LDV Forum manuscript No.
(will be inserted by the editor)

Formulating representative features with respect to document genre classification

Yunhyong Kim · Seamus Ross

Received: date / Accepted: date

Abstract Genre classification (e.g. whether a document is a scientific article or magazine article) is closely bound to the physical and conceptual structure of document as well as the level of depth involved in the text. Hence, it provides a means of ranking documents retrieved by search tools according to metrics other than topical similarity. Moreover, the structural information derived from genre classification can be used to locate target information within the text. In previous studies, the detection of genre classes has been attempted by using some normalised frequency of terms or combinations of terms in the document (here, we are using term as a reference to words, phrases, syntactic units, sentences and paragraphs, as well as other patterns derived from deeper linguistic or semantic analysis). These approaches largely neglect how the term is distributed throughout the document. Here, we report the results of automated experiments based on distributive statistics of words in order to present evidence that term distribution pattern is a better indicator of genre class than term frequency.

Keywords document classification · genre · document representation · word distribution

Y. Kim
HATII, 11 University Gardens, Glasgow, G12 8QH, UK
Tel.: +44-141-3308594
Fax: +44-141-3303788
E-mail: y.kim@hatii.arts.gla.ac.uk

S. Ross
Same institute as above
Tel.: +44-141-3303635
Fax: +44-141-3303788
E-mail: s.ross@hatii.arts.gla.ac.uk

1 Introduction

This paper examines the role of word distribution in classifying documents according to genre. Document classification is one of the most fundamental steps in enabling the search, selection, and ranking of digital material according to its relevance in answering a predefined search. As such it is a valuable means of knowledge discovery and an essential part of the effective and efficient management of digital documents in a repository, library, or archive. Document classification has previously been dominated by the classification of documents according to topic. Recently, however, there has been a growing interest in the classification of documents with respect to factors other than topic, for example, genre (e.g. scientific papers, emails and news reports). The interest in the genre of documents reflects the limitations of relevance measurements based on topic. Topic alone does not provide insight into whether or not a retrieved document is appropriate for your purpose; a document with the same topic may be created with different objectives resulting in different levels of usefulness as a source of information (e.g. compare an advertisement about a camera to a product review of the same camera). The objectives of document creation define the functional requirements of the document (e.g. to narrate, to argue against, to argue for, to present research results) that characterise its genre, and the structures found within the document are designed to meet these functional requirements. Therefore, the structural classification of documents is a fundamental component in detecting genre. Classical models of document classification largely depend on term frequency weighting and counting instances of specified linguistic constructs. The former does not reflect much document structure and the latter results in a highly language dependent

model that incorporates some local conceptual structure but largely disregards the global conceptual or physical structure of the document and its components. In this paper,

- we describe an approach to document representation that incorporates more document structure by considering how strings are distributed throughout the document (Section 2.2), and,
- give evidence that this approach is better than the bag-of-words approach by comparing it against the rainbow classifier [18] (Section 5.2).

Being able to bind together tools trained to retrieve information within selected structural domains is crucial to automating the ingest, management and preservation of material in digital repositories [20]. This is especially true where metadata describing the technical characteristics, function, source and content of digital material play a core role in the efficient and effective management and re-use of the same (cf. [20]). As we have discussed in earlier papers (e.g. [25]), the manual collection of metadata is labour-intensive, costly and susceptible to variation in quality and precision across different actors; automating the process of semantic metadata extraction is, therefore, essential. Past efforts (e.g. [3], [6], [10], [11], [13], [22]) to extract metadata automatically from digital documents have relied heavily on the structure that characterises the genre class to which the document under consideration belongs. The reliance of these methods on document structure emphasises the benefits of constructing a tool that enables automated genre classification. An effective automated genre classifier would function as an overarching tool for integrating genre-specific tools and, in any case, provide a first-level classification of documents into those of a similar structure, which would facilitate the extraction of further information.

The vast number of different contexts in which genre have emerged across classification attempts illustrate that genre is a high-level, context-dependent concept (cf. literature review in [21]). Genre has been referred to as aspects of the text described by level of information or degree of elaboration, persuasion and abstraction [4], as well as, to common document forms such as FAQ, Job Description, Editorial or Reportage [12], [14]. In some cases, genre has been used to describe the classification of a document according to whether or not it is a narrative and whether it is intended for an audience of specialists [14], and whether it is fact or opinion, and, in the case of opinion, whether it is positive or negative [9]. On occasion it has been used to describe membership to selected journals and brochures [1], and, to denote similar feature cluster groups ([2], [19]).

A prevailing notion in earlier analyses is that genre classification is a task independent from subject classification. While this may be true on a conceptual level, there is reason to believe that this may not be a statistically sound approach. For example, the topic of algebraic variety, a well-known subject area in higher mathematics, would not be expected to appear as frequently in the genre class Reportage as it would in the genre class Research Article. In fact, preliminary results from a recent experiment, classifying documents belonging to ten genre classes into twenty newsgroup topic classes, shows that, while there are genre classes whose documents are randomly distributed across the twenty topics (e.g. Poem), there are also genres 95% of whose documents are classified into only four newsgroup topics (e.g. Minutes).

Given these examples where genre is interactively intertwined with topic, it would seem beneficial to build a general classification model that encompasses both tasks. With this in mind, in Section 2, we would like to introduce genre classification, not as a classification task distinct from topic classification, but as a point in a continuum of classifications, emphasising both genre classification and topic classification as a special case of a general abstract classification model.

In Section 4, we introduce the dataset and investigate the agreement between human labellers in classifying the data, a discussion that demonstrates the complexity of genre classification and leads to a measurement of the cleanliness of the dataset which helps to set the standards against which to compare the automated experiments. We further compare the support vector machine classification based on the document representation introduced in Section 2.2 to the baseline support vector machine rainbow classifier [18]. This comparison, using our abstract classification model, will show its effectiveness in performing automated classification.

The discussion here is a result of examining PDF documents. The study was limited to PDF documents because of the popularity this format has across library, archival, commercial and private sectors. This popularity implies that a classification tool developed for this format is likely to have widespread immediate usefulness. 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 text.

2 Defining genre classification

2.1 Document representation in conventional text classification

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

$$TF \times IDF \times N \quad (1)$$

where TF denotes the frequency of the term in the document, IDF denotes the number of documents in the collection containing the term, and N denotes a normalisation factor dependent on the length of the document. The calculation method of each of these terms differs according to the research or application in question. This model is based on the notion that: if a term appears frequently in a document, it is likely to be a characterising feature of the document; if a term appears across several documents, then it is not likely to be a strong feature in distinguishing any one of those documents from the others; and if the same term appears in equal numbers within a short document and a long one, then it is likely to be a stronger feature of the short document. While it may be considered a gross simplification to represent all the various classification methods by this one description, it still seems true that the basic principles that drive various text classification methods are closely related to this model. In a subject classification task, the term may surface as words or N-grams (N consecutive words or characters), while in other classification tasks term may manifest itself also as functional groups of words (e.g. verb) or combinations of such words and phrases and groups. Nevertheless, the mechanism driving the classification is largely dependent on counting patterns, and weighing the number against the pattern count throughout the collection being examined. The location of patterns, the relationship between instances of the patterns, and the interplay between different types of patterns are largely bypassed and only represented implicitly through the same counting process.

2.2 Harmonic descriptor representation (HDR) of documents

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

Because of its static appearance, a document is often misunderstood to be time independent, but the interpretation of each symbol is possible only as a consequence of its temporal relationship to other symbols. In this light, document classification can be considered to be a subtask of signal processing. Viewed in this way, an accurate measure of term frequency is expressed by how many times a symbol occurs with respect to time. The term weight calculated in Section 2.1 presents no awareness of the role of temporal progression in the semantic analysis of the document. That is, if the word "clock" were to appear in two documents ten times, then the weight of this word would be equal with respect to both documents: the fact that the word appears only in the first half of the document with respect to one of the documents in contrast to being evenly distributed throughout the document (which may be the case with respect to the other document) would be disregarded. A proper consideration of the time dimension would suggest "clock" in the first document as a signal having twice the frequency of that of the second document, but lasting only half the length of time. Time should not be taken to be the length of the text. Although the two are closely related, the length of the text is not equivalent to the tempo of the piece of writing, beginning with an introduction and ending with a conclusion. To understand the notion of time, we will compare a document to a string of a musical instrument. An occurrence of a symbol within the document partitions the document into two parts. If the two partitions are equal in length, then the phase division is akin to a harmonic with twice the frequency of the fundamental of the string (the document with zero occurrence of the symbol). If the division is not equal, then the frequency can not be considered to be uniform throughout the document.

In the case of topic detection, a loose application of time (e.g. taking the frequency to be uniform throughout the document) may be sufficient to capture salient vocabulary, but in other types of classification, where the main interest lies in the physical or conceptual structure of the object, the lack of temporal and relational placement of symbols contributes to a considerable loss of information. To fill this gap, we propose incorporating the symbols range and period as an effective means of characterising the symbol with respect to document structure. We define range as the interval between the initial and ultimate occurrence of the symbol, and period as the time duration between two consecutive occurrences of the symbol. When the symbol occurs at regular intervals, the resulting signal in the document

is akin to a harmonic of the document as a wave. Brookstein, Klein and Raita [5] observed that content-bearing words would clump together and therefore result in non-harmonic behaviour. In contrast to the content-bearing words that they discuss, our research focuses on words that may be indicative of style and structure. We observe that document structure is captured by words displaying both harmonic and non-harmonic behaviour; harmonic words define the physical structure of the document, while non-harmonic words define conceptual landmarks or structure. In our description, we attempt to capture the degree of non-harmonic behaviour using three quantities derived from the range and period of each symbol:

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

The more harmonic the behaviour of a symbol, the closer AP will be to 1. In Figure 1, we display an example of six documents (D1 to D6) of different lengths, portrayed as blue strips where the top of the strip is the beginning of the document. Occurrences of symbols in the documents (s1 to s7) have been represented as horizontal lines across the strips. The period between two consecutive occurrences have been indicated to be x. This example will be used in Figures 2, 3, and 4 to demonstrate how FP, LP, and AP change under different conditions.

We present in Figure 2, a graph illustrating how FP, LP and AP change as the position of a symbol occurring once in $D1$ (see Figure 1) changes from $s1$ to $s7$. In Figure 3, we show how FP, LP and AP for a symbol occurring twice in D1 change with respect to the period between the two instances, as the second occurrence of

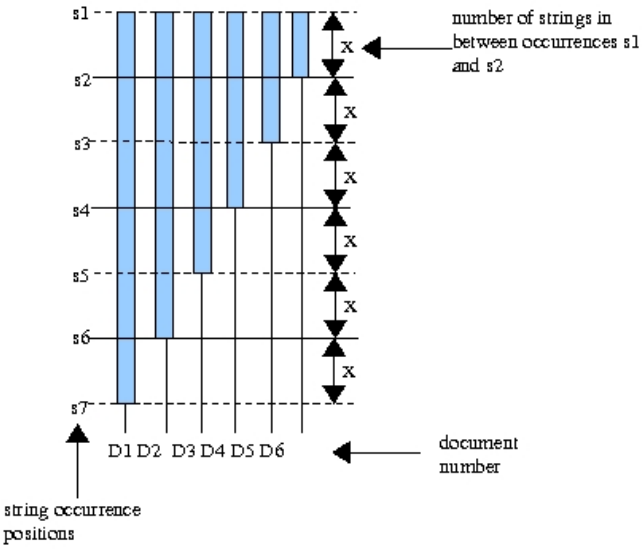

Fig. 1 Example of symbol occurrence in six documents of different lengths.

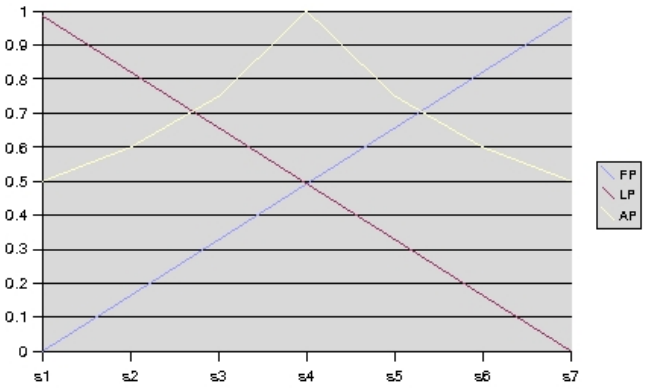

Fig. 2 FP, LP, and AP with respect to the position (X-axis) of a single occurrence of a symbol in D1.

the symbol moves away from the first occurrence. Finally, the graph in Figure 4 presents how FP, LP and AP, for a symbol occurring once halfway between $s1$ and $s2$, change as the document length varies.

Given a document, each word or symbol in the document is associated to their FP, LP and AP values. By taking all the words in a collection or by using a pre-compiled list of indicative words (say, in either case, the resulting word list is of size N), each document can be represented as a vector of dimension 3N, where each term in the vector is the FP, LP, or AP value of each word. In our model we pre-compiled a list of words from a sample dataset (which is discarded from the test dataset after the words are collected) by aggregating a list of words that appear in 75% of all the documents in at least one genre class in the sample dataset.

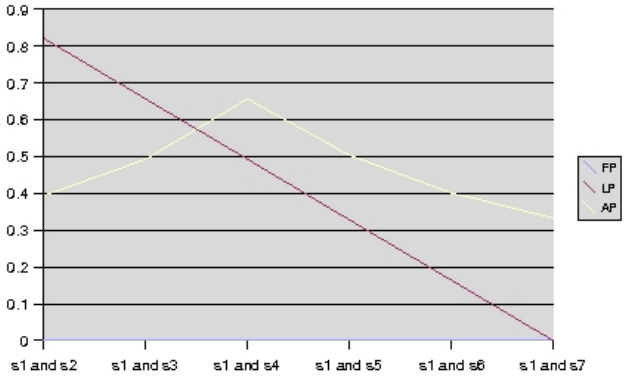

Fig. 3 FP, LP, and AP for a symbol occurring twice in D1 as the period between the two instances become larger.

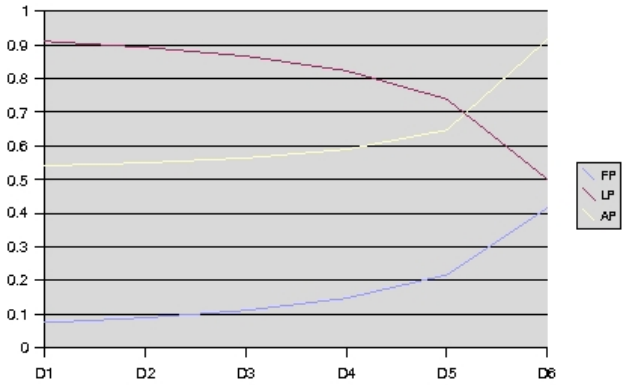

Fig. 4 FP, LP, and AP for a symbol occurring once in the same position relative to the beginning of different length documents.

The relevance of term distribution has been mentioned by others including Manning et al. [17], and, more recently, by De Roeck et al. [7] who carried out a study of profiling datasets to determine the degree of homogeneity or heterogeneity in the distribution of frequent terms. However, there have only been few explicit implementations of the measurement for the purpose of automated classification, and most of these previous analyses have been based on a count of words in selected chunks of the texts. The model presented here compares relative distances between term instances. The latter approach views the entire document as a time dependent whole, and does not involve arbitrary choices of chunk sizes.

2.3 Genre classification

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 fulfil its functional intention(s) (e.g. to describe, to inform and to argue, to advertise) within its target environment (e.g. the user community, publisher and creator), much the same as the structure of an organism evolves to optimise its survival function in the natural environment (cf. [16]). As a consequence, genre reflects one or more of the following:

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

The model described in Section 2.1, while effective in distinguishing some intentional and interpretive aspects of genre, seems insufficient to capture distinguishing features in the case of prescriptive, conceptual or physical structure. Such structure can be characterised even by low frequency terms of the class (e.g. single occurrence of "minutes" in the title of meeting minutes, or headings in a curriculum vitae), and the distributional pattern of words throughout the document (variation of density) is often bound to its class (e.g. the even distribution of wh-words in a FAQ sheet). The last observation is a generalisation of the observation by Brookstein, Klein and Raita [5], who noted the clumping properties of content-bearing words and their role in text classification. In contrast to the content-bearing words that they discuss, we are interested also in words indicative of style and structure. These words can exhibit both clumping and uniform distribution properties. We present evidence that documents of each genre class display distinctive distributional characteristics.

Here we have adopted the genre schema of seventy classes (KRYS I corpus)introduced in Kim and Ross [15], [16], constructed to represent these aspects from different perspectives, as well as Santini's data set ([21]) consisting of seven webpage classes. We use twenty four classes from KRYS I and Santini's dataset as a sample testbed, altogether consisting of 3,452 documents in thirty-one genres, to test the harmonic descriptor representation of documents described above. The test was initially confined to thirty-one genres in order to limit the computation time. The twenty-four classes from KRYS I were selected more or less at random

Table 1 Number of words found in seven out of ten documents belonging to three genres (top row) with respect to word type (left column). Median length of documents in each genre are expressed in the parentheses next to the genre label as number of bytes.

	Poem (1718)	Letter (4265)	Thesis (132993)
Article	2	2	3
Wh-word	0	0	6
Modal	0	1	9
Have Verb	0	1	3
Be Verb	1	3	7
Verb	0	0	29
Noun	0	0	46
Subject Pronoun	2	1	4
Object Pronoun	0	0	1
Possessive Pronoun	0	0	0
Possessive Adjective	0	0	2
Adjective	1	1	43
Adverb	0	1	29
Quantifier	0	1	9
Demonstrative	1	2	6
Conjunction	1	3	9
Preposition	5	8	20
Punctuation	2	3	4
Other	1	1	12

apart from an effort to select a porportion of classes from each of the ten genre groups presented in [16].

In Section 5, we will compare support vector machine (SVM) classification using the harmonic descriptor representation of documents (this is modelled using Weka machine learning software [24]) against the SVM classification performed using the Bow Toolkit rainbow text classifier developed by MacCallum [18], and the classification attempts of Santini [21], to show that the performance is consistently better when using the new description. The reason we have selected SVM as the classification method is that it showed the best results for rainbow when compared with Rocchio/TFIDF and Naive Bayes.

The symbols (selected to be words in the experiments here) we will examine with respect to SVM HDR in the experiments were compiled by examining a sample dataset, a small slice of the corpus set aside, for words that appear in a large number of documents (but not necessarily frequently in any one document) in each genre. The list is intended to represent a set of words prolific within at least one genre in the collection. The words collected are expected to include stop words and hmtl tags. As a sample, we present the number of words found to be prolific in the genre classes Poem, Letter and Thesis with respect to word type (WT), after examining ten random documents in each class (Table 1). Most of the numbers in Table 1 are not very illuminating by itself in that the median lengths of documents belonging to Poem, Letter and Thesis are 1718, 4265, and 132994, respectively (in bytes), that is, we expect the numbers to be increasing in that oder for each type of word. However, we immediately notice an exception in this pattern with respect to subject pronouns, and, closer examination of the actual words show that at least one of the two subject pronouns found to be prolific in poems (i.e. "you" and "I") is not found to be as prolific in letters (i.e. "it") and theses (i.e. "I", "we", "they", "it"). Further, the word "Dear" is only found to be prolific within letters.

To illustrate how the FP, LP and AP of the HDR description varies across documents of the same genre we present a snapshot of these values with respect to the word "whose" across 90 poems, 100 theses, 91 letters and 91 technical reports in Figure 5. The segments corresponding to the documents belonging each genre are indicated at the bottom of the figure. The figure shows that FP, LP, and AP are similar for documents belonging to the same genre but diverge as we move across documents belonging to different genres.

3 Classifiers

We used two different classifiers in the experiments described in Section 5: the support vector machine (SVM) rainbow text classifier [18] and the SVM harmonic descriptor representation (HDR) classifier modelled using the Weka machine learning toolkit [24]. The rainbow text classifier, included in the BOW toolkit developed by Andrew McCallum [18], 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 SVM, which has been proven to be effective in other text classification tasks [23]. As we mentioned at the end of Section 2.2, in our model, the HDR uses a pre-compiled list of words. For the experiments reported in Section 5, we set aside ten random documents from each of the genres in the dataset and collected all the words that appear in more than 75% of the documents in each genre. This list consists of 2,477 words.

4 Datasets

A comparison of automated classification methods on a dataset that has not been tested for human agreement can give misleading information as human agreement analysis conveys to us how clean the dataset is and the nature of the genre class schema of the dataset. The experiments reported here were carried out on a collection consisting of the genres in Table 2 (numbers of documents in each genre, excluding those used to construct the word list in the previous section, are indicated in

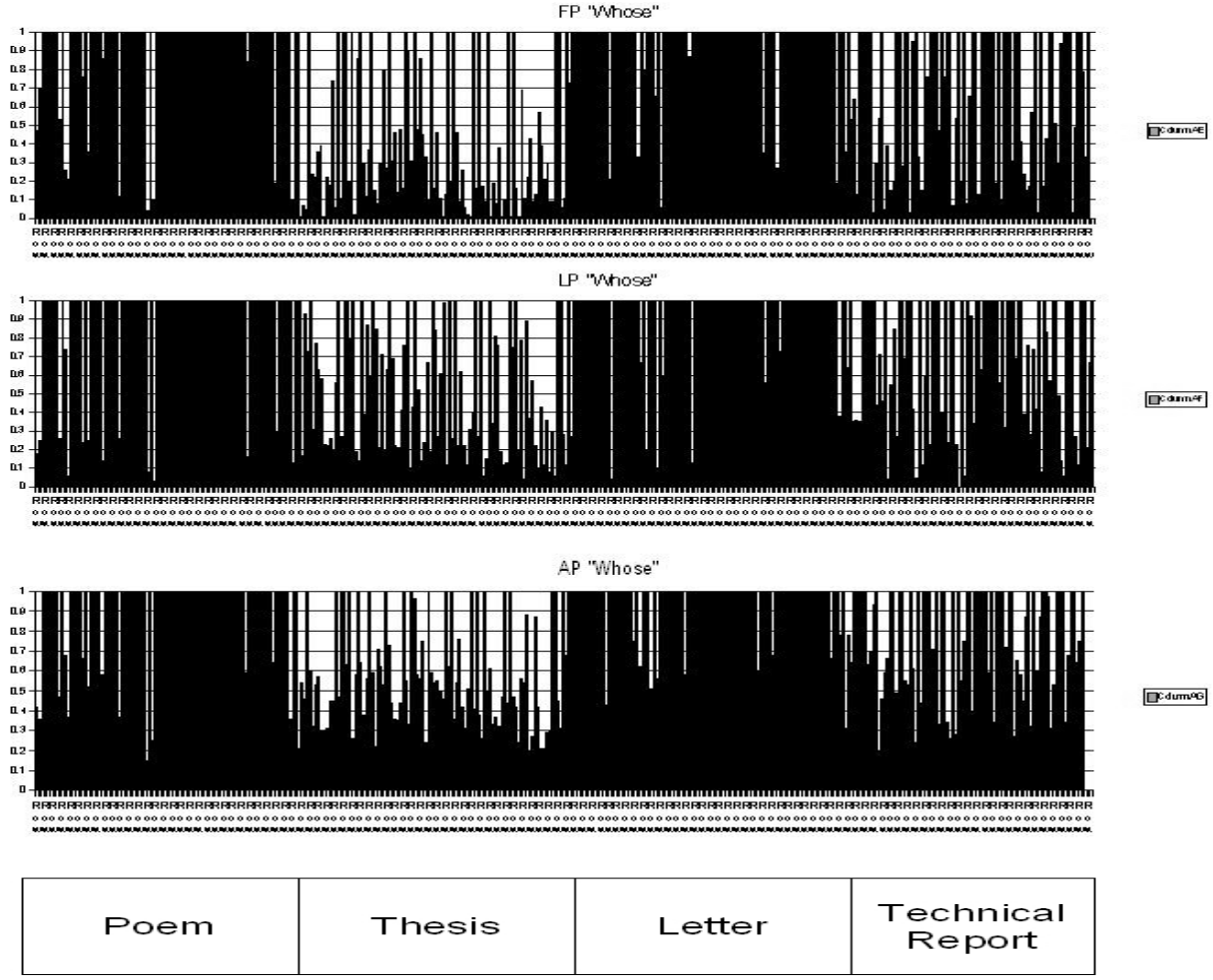

Fig. 5 Example of FP (top), LP (middle), and AP (bottom) values with respect to the word "whose" across documents belonging to four distinct genres (the documents corresponding to each of these genres are noted by segmentation indicated at the bottom of the figure).

parentheses). The dataset for the twenty-four document genres were collected by:

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

None of the labellers were given a definition for the genres in the schema. This was partly to establish whether there was already a well understood genre vocabulary. The human performance was examined by taking the number of labels given by a single labeller in agreement with the other two labellers over the total number of documents on which the other two labellers agreed. The three numbers obtained in this way are 0.675, 0.73 and 0.829. Although the difference between the lowest and the highest recall is a noticeable 14 per cent, this should be viewed with the knowledge that the highest recall is the result of student classification while the lowest recall is that of secretary classification.

The dataset for the seven webpage genres was obtained from Santini's collection available at her website([21]).

5 Results

The performance will be evaluated using one or more of three conventional metrics: accuracy, precision and recall. To re-visit the definition for these terms, let N be the total number of documents in the test data, N_c

Table 2 Scope of genres

Creative	Book of Fiction(29) Poem(90)
Determined by user context	Email(90) Exam/Worksheet (90) Form (90) Handbook (90) Letter (91) Minutes (99) Resumé/CV Sheet Music (90) Speech Transcript (91) Technical Manual (90)
Determined by organisational prescription	Abstract (89) Academic Monograph (99) Advertisement (90) Business Report (100) Magazine Article (90) Scientific Article (90) Memo (90) Periodicals (67) Poster (90) Slides (90) Technical Report (91) Thesis (100)
Webpage genres	Blog (190) Eshop (190) FAQ (190) Front Page (190) List (190) Personal Home Page (190) Search Page (190)

the number of documents in the class C, $TP(C)$ the number of documents correctly predicted to be a member of class C, and $FP(C)$ the number of documents incorrectly predicted as belonging to class C. Accuracy, A, is defined to be:

$$A = \frac{\sum TP(C)}{N}, \tag{2}$$

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

$$P(C) = \frac{TP(C)}{TP(C) + FP(C)}, \tag{3}$$

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

$$R(C) = \frac{TP(C)}{N_c}. \tag{4}$$

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

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

5.1 Overall accuracy

The figures in Table 3 are the overall accuracies of the support vector machine rainbow classifier (SVM rainbow), the support vector HDR classifier (SVM HDR) and average human agreement. The classifier we are considering to be a baseline classifier in this comparison is the SVM rainbow classifier. The human agreement is included to indicate the cleanliness level of the dataset being used.

Table 3 Overall accuracy across all the genre classes.

Classifier	SVM rainbow	SVM HDR	Human
Overall accuracy	0.73	0.80	0.74

The numbers in Table 3 suggest that the performance level of the SVM rainbow classifier is already comparable to the average performance of three human labellers, and shows that the SVM HDR improves on the SVM rainbow classifier by 7%.

To test the limits on a cleaner dataset, we analysed the classification results with respect to Santini's webpage corpus. This is the overall accuracy of the classification when the recall of the documents belonging to the webpage genre classes is calculated upon the classification of the entire dataset into thirty-one classes. There is a slight increase of 0.002 when the webpage classes are classified on their own. The results are shown in Table 4: the numbers suggest that SVM HDR is a strong contender in webpage genre classification.

Table 4 Overall accuracy of classifiers across webpage genres (Blog, Personal Home Page, FAQ, List, Search Page, EShop, Front Page).

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

5.2 Precision and recall

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

In Figures 6 and 7, we present the recall and precision of SVM rainbow and SVM HDR with respect to each of our classes. The graphs show that SVM HDR outperforms SVM rainbow with respect to most of the classes in both recall and precision. The recall

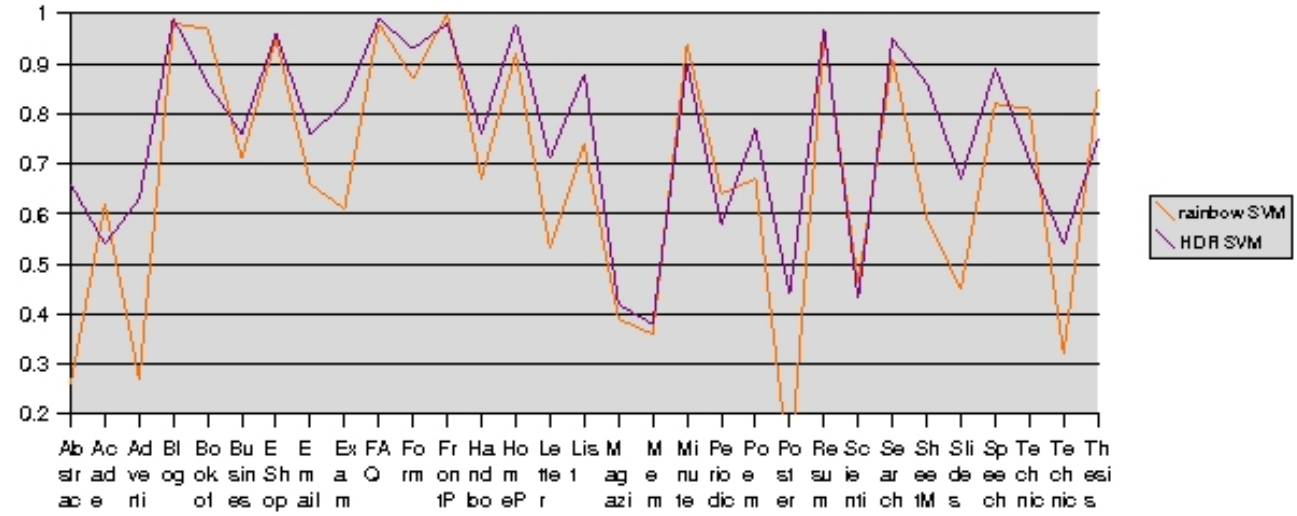

Fig. 6 Recall: a comparison, SVM rainbow and SVM HDR.

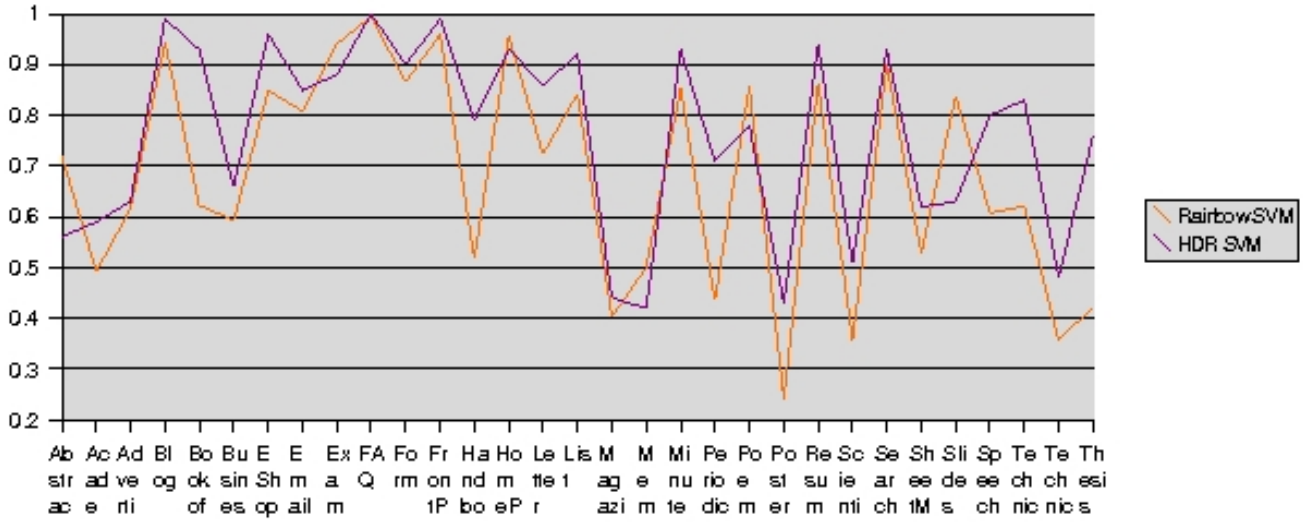

Fig. 7 Precision: a comparison, SVM rainbow and SVM HDR.

of SVM rainbow with respect to Academic Monograph, Book of Fiction, Front Page (of a website), Minutes, periodicals, Technical Manual and Thesis is Marginally higher than SVM HDR and the precision of SVM rainbow with respect to Abstract, Exam/Worksheet, Home Page, Poem, and Slides is somewhat higher than that of SVM HDR. However, with respect to the majority of the classes, SVM HDR outperforms SVM rainbow.

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

The graph in Figure 8 presents the F-measures of SVM rainbow and SVM HDR with respect to each class. This graph shows that the F-measures of SVM HDR are greater than those of SVM rainbow with respect to every class except the class Memo. With respect to Memo, the difference is 0.02 in favour of SVM

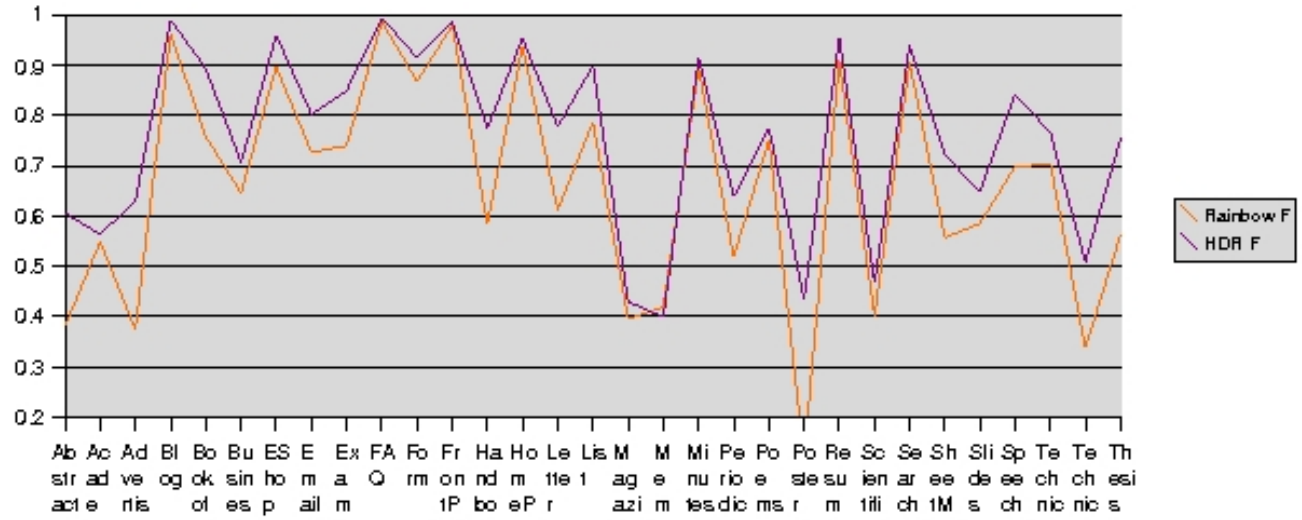

Fig. 8 F-measure: a comparison, SVM rainbow and SVM HDR.

rainbow. Latest experiments using HDR to analyse a newgroup dataset of 19597 documents in twenty topical classes (obtained from McCallum's website1), show that the same SVM HDR model is also promising in topic classification, with an overall accuracy of over 95% (detailed report of this experiment available shortly). A list of eighty-two words was compiled from 400 documents (20 documents from each genre) set aside from the original 19997 documents for this experiment. We have also calculated the F-measures of SVM HDR with respect to the classes in this dataset to find them all greater than the best results (overall accuracy 93.7%) of the rainbow classifier. The details of this experiment will be published shortly.

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

6 Conclusions

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

The results with respect to Santini's dataset present evidence that SVM HDR might be superior to classifiers that rely on counts of terms or patterns. This conjecture is again supported by comparing the result with a recent report by Dong et al. [8] on the classification of Santini's data belonging to four genres (best overall accuracy approximately 96 per cent). Although their numbers are similar to ours, it must be noted that the accuracy presented in our paper is that obtained from a classification across seven webpage genres not classified in isolation but classified when accompanied by a classification of twenty-four additional document genres.

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

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

- examining the resolution mentioned at the end of Section 5.2;
- looking at, and comparing, other forms of symbols apart from words; and
- considering ways in which the two approaches might be integrated

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

Acknowledgements The work presented in this paper was supported by DELOS: Network of Excellence on Digital Libraries[1] (G038-507618), funded under the European Commissions IST Sixth Framework Programme,and the UK's Digital Curation Centre (DCC)[2], funded by the Joint Information Systems Committee (JISC)[3] and the e-Science Core Programme of the Engineering and Physical Sciences Research Council (EPSRC)[4][GR/T07374/01].

References

1. Bagdanov, A. and Worring, M. Fine-grained document genre classification using first order random graphs. In Proceedings Sixth International Conference on Document Analysis and Recognition (ICDAR2001), 79-90. (2001)
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, 1216-1220.(2005)
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, Center 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. Bookstein, A., Klein, S.T. and Raita, T. Clumping properties of content-bearing words. Journal of the American Society of Information Science, 1998, 49(2): 102-114.(1998)
6. dc-dot, UKOLN Dublin Core metadata editor. http://www.ukoln.ac.uk/metadata/dcdot/
7. De Roeck, A., Sarkar, A. and Garthwaite, P. Frequent Term Distribution Measures for Dataset Profiling. Technical Report 2004/06. Faculty of Mathematics and Computing, Open University. Milton Keynes, UK. (2004) http://computing-reports.open.ac.uk/index.php/
8. Dong, L., Watters, C., Duffy, J. and Shepherd, M. An Examination of Genre Attributes for Web Page Classification. In Proceedings 41st Hawaiian International Conference on System Sciences, IEEE Computer Society Press, ISBN-13: 978-0-7695-3075-8, ISBN-10: 0-7695-3075-3, ISSN: 1530-1605. (2008)
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 International Conference on Digital Libraries, 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 Conference on Digital Libraries, 37-48. (2003)
12. Karlgren, J. and Cutting, D. Recognizing text genres with simple metric using discriminant analysis. In Proceedings 15th Conference on Computational Linguistics, 2: 1071-1075. (1994)
13. Ke, S.W. and Bowerman, C. Perc: A personal email classifier. In Proceedings 28th European Conference on Information Retrieval (ECIR 2006), 460-463. (2006)
14. Kessler, G., Nunberg, B. and Schuetze, H. 1997. Automatic detection of text genre. In Proceedings 35th Annual Meeting ACL, 32-38.
15. Kim, Y. and Ross, S. Detecting family resemblance: Automated genre classification. CODATA Data Science Journal, 6: S172-S183. ISSN: 1683-1470. (2007)
16. Kim, Y. and Ross, S. Searching for Ground truth: a stepping stone in automated genre classification. LNCS 4877, 248-261, Springer. DOI: 10.1007/978-3-540-77088-6 (2007) http://www.springerlink.com/content/lt760613m2731723/fulltext.pdf
17. Manning, C. and Schutze, H. Foundations of Statistical Language Processing, MIT Press. Cambridge, MA. (1999)
18. McCallum,A. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. (1996) http://www.cs.cmu.edu/~1mccallum/bow
19. Rauber, A. and Muller-Kogler, A. Integrating automatic genre analysis into digital libraries. In Proceedings ACM/IEEE Joint Conference on Digital Libraries, 1-10, Roanoke, VA. (2001) http://doi.acm.org/10.1145/379437.379439
20. Ross, S. and Hedstrom, M. Preservation research and sustainable digital libraries. International Journal of Digital Libraries, v 5.4, 317-325. DOI: 10.1007/s00799-004-0099-3. (2005) http://eprints.erpanet.org/archive/00000095/
21. Santini, M. PhD thesis, University of Brighton, Brighton (UK). (2007) http://www.itri.brighton.ac.uk/~Marina.Santini/MSantini_PhD_Thesis.zip
22. 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
23. Yang, Y., Zhang, J. and Kisiel, B. A scalability analysis of classifiers in text categorization. In Proceedings 26th annual international ACM SIGIR conference on research and development information retrieval, 96-103. ISBN: 1-58113-646-3, 96-103. (2003)
24. Witten, H.I. and Frank, E. Data mining: Practical machine learning tools and techniques. 2nd edition, Morgan Kaufmann, San Francisco. (2005)
25. Details omitted to anonymise paper.

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