/00000095/

[20] Santini, M. State-of-the-art on Automatic Genre Identification, Technical Report ITRI-04-03, ITRI, University of Brighton, UK, 2004.

[21] Santini, M. Towards a Zero-to-Multi-Genre Classification Scheme, *Journée ATALA "Typologies de textes pour le traitement automatique"* (9 décembre 2006), Paris, http://www.nltg.brighton.ac.uk/home/Marina.Santini/marina_santini_ATALA2006.pdf

[22] Santini, M. Characterizing Genres of Web Pages: Genre Hybridism and Individualization, *40th Annual Hawaii International Conference on System Sciences (HICSS'07), http://csdl2.computer.org/comp/proceedings/hicss/2007/2755/00/27550071.pdf.*

[23] Thoma, G. Automating the production of bibliographic records. Technical report, *Lister Hill National Center for Biomedical Communication, US National Library of Medicine*, 2001, http://archive.nlm.nih.gov/pubs/thoma/mars2001.php

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