Searching for Ground Truth: A Stepping Stone in Automating Genre Classification

Yunhyong Kim and Seamus Ross

Digital Curation Centre (DCC)
&
Humanities Adavanced Technology Information Institute (HATII)
University of Glasgow
Glasgow, UK
{y.kim,s.ross}@hatii.arts.gla.ac.uk

Abstract. This paper examines genre classification of documents and its role in enabling the effective automated management of digital doc- uments by digital libraries and other repositories. We have previously presented genre classification as a valuable step toward achieving auto- mated extraction of descriptive metadata for digital material. Here, we present results from experiments using human labellers, conducted to as- sist in genre characterisation and the prediction of obstacles which need to be overcome by an automated system, and to contribute to the process of creating a solid testbed corpus for extending automated genre clas- sification and testing metadata extraction tools across genres. We also describe the performance of two classifiers based on image and stylistic modeling features in labelling the data resulting from the agreement of three human labellers across fifteen genre classes.

Keywords: information extraction, genre classification, automated meta- data extraction, metadata, digital library, data management.

1 Introduction

As digital resources become increasingly common as a form of information in our everyday life, the task of storing, managing, and utilising this informa- tion becomes increasingly important. Managing digital objects not only involves storage, efficient search, and retrieval of objects - tasks already expected by traditional libraries - but also involves ensuring the continuation of technologi- cal requirements, tracking of versions, linking and networking of independently produced objects, and selecting objects and resources for retention from a del- uge of objects being created and distributed. Knowledge representation, em- bodying the core information about an object, e.g. metadata summarising the technical requirements, function, source, and content of data, play a crucial role in the efficient and effective management and use of digital materials (cf. [22]), making it easier to tame the resources within. It has been noted that the manual collection of such information is costly and labour-intensive and that a

collaborative effort to automate the extraction or creation of such information would be undoubtedly necessary [1].

There have been several efforts (e.g. [11], [12], [23], DC-dot metadata editor[2], [3] and [14]) to automatically extract relevant metadata from selected genres (e.g. scientific articles, webpages and emails). These often play heavily on the structure of the document, which characterises the genre to which the document belongs. It seems, therefore, reasonable to employ automated genre classification to bind these genre-dependent tools. However, there is a distinct lack of consolidated corpora on which automated genre classification and the transferability or integrability of tools across genres can be tested. One of the reasons such corpora have not yet been constructed relates to the elusive nature of genre classification, which seems to take on a different guise in independent researches. Biber's analysis ([5]) tried to capture five genre dimensions (information, narration, elaboration, persuasion, abstraction) of text, while others ([13], [6]) examined popularly recognised genre classes such as FAQ, Job Description, Editorial or Reportage. Genre has been used to describe stylistic aspects (objectivity, intended level of audience, positive or negative opinion, whether it is a narrative) of a document ([10], [15]), or even to describe selected journal and brochure ti- tles ([1]). Others ([21], [2]) have clustered documents into similar feature groups, without attempting to label the samples with genre facets or classes.

The difficulty of defining genre is already emphasised in the literature, and many proposals have been reasonably suggested. However, very little active search for ground truth in human agreement over genre classification has been conducted to scope for a useful genre schema and corpus. To shed some light on the situation, we have undertaken experiments to analyse human agreement over genre classification: the agreement analysis will establish the degree of agreement that can be reached by several human labellers in genre classification, isolate the conditions that give meaning to genre classification, and provide a statistically well understood corpus. The corpus will also function as a testbed for examining transferability of tools tailored to work in a small number of genres to other gen- res, and constructing metadata extraction tools which integrate tools developed independently for different genres. In addition, a study of human performance in genre classification provides a means of scoping new emerging genres, and helps us to grasp the history of genre development. To this end, we have constructed a schema of seventy genres (Section 2) and present results in document collection and categorisation by human labellers in Section 3.

Genre classification, in its most general understanding, is the categorisation of documents according to their structural (e.g. the existence of a title page, chapter, section) and functional (e.g. to record, to inform) properties. The two are, however, not divorced from each other: the structure evolves to optimise the functional requirements of the document within the environment (e.g. the target

[1] Issues addressed in The Cedars Project at the University of Leeds: http://www.leeds.ac.uk/cedars/guideto/collmanagemnet/guidetocolman.pdf

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

community, publisher and creator), just as the structure of organisms evolves to meet their survival functions within the natural environment. And, just as the functional aspect of an organism is central to its survival, the functional properties of a digital object is the crucial driving force of document genre. The functional aspect of the document, however, is a high level concept which is inferred from selected structural aspects of the document, and, in turn, the structural aspects are defined by lower level features which constitute genes in the DNA of the document. Unlike organisms, we have not even come close to identifying the DNA sequence of a digital document, let alone parsing the sequence into genes to understand how they are expressed to create semantic information (e.g. genre or subject). Accordingly, automated classification has traditionally taken to examining a large pot of related and unrelated features, to be refined by selection or creation algorithms to distinguish between a small number of predefined classes. This method might result in three immediately noticeable problems:

– The reason for specific selections and creations of features remains opaque. – Features will be selected to conform to the unavoidable bias in the training data.
– The performance of the tool on a new set of classes is unpredictable, and most likely, the tool will have to be reconstructed by re-running feature selection over new data.

To address these points, we propose grouping features according to similar type (e.g. those which come together to describe a well-defined aspect of document structure) in analogy to genes. This makes it easier to identify the reasons behind errors and see if success is an artefact of unrepresentative data. We also propose that a study of a wider variety of genre classes may be necessary. A classifier which performs well to distinguish three classes can be expected to perform well to distinguish two of the three classes; whereas the behaviour of a classifier which recognises two classes in distinguishing three classes is less predictable. The amount of information the class of a document encompasses is in direct relationship to the number of other classes to which it could belong. By building a system which can detect selected genre classes from a vast range of classes, we are building a better informed system.

We have previously identified ([16]) five feature types: image features (e.g white space analysis; cf. [1]), stylistic features (e.g. word, sentence, block statistics; cf. [21]), language modelling features (e.g. Bag-of-Words and N-gram models), semantic features (e.g. number of subjective noun phrases) and source or domain knowledge features (e.g. file name, journal name, web address, techni- cal format structures, institutional affiliations, other works by the author). We reported preliminary results of classifiers built on two or more of the first three feature types on a privately labelled corpus ([16], [18],[19]). In this paper, we look at the performance of a classifier modeled on the first two types of features on new data labelled by three human labellers, as further study of the correlation between genres and feature types.

2 Genre Schema

In this paper we are working with seventy genres which have been organised into ten groups (Table 1). The schema was constructed from an examination of PDF documents gathered from the internet using a list of random search words. The schema captures a wide range of commonly used genres. The aim is to initially vie for a coverage of as many genres as possible rather than to employ a well established 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, the hope is that when you view the entire path as genres, e.g. Evidential Document - Legal Proceedings versus Other Functional Document Legal Order, the distinction will become clearer. The schema will form a fundamental field to be harvested for further refinement. It will be adjusted to exclude ill-defined genres depending on emerging results of the human labelling experiments described in Section 3.

Table 1. Genre schema (numbers in parentheses are assigned database IDs)

Book		
Academic Monograph (2)	Book of Poetry (4)	Other Book (6)
Book of Fiction (3)	Handbook (5)	
Article		
Abstract (8)	Other Research (10)	News Report (12)
Scientific Article (9)	Magazine Article (11)	
Short Composition		
Fictional Piece (14)	Dramatic Script (16)	Short Biographical Sketch (18)
Poems (15)	Essay (17)	Review (19)
Serial		
Periodicals (News, Mag) (21)	Conference Proceeding (23)	
Journals (22)	Newsletter (24)	
Correspondence		
Email (26)	Memo (29)	
Letter (27)	Telegram (30)	
Treatise		
Thesis (32)	Technical Report (34)	Technical Manual (36)
Business/Operational Rept (33)	Miscellaneous Report (35)	
Information Structure		
List (38)	Table (41)	Programme (44)
Catalogue (39)	Menu (42)	Questionnaire (45)
Raw Data (40)	Form (43)	FAQ (46)
Evidential Document		
Minutes (48)	Financial Record (50)	Slip (52)
Legal Proceedings (49)	Receipt (51)	Contract (53)
Visual Document		
Artwork (55)	Graph (58)	Poster (61)
Card (56)	Diagram (59)	Comics (62)
Chart (57)	Sheet Music (60)	
Other Functional Document		
Guideline (64)	Product Description (70)	Forum Discussion (76)
Regulations (65)	Advertisement (71)	Interview (77)
Manual (66)	Announcement (72)	Notice (78)
Grant/Project Proposal (67)	Appeal/Propaganda (73)	Resume/ CV (79)
Legal Proposal/Order (68)	Exam or Worksheet (74)	Slides (80)
Job/Course/Project Desc. (69)	Factsheet (75)	Speech Transcript (81)

3 Human Labelling Experiment

We have undertaken two human document genre classification experiments in this research. First we had students retrieve sample documents of the seventy genres in Table 1 (Document Retrieval Exercise), and subsequently had them re-assigned with genres from the same schema (Reclassification) to validate, quantify, or examine agreement over its membership to any one genre.

Document Retrieval Exercise: In this experiment, university students were assigned genres and asked to retrieve 100 samples of PDF files belonging to their assigned genre written in English. They were also asked to give reasons for including the particular sample in the set and asked not to retrieve more than one document from each source. They were not introduced to pre-defined notions of the genre before retrieval.

Reclassification: Two people from a secretarial background were employed to reclassify the retrieved documents. They 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. The number of items which have been stored in the database is described in Table 2.

At first, it may seem odd not to provide definitions for the genres in the schema. However, note that it is not true that every genre class requires the same amount of detail in its definition to achieve the same level of precision. In fact, as we will see, the level of agreement on some genres is high regardless of the lack of definition.

Table 2. Database composition (left) and Agreement of Labellers (right)

Total	with three labels	with two labels	damaged
5485	5373	103	9

Labellers	Agreed
student & secretary A	2745
student & secretary B	2974
secretary A & B	2422
all labellers	2008

Some of the collected data can not be considered to be examples of the genre. For instance, some students introduced articles about email into the database as samples of the genre Email. Others submitted empty receipt forms as samples of the genre Receipt. The genre Card was also heavily populated with forms (e.g. unfilled identity cards). While the first set of errors are due to a misunderstanding of the instructions and stems from the fact that emails are hard to find in PDF format, the latter sets of errors are due to differing opinions of the genre definition. These items were not removed from the database because:

– this would introduce the bias of the remover into the database; and, –
documents which have been included erroneously will be automatically filtered out of the collection once reclassification labels and agreement data are acquired.

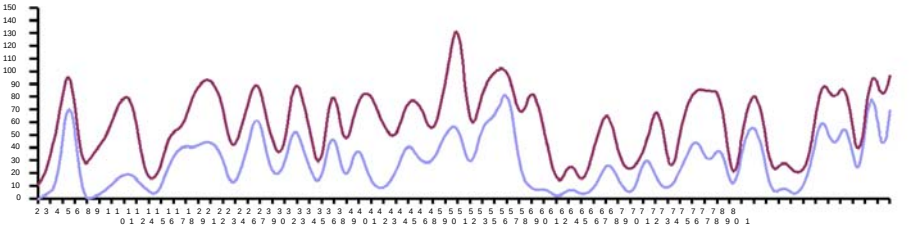

Fig. 1. Two labeller agreement (top graph) versus three labeller agreement (bottom graph)

Full analysis of these errors will be carried out before the release of the corpus, at which time, the rationale presented by students to justify the inclusion of particular items will also be analysed for elements that might characterise genres. In Figure 1, we have presented the numbers of documents in each of the seventy genres on which labellers have agreed. The graph exhibiting higher numbers presents the number of documents on which at least two labellers have assigned the same label, and the lower level graph displays the number of documents on which all three labellers have assigned the same label. The genre classes are indicated as numbers (to save space) along the bottom of the graph, indicating the assigned genre IDs given in Table 1. Note that there is a large discrepancy between the agreement with respect to the genre Form (43), but that selected genres such as Handbook (5), Minutes (48) and Resume/CV(79) show little difference between the two labeller agreement and the three labeller agreement, suggesting the latter genres as less context-dependent genres.

4 Case Study

In this section we will look at the student performance on documents for which secretaries have given the same label. There are 2422 items on which the decision of the secretaries concurred. The figures in Table 2 show the number of documents on which different groups of labellers have agreed. The statistics in Table 2 show that there is more agreement between the student and either of the secretaries than between the two secretaries. A possible explanation for the discrepancy could be that secretaries are trained to identify strictly defined properties of a limited number of genres, while students detect broadly defined properties of a vast range of genres. Further labellers and studies will be required to make any decisive conclusions.

4.1 Precision Versus Recall

In this section we present the recall and precision of student classification on the data which was given the same label by the secretaries. The results are shown in Table 3. Compared to some other classification tasks (e.g. classification of

Table 3. Human Labelling: Overall accuracy: 82.9%

Genre group	Genre	no. of items	Recall(%)	Precision(%)
Book	Academic Monograph	3	0	0
	Book of Fiction	8	37	100
	Book of Poetry	12	67	23
	Handbook	105	88	100
	Other Book	0	0	0
Article	Abstract	1	100	8
	Scientific Research Article	15	47	32
	Other Research Article	36	50	69
	Magazine Article	40	50	61
	News Report	9	89	89
Short Composition	Fictional Piece	1	100	33
	Poems	37	78	91
	Dramatic Script	43	98	100
	Essay	59	68	89
	Short Biographical Sketch	46	100	98
	Review	46	85	83
Serial	Periodicals (Newspaper, Magazine)	21	29	100
	Journals	34	91	86
	Conference Proceedings	76	96	99
	Newsletter	28	71	80
Correspondence	Email	21	90	70
	Letter	67	93	100
	Memo	29	93	71
	Telegram	7	100	78
Treatise	Thesis	66	89	98
	Business/Operational Report	12	75	36
	Technical Report	52	88	94
	Miscellaneous Report	38	34	81
	Technical Manual	7	86	27
Information Structure	List	26	73	86
	Catalogue	51	90	90
	Raw Data	40	73	91
	Table Calendar	30	93	68
	Menu	52	100	96
	Form	114	53	100
	Programme	29	66	100
	Questionnaire	61	98	91
	FAQ	71	90	98
Evidential Document	Minutes	94	97	100
	Legal Proceedings	36	50	58
	Financial Record	7	86	75
	Receipt	8	100	21
	Slips	0	0	0
	Contract	10	90	82
Visual Document	Artwork	2	100	13
	Card	9	100	35
	Chart	39	82	74
	Graph	14	71	48
	Diagram	6	33	18
	Sheet Music	37	100	100
	Poster	23	48	85
	Comics	7	100	27
Other Functional Document	Guideline	48	58	93
	Regulations	53	94	91
	Manual	43	60	96
	Grantor Project Proposal	45	98	81
	Legal Appeal/Proposal/Order	0	0	0
	Job/Course/Project Description	62	89	96
	Product/Application Description	56	100	89
	Advertisement	6	33	25
	Announcement	12	83	56
	Appeal/Propaganda	1	100	25
	Exam/Worksheet	22	81	90
	Factsheet	80	86	93
	Forum Discussion	38	97	79
	Interview	64	98	97
	Notice	9	89	89
	Resume/CV	100	98	100
	Slides	27	85	92
	Speech Transcript	71	97	96

pronouns in [17]), the overall accuracy of 82.9% is a low percentage. However, as genre classification is a task involving high level conceptual analysis, this seems a reasonable ageement level. Having said this, the agreement within Scientific Research Article is unexpectedly low. There could be at least two reasons for such discord between the labellers. For example, there might have been a misunderstanding which caused poor quality in the initial document retrieval

exercise, or certain genres might inherently be dependent on experience or training and are not clearly recognisable by members of other communities. Upon examination of the documents, it seems to be that both reasons are in play. For instance, numerous forms were labelled as receipts by students, under the impression that receipts which have not been filled are still receipts. Those with a secretarial background did not share this notion. Likewise, some articles on the subject of email were retrieved as samples of the class Email by the students. On the other hand, there was only a single example out of one hundred abstracts collected by students which the secretaries, who are not necessarily academically inclined, agreed as being an abstract. Nevertheless, the results are encouraging in that an 82.9% overall agreement along with the high precision rate of many genres suggest that, even without giving extensive definitions of each genre class, a reasonable agreement is already achieved with common genre terms. It should be mentioned, however, that each secretary's overall accuracy on the agreement data of the other two labellers was also examined and found to be lower at 73.2% and 67.5%.

4.2 Disagreement Analysis

The groups in Table 4 represent cluster of genres for which frequent cross labelling was observed. The groups in Table 4 are not exclusive of other confusion. The table is meant to convey the clusters of the most confused genre classes. It should also be noted that two genres may be included in the same cluster, but the frequency at which one is labelled as the other may not be comparable in both directions. For instance, Manual was often given the label Technical Manual but not vice versa. The confusion between Receipt and Form is due to perceiving a receipt form prior to its completion as a sample of Receipt. The groups in Table 4 suggest that most of the confusion arises within the genre groups (cf. Table 1), which seems to add partial value to our genre schema.

Table 4. Genre cross-labelling cluster groups

Group	Genres
Group A	Book of Fiction, Poetry Book, Fictional Piece, Poems
Group B	Magazine Article, Scientific Research Article, Other Research Article Group C Technical
Report, Business/Operational Report, Miscellaneous Report Group D Posters, Artwork,	
Advertisement	
Group E	Diagram, Graph, Chart
Group F	Form, Receipt
Group G	Handbook, Technical Manual, Manual Group H
List, Catalogue, Raw Data, Table	
Group I	Legal Proceedings, Legal Appeal/Proposal/Order

4.3 Improving the Corpus

Acquiring a representative corpus is difficult ([4]). Part of the reason for this is because representativeness is meaningful only within the context of the task

to be performed. For example, a well known part-of-speech tagger ([9]), trained on the well-designed Penn Treebank Wall Street Journal corpus ([20]), fails to tag instances of He (Helium) in Astronomy articles correctly ([17]) because the training data failed to be representative of astronomy articles - the task do- main. As purposes and domains change, we propose that a well-designed corpus should not emphasise representativeness but be based on the level of annotation, qualifications, and consolidation. Most existing corpora are designed to hold a number of selected categories populated by samples from well-defined sources, upon the agreement of expert knowledge of the categories. Here we would like to propose the construction of a different type of corpus. We set forth the following principles:

- every member of the database must be accompanied by a vector of dimension N (the size of the final genre schema) indicating the number of times each genre was assigned to the item by human labellers, and,
- labellers from a selected number of characterising groups should be employed to label the data, and each instance of a genre assignment should be qualified by the group of the labeller.

The selection of labellers determines the classification standard or the policy one wishes to represent in an automated classification. If the objective is to model genre classification based on common sense, a large number of labellers from a diverse set of backgrounds should be represented. But, if the objective is to design a classifier for specialists of a selected domain, this corpus is likely to prove inadequate for representing the domain population. A corpus built on the above principles would provide us with greater scope for analysis, for achieving representativeness of different populations, and for fine tuning an automated system, by making transparent:

- the confidence level of each item's membership in each genre class, –
the labeller's possible bias by indicating the labeller background, and,
- the fact that classification is not a binary decision (deciding whether or not an item is a sample of a class) but a selection of several probable options.

5 Experiments

5.1 Data

The dataset used in this sections's experiments consists of the data on which all labellers have agreed in the human labelling experiment described in Section 3. The experiment was conducted over only sixteen of the seventy genres presented in Table 1. The range of genres was limited to be more easily comparable to earlier experiments in [16], [18], [19]. The results of experiments on the full range of genres will be available after further analysis of the human experiments have been carried out.

5.2 Classifiers

In [16], we reported results on using the Nave Bayes model to detect instances of Periodicals, Scientific Research Article, Thesis, Business Report, and Forms from a pool of documents belonging to nineteen genres. In this paper we have abandoned the Nave Bayes Model. The Nave Bayes Model was only chosen as a preliminary testing ground as it is one of the most basic probabilistic models available. In reality, Nave Bayes is well known to have problems when dealing with features which are not independent and, in the current context, we want to identify features of one genetic feature type, i.e. features which are dependent on each other, which makes Nave Bayes an inappropriate choice. In its place we have chosen the Random Forest method ([7]), which has been presented as being effective when dealing with imbalanced data ([8]). We have examined two classifiers in this paper:

Image classifier: The first page of the document was sectioned into a sixty- two by sixty-two grid. Each region on the grid is examined for non-white pixels, where non-white pixel is defined to be those of a value less than 245. All regions with non-white pixels are labelled 1, while those which are completely white are labelled 0. The choice of sixty-two to define the size of the grid reflects the fact that the level of granularity seemed to be the coarsest level at which some of the documents were recognisable as belonging to specific genres even by the human eye. The resulting vector was then probabilistically modeled via the Random Forrest Decision method, with nineteen trees using the Weka Machine Learning Toolkit([24]). The motivation for this classifier comes from the recognition that certain genres have more (or less) white space in the first page (e.g. the title page of the book), and that the page is often more strictly formatted (e.g. slides for a conference presentation) to catch the attention of the reader (e.g. the reverse colouring on a magazine cover) and introduce them to the type of document at hand without detailed examination of the content. Note that another advantage of white space analysis is that it is easily applicable to documents of any lan- guage and does not depend heavily on character encoding and the accessibility of content.

Style classifier: From a previously collected data set, the union of all words found in the first page, of 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. The collection of vectors is modeled again via the Random Forrest Decision method with nineteen trees using the Weka toolkit([24]). The feature are different from the classifiers in [16] and [17] which also incorporated the number of words, font sizes and variations. This classifier is intended to capture frequency of words common to all genres as well as words which only appear in some genres. The contention of this paper is that even words which appear in a wide variety of genres may be a significant metric, when the frequency is also taken into consideration. A typical example of its weight is embodied in the fact that forms are less likely to contain as many definite or indefinite articles as theses. The two classifiers were used to predict

Table 5. Image classifier: overall accuracy 38.37%

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

Table 6. Style classifier: overall accuracy 69.96%

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

the genres of documents spanning over sixteen genres. The genres that were examined and the results are given in Section 6.

6 Results

The results of the image classifier in Table 5 do not show the same level of accuracy level as the results previously given in [19]. However, the results in our previous work was of binary classification. As distinctions between larger number of genres have to be made in the current context, it is more likely that any single class resembles another without sharing its identity.

The style classifier (cf. Table 6) shows a surprisingly high level of accuracy on the new data, suggesting that the frequency of words may be a key feature in detecting genres. The prediction of Minutes is particularly noticeable. Parallel to the results in [16] and [19], Periodicals are better recognised by the image classifier than the style classifier. Slides also seem to show the same tendency. As might be expected, genres which depend heavily on the content such as Technical

Report fare much better with the word frequency model. Another observation to be made from the results of Tables 5 and 6 is that the image classifier seems to fare better on a small amount of data (e.g. periodicals, book of fiction).

7 Error Analysis

For a thorough error analysis, a well-designed experimental corpus is required. Until the human labelling experiment in Section 3 is taken forward to include sufficient data and labels from more labellers for in-depth analysis, we can not claim to have the necessary corpus. Nevertheless, many of the errors can already seen to be due to a lack of data (e.g. Book of Fiction), while others seem inexorably linked to the fact that semantic content plays a heavier role than surface styles and structure (e.g. Guideline). An immediately recognisable flaw in the image representation of the document is that it is too strictly dependent on the exact location of non-white space. Ideally, we would like to detect the topol- ogy of the image representation such as the existence of lines, closed loops and other shapes. The location is only loosely relevant. The current representation is too rigid and should be modified to represent the general topology, rather than point-fixed pixel values. Also, more sophisticated linguistic pattern analysis is envisioned to be necessary for the next stage of the stylistic word frequency model.

8 Conclusions

The results in this paper can be summarised by the following:

- Genre classification as an abstract task is ill-defined: there is much disagreement even between human labellers and a detailed study of further human experiments are required to determine conditions which make the task meaningful.
- A fair amount of automated genre classification can be achieved by examining the frequency of genre words.
- The image of the first page alone seems to perform better classification than style, when only a small amount of training data is availabble.
- The performance of the image classifier appears to complement the performance of the style classifier.

It is evident that research in this area is still in its infancy. There is much to do. As we have noted elsewhere, the other classifiers based on language modeling, semantic analysis and domain knowledge should be tested for further comparison. Furthermore, proper error analysis and further gathering of documents and human labelling analysis is required to establish a well designed corpus. To maximise sustainability in an environment where technology changes at a rapid rate, the technical format information (e.g. PDF specification, or metadata extracted by pdfinfo) should only be included in the extraction tool algorithm at the last

stage of improvement. The classification of documents into a small number of types has its limits. To be able to utilise these classifiers constructed under different conditions in the larger context of information management, we need to be able to construct systems that can group classifiers into clusters of similar tasks, or more specifically, into clusters of co-dependent classifiers.

Acknowledgments

This research is collaborative. DELOS: Network of Excellence on Digital Libraries (G038-507618)[3] funded under the European Commissions IST 6th Framework Programme provides a key framework and support, as does the UK's Digital Curation Centre. The DCC[4] is supported by a grant from the Joint Information Systems Committee (JISC)[5] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)[6]. The EPSRC supports (GR/T07374/01) the DCCs research programme. We would also like to thank Andrew McHugh, Adam Rusbridge, and Laura Brouard, who organised the web support and supervised the administrative process for the human labelling experiments.

References

1. Bagdanov, A., Worring, M.: Fine-grained document genre classification using first order random graphs. In: Proceedings 6th International Conference on Document Analysis and Recognition, pp. 79–83 (2001) ISBN 0-7695-1263-1
2. Barbu, E., Heroux, P., Adam, S., Turpin, E.: Clustering document images using a bag of symbols representation. In: Proceedings 8th International Conference on Document Analysis and Recognition, pp. 1216-1220 (2005) ISBN ISSN 1520-5263
3. Bekkerman, R., McCallum, A., Huang, G.: Automatic categorization of email into folders. benchmark experiments on enron and sri corpora. In: Bekkerman, R., McCallum, A., Huang, G. (eds.) Technical Report IR-418, Centre for Intelligent Information Retrieval, UMASS (2004)
4. Biber, D.: Representativeness in Corpus Design. Literary and Linguistic Comput- ing 8(4), 243–257 (1993)
5. Biber, D.: Dimensions of Register Variation:a Cross-Linguistic Comparison. Cambridge University Press, New York (1995)
6. Boese, E.S.: Stereotyping the web: genre classification of web documents. Master's thesis, Colorado State University (2005)
7. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
8. Chao, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data (2004), http://www.stat.berkeley.edu/~ breiman/RandomForests/
9. Curran, J., Clark, S.: Investigating GIS and Smoothing for Maximum Entropy Taggers. In: Proceedings Aunnual Meeting European Chapter of the Assoc. of Computational Linguistics, pp. 91–98 (2003)

[3] http://www.delos.info
[4] http://www.dcc.ac.uk
[5] http://www.jisc.ac.uk
[6] http://www.epsrc.ac.uk

10. Finn, A., Kushmerick, N.: Learning to classify documents according to genre. Jour- nal of American Society for Information Science and Technology 57(11), 1506–1518 (2006)

11. Giuffrida, G., Shek, E., Yang, J.: Knowledge-based metadata extraction from postscript file. In: Proceedings 5th ACM Intl. Conf. Digital Libraries, pp. 77–84. ACM Press, New York (2000)
12. Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic doc- ument metadata extraction using support vector machines. In: 3rd ACM/IEEECS Conf. Digital Libraries, pp. 37–48 (2003)
13. Karlgren, J., Cutting, D.: Recognizing text genres with simple metric using dis- criminant analysis. Proceedings 15th Conf. Comp. Ling. 2, 1071–1075 (1994)
14. Ke, S.W., Bowerman, C.: Perc: A personal email classifier. In: Lalmas, M., Mac- Farlane, A., R罕ger, S., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 460–463. Springer, Heidelberg (2006)
15. Kessler, G., Nunberg, B., Schuetze, H.: Automatic detection of text genre. In: Proceedings 35th Ann., pp. 32–38 (1997)
16. Kim, Y., Ross, S.: Genre classification in automated ingest and appraisal metadata. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 63–74. Springer, Heidelberg (2006)
17. Kim, Y., Webber, B.: Implicit reference to citations: A study of astronomy pa- pers. Presentation at the 20th CODATA international Conference, Beijing, China. http://eprints.erpanet.org/ paper id 115 (2006), http://eprints.erpanet.org/paperid115
18. Kim, Y., Ross, S.: Detecting family resemblance: Automated genre classification. Data Science 6, S172–S183 (2007), http://www.jstage.jst.go.jp/article/dsj/6/0/s172/ pdf
19. Kim, Y., Ross, S.: The Naming of Cats: Automated genre classification. Interna- tional Journal for Digital Curation 2(1) (2007), http://www.ijdc.net/./ijdc/article/view/24
20. Marcus, M.P., Santorini, B., Mareinkiewicz, M.A.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2), 313–330 (1994)
21. Rauber, A., M罕ller-K矣gler, A.: Integrating automatic genre analysis into digital libraries. In: Proceedings ACM/IEEE Joint Conf. Digital Libraries, Roanoke, VA, pp. 1–10 (2001)
22. Ross, S., Hedstrom, M.: Preservation research and sustainable digital libraries. International Journal of Digital Libraries, (2005) DOI: 10.1007/s00799-004-0099-3
23. Thoma, G.: Automating the production of bibliographic records. Technical report, Lister Hill National Center for Biomedical Communication, US National Library of Medicine (2001)
24. Witten, H.I., Frank, E.: Data mining: Practical machine learning tools and tech- niques. Morgan Kaufmann, San Francisco (2005)