Recognizing Text Genres with Simple Metrics Using Discriminant Analysis

Jussi Karlgren
jussi@sics.se
Swedish Institute of Computer Science
Box 1263, S – 164 28 Kista, Stockholm, Sweden

Douglass Cutting
cutting@apple.com
Apple Computer
Cupertino, CA 95014, USA

Abstract

A simple method for categorizing texts into pre-determined text genre categories using the statistical standard technique of discriminant analysis is demonstrated with application to the Brown corpus. Discriminant analysis makes it possible use a large number of parameters that may be specific for a certain corpus or information stream, and combine them into a small number of functions, with the parameters weighted on basis of how useful they are for discriminating text genres. An application to information retrieval is discussed.

Text Types

There are different types of text. Texts "about" the same thing may be in differing genres, of different types, and of varying quality. Texts vary along several parameters, all relevant for the general information retrieval problem of matching reader needs and texts. Given this variation, in a text retrieval context the problems are (i) identifying genres, and (ii) choosing criteria to cluster texts of the same genre, with predictable precision and recall. This should not be confused with the issue of identifying topics, and choosing criteria that discriminate one topic from another. Although not orthogonal to genre-dependent variation, the variation that relates directly to content and topic is along other dimensions. Naturally, there is co-variance. Texts about certain topics may only occur in certain genres, and texts in certain genres may only treat certain topics; most topics do, however, occur in several genres, which is what interests us here.

Douglas Biber has studied text variation along several parameters, and found that texts can be considered to vary along five dimensions. In his study, he clusters features according to covariance, to find underlying dimensions (1989). We wish to find a method for identifying easily computable parameters that rapidly classify previously unseen texts in general classes and along a small set – smaller than Biber's five – of dimensions, such that they can be explained in intuitively simple terms to the user of an information retrieval application. Our aim is to take a set of texts that *has* been selected by some sort of crude semantic analysis such as is typically performed by an information retrieval system and partition it *further* by genre or text type, and to display this variation as simply as possible in one or two dimensions.

Experiment 1	Experiment 2	Experiment 3 (Brown categories)
I. Informative	1. Press	A. Press: reportage
		B. Press: editorial
		C. Press: reviews
	4. Misc	D. Religion
		E. Skills and Hobbies
		F. Popular Lore
		G. Belles Lettres, etc.
	2. Non-fiction	H. Gov. doc. & misc.
		J. Learned
II. Imaginative	3. Fiction	K. General Fiction
		L. Mystery
		M. Science Fiction
		N. Adv. & Western
		P. Romance
		R. Humor

Table 1: Categories in the Brown Corpus

Method

We start by using features similar to those first investigated by Biber, but we concentrate on those that are easy to compute assuming we have a part of speech tagger (Cutting *et al*, 1992; Church, 1988), such as such as third person pronoun occurrence rate as opposed to 'general hedges' (Biber, 1989). More and more of Biber's features will be available with the advent of more proficient analysis programs, for instance if complete surface syntactic parsing were performed before categorization (Voutilainen & Tapanainen, 1993).

We then use discriminant analysis, a technique from descriptive statistics. Discriminant analysis takes a set of precategorized individuals and data on their variation on a number of parameters, and works out a set *discriminant functions* which distinguishes between the groups. These functions can then be used to predict the category memberships of new individuals based on their parameter scores (Tatsuoka, 1971; Mustonen, 1965).

Evaluation

For data we used the Brown corpus of English text samples of uniform length, categorized in several categories

Variable	Range
Adverb count	19 – 157
Character count	7601 – 12143
Long word count (> 6 chars)	168 – 838
Preposition count	151 – 433
Second person pronoun count	0 – 89
"Therefore" count	0 – 11
Words per sentence average	8.2 – 53.2
Chars / sentence average	34.6 – 266.3
First person pronoun count	0 – 156
"Me" count	0 – 30
Present participle count	6 – 101
Sentence count	40 – 236
Type / token ratio	14.3 – 53.0
"I" count	0 – 120
Character per word average	3.8 – 5.8
"It" count	1 – 53
Noun count	243 – 751
Present verb count	0 – 79
"That" count	1 – 72
"Which" count	0 – 40

Table 2: Parameters for Discriminant Analysis

Category	Items	Errors
I. Informative	374	16 (4 %)
II. Imaginative	126	6 (5 %)
Total	500	22 (4 %)

Table 3: Categorization in Two Categories

as seen in table 1. We ran discriminant analysis on the texts in the corpus using several different features as seen in table 2. We used the SPSS system for statistical data analysis, which has as one of its features a complete discriminant analysis (SPSS, 1990). The discriminant function extracted from the data by the analysis is a linear combination of the parameters. To categorize a set into N categories $N-1$ functions need to be determined. However, if we are content with being able to plot all categories on a two-dimensional plane, which probably is what we want to do, for ease of exposition, we only use the two first and most significant functions.

2 categories

In the case of two categories, only one function is necessary for determining the category of an item. The function classified 478 cases correctly and misclassified 22, out of the 500 cases, as shown in table 3 and figure 1.

4 categories

Using the three functions extracted, 366 cases were correctly classified, and 134 cases were misclassified, out of the 500 cases, as can be seen in table 4 and figure 2. "Miscellaneous", the most problematic category, is a loose grouping of different informative texts. The single most problematic subsubset of texts is a subset of eighteen non-fiction texts labeled "learned/humanities". Sixteen of them were misclassified, thirteen as "miscellaneous".

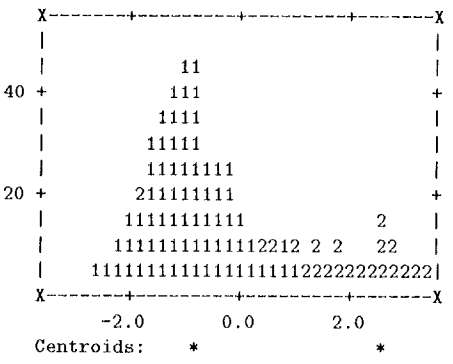

Figure 1: Distribution, 2 Categories

Category	Items	Errors
1. Press	88	15 (17 %)
2. Non-fiction	110	28 (25 %)
3. Fiction	126	6 (5 %)
4. Misc.	176	68 (47 %)
Total	500	134 (27 %)

Table 4: Categorization in Four Categories

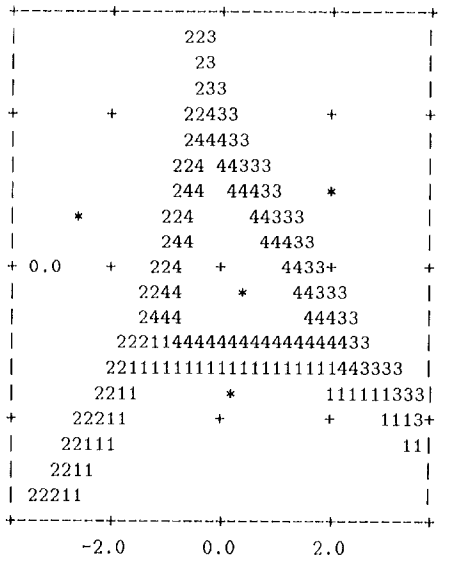

Figure 2: Distribution, 4 Categories

15 (or 10) categories

Using the fourteen functions extracted, 258 cases were correctly classified and 242 cases misclassified out of the 500 cases, as shown in table 5. Trying to distinguish between the different types of fiction is expensive in terms of errors. If the fiction subcategories were collapsed there only would be ten categories, and the error rate for the categorization would improve as shown in the "revised total" record of the table. The "learned/humanities" subcategory is, as before, problematic: only two of the eighteen items were correctly classified. The others were most often misclassified as "Religion" or "Belles Lettres".

Validation of the Technique

It is important to note that this experiment does not claim to show *how* genres in fact differ. What we show is that this sort of technique *can be* used to determine which parameters to use, given a set of them. We did not use a test set disjoint from the training set, and we do not claim that the functions we had the method extract from the data are useful in themselves. We discuss how well this method categorizes a set text, given a set of categories, and given a set of parameters.

The error rates climb steeply with the number of categories tested for in the corpus we used. This may have to do with how the categories are chosen and defined. For instance, distinguishing between different types of fiction by formal or stylistic criteria of this kind may just be something we should not attempt: the fiction types are naturally defined in terms of their content, after all.

The statistical technique of *factor analysis* can be used to discover categories, like Biber has done. The problem with using automatically derived categories is that even if they are in a sense real, meaning that they are supported by data, they may be difficult to explain for the unenthusiastic layman if the aim is to use the technique in retrieval tools.

Other criteria that should be studied are second and higher order statistics on the respective parameters. Certain parameters probably *vary more* in certain text types than others, and they may have a *skewed distribution* as well. This is not difficult to determine, although the standard methods do not support automatic determination of standard deviation or skewness as discrimination criteria. Together with the investigation of several hitherto untried parameters, this is a next step.

Readability Indexing

Not unrelated to the study of genre is the study of *readability* which aims to categorize texts according to their suitability for assumed sets of assumed readers. There is a wealth of formulæ to compute readability. Most commonly they combine easily computed text measures, typically average or sampled average sentence length combined with similarly computed word length, or incidence of words not on a specified "easy word list" (Chall, 1948; Klare, 1963). In spite of Chall's warnings about injudicious application to writing tasks, readability measurement has naively come to be used as a prescriptive metric of good writing as a tool for writers, and has thus come into some disrepute among text researchers. Our small study confirms the basic findings of the early readability studies: the most important factors of the ones we tested are word length, sentence length, and different derivatives of these two parameters. As long as readability indexing schemes are used in descriptive applications they work well to discriminate between text types.

Application

The technique shows practical promise. The territorial maps shown in figures 1, 2, and 3 are intuitively useful tools for displaying what type a particular text is, compared with other existing texts. The technique demonstrated above has an obvious application in information retrieval, for picking out interesting texts, if content based methods select a too large set for easy manipulation and browsing (Cutting *et al*, 1992).

In any specific application area it will be unlikely that the text database to be accessed will be completely free form. The texts under consideration will probably be specific in some way. General text types may be useful, but quite probably there will be a domain- or field-specific text typology. In an envisioned application, a user will employ a cascade of filters starting with filtering by topic, and continuing with filters by genre or text type, and ending by filters for text quality, or other tentative finer-grained qualifications.

The IntFilter Project

The IntFilter Project at the departments of Computer and Systems Sciences, Computational Linguistics, and Psychology at Stockholm University is at present studying texts on the USENET News conferencing system. The project at present studies texts which appear on several different types of USENET News conferences, and investigates how well the classification criteria and categories that experienced USENET News users report using (IntFilter, 1993) can be used by a newsreader system. To do this the project applies the method described here. The project uses categories such as "query", "comment", "announcement", "FAQ", and so forth, categorizing them using parameters such as different types of length measures, form word content, quote level, percentage quoted text and other USENET News specific parameters.

Acknowledgements

Thanks to Hans Karlgren, Gunnel Källgren, Geoff Nunberg, Jan Pedersen, and the Coling referees, who all have contributed with suggestions and methodological discussions.

Category	Items	Errors	Miss
A. Press: reportage	44	11 (25 %)	F
B. Press: editorial	27	8 (30 %)	A
C. Press: reviews	17	4 (24 %)	B
D. Religion	17	8 (47 %)	G
E. Skills and Hobbies	36	17 (47 %)	J
F. Popular Lore	48	32 (67 %)	G,E
G. Belles Lettres, Biographies etc.	75	49 (65 %)	D,B,A
H. Government documents & misc.	30	9 (30 %)	J
J. Learned	80	32 (40 %)	H,D,G,F
K. General Fiction	29	16 (55 %)	fiction
L. Mystery	24	12 (50 %)	-"-
M. Science Fiction	6	1 (17 %)	-"-
N. Adventure and Western	29	18 (62 %)	-"-
P. Romance	29	22 (76 %)	-"-
R. Humor	9	3 (33 %)	-"-
Total	500	242 (48 %)	
Fiction (From previous table)	126	6 (5 %)	
Revised total	500	178 (35 %)	

Table 5: Categorization in 15 Categories

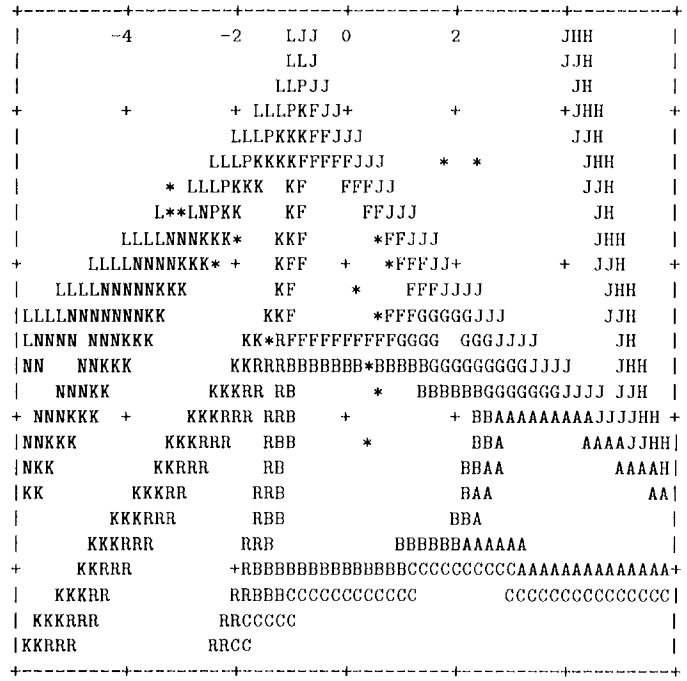

Figure 3: Distribution, 15 Categories — * Indicates a group centroid.

References

Douglas Biber 1989. "A typology of English texts", *Linguistics, 27*:3-43.

Jeanne S. Chall 1948. *Readability*, Ohio State Univ.

Kenneth Church 1988. "A Stochastic Parts of Speech and Noun Phrase Parser for Unrestricted Text", *Procs. 2nd ANLP*, Austin.

Douglass Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun 1992. "A Practical Part-of-Speech Tagger", *Procs. 3rd ANLP*, Trento.

Douglass Cutting, D. Karger, Jan Pedersen, and John Tukey 1992. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections" *Procs. SIGIR'92.*

IntFilter 1993.
Working Papers of the IntFilter Project, available by gopher from `dsv.su.se:/pub/IntFilter`.

George R. Klare 1963. *The Measurement of Readability*, Iowa Univ press.

W. N. Francis and F. Kučera 1982. *Frequency Analysis of English Usage*, Houghton Mifflin.

Seppo Mustonen 1965. "Multiple Discriminant Analysis in Linguistic Problems", *Statistical Methods in Linguistics, 4*:37-44.

M. M. Tatsuoka 1971. *Multivariate Analysis*, New York:John Wiley & Sons.

Atro Voutilainen and Pasi Tapanainen 1993. "Ambiguity resolution in a reductionistic parser", *Procs. 6th European ACL*, Utrecht.

SPSS 1990. *The SPSS Reference Guide*, Chicago: SPSS Inc.