Frequent Term Distribution Measures for Dataset Profiling

Anne De Roeck[1], Avik Sarkar[1], Paul Garthwaite[2]

The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
A.DeRoeck@open.ac.uk, A.Sarkar@open.ac.uk, P.H.Garthwaite@open.ac.uk

Abstract

We motivate the need for dataset profiling in the context of evaluation, and show that textual datasets differ in ways that challenge assumptions about the applicability of techniques. We set out some criteria for useful profiling measures. We argue that distribution patterns of frequent words are useful in profiling genre, and report on a series of experiments with χ^2 based measures on the TIPSTER collection, and on textual intranet data. Findings show substantial differences in the distribution of very frequent terms across datasets.

Evaluation and Dataset Profiles

There is a substantial literature to suggest that the characteristics of a particular corpus or dataset (including genre) will influence the behaviour and performance of Language Engineering (LE) and Information Retrieval (IE) applications and techniques in significant ways. Standard textbooks state, for instance, that keyword based retrieval works better on long documents (Jurafsky and Martin 2000), and that some techniques, such as LSA, work better on datasets with heterogenous vocabulary (Manning and Schuetze 1999). Stemming improves performance for short documents (Krovetz 1993), but not in general (Harman 1991).

Recently, Barbu and Mitkov (2001) have pointed out the impact of the evaluation corpus for anaphora resolution algorithms, and Donaway et al (2000) pursue a related argument for automatic summarization. However, the precise nature of this dependency between dataset and performance remains vague in the absence of established methodologies and measures for profiling datasets. Evaluations of systems and techniques are reported without reference to the characteristics of the collections on which they were performed. Yet, making profiling information available would have several methodological and practical benefits. It would add a dimension to the significance of evaluation results, which could be interpreted in the context of different collections. It would also help researchers and developers in estimating the distance between the type of dataset used for development and evaluation of a system or technique, and the type of dataset on which it is deployed in a practical setting.

In this paper, we set out some criteria for useful measures, and develop one such measure which aims to profile the degree of heterogeneity in the distribution of very frequent terms in different collections. We first formulate a "homogeneity" assumption, which we defeat, for each dataset, by means of an experimental regime based on the χ^2 test (with p-value). Experiments profiled the TIPSTER sub-collections, and a dataset harvested from the Open University intranet. We conclude with a brief evaluation of our measure against the initial criteria.

Developing measures

What makes a good measure?

To be useful, profiling measures have to meet both practical and methodological requirements. Ideally

(i) the (collection of) measures have to profile testable features that are relevant to a range of language processing, search and retrieval techniques;

(ii) they have to be sufficiently diverse and fine grained to allow complex profiles that reflect combinations of a range of relevant features;

(iii) they have to be cheap to implement and run, so they can be used in practical development, over large datasets.

Why measure very frequent term distribution?

The question arises which features to measure. Term distribution patterns for high frequency words are a good starting point. Frequency based measures are cheap to implement (criterion iii). Regarding criteria (i) and (ii), term distribution patterns have been associated with fine grained genre and language modelling (Kilgariff 1997; Rose et al 1997), and with a number of techniques relevant to information retrieval. Katz (1996), for instance, argues convincingly for the identification of (high frequency) function and (rarer) content words with specific distribution patterns. The performance of established retrieval and categorisation techniques improves where stop-word identification takes account of collection specific term distributions (Wilbur and Sirotkin 1992, Yang et al. 1996). Finally, where they are function words, very frequent terms provide large amounts of evidence in almost any textual dataset, and allow a readily available point of comparison between collections.

Some initial sampling we conducted shows that datasets do differ in ways which challenge assumptions about frequent term distribution. The 50 most frequent terms in each of the TIPSTER datasets contain several examples of domain-dependent non-function words. Table 1, for instance, lists a short description of each dataset. The occurrence of "software" in position 21 of the ZF frequency list, and "invention" in position 26 in the PAT frequency list are a case in point.

[1] Computing Department
[2] Statistics Department

The 10 most frequent terms in each of the TIPSTER collections (Table 2) appear to be function words, with a high degree of overlap between datasets. On the other hand, when we compare these with the 10 most frequent terms on our university intranet, we find unexpected results. Not only does the intranet list contain non-function words (eg the word "report", as well as various one character terms which are not English morphemes), we found that these also have a cumulative probability distribution that is, surprisingly, a step function[3]. In short, these points argue for the development of a collection of measures to profile term distribution as it occurs in actual datasets, for use in practical settings. Very frequent terms in particular deserve our attention.

Dataset	Contents of the documents
AP	AP Newswire stories - 1989.
DOE	Short abstracts - Department of Energy.
FR	Issues of the Federal Register (1989)
PAT	U.S. Patent Documents 1983-1991.
SJM	Stories - San Jose Mercury News (1991).
WSJ	Stories - Wall Street Journal (1987-1989).
ZF	Information – Computer Select disks 1989/1990, Ziff-Davis Publishing
OU	The Open University intranet web-pages.

Table 1: Content for each TIPSTER dataset, and OU

Dataset	10 Most Frequent Terms
AP	the, of, to, a, in, and, said, s, for, that.
DOE	the, of, and, in, a, to, is, for, with, are..
FR	the, of, to, and, a, in, for, or, that, be.
PAT	the, of, a, and, to, in, is, for, said, as.
SJM	the, a of, to, and, in, s, for, that, is.
WSJ	the, of, to, a, in, and, s, that, for, is.
ZF	the, m, p, and, to, of, a, in, is, for.
OU	the, of, to, a, and, j, in, k, is, report.

Table 2: 10 most frequent terms for each dataset

Defeating the homogeneity assumption

We investigate the behaviour of very frequent terms across datasets, by formulating a hypothesis: that very frequent terms distribute homogeneously. We then develop a method for defeating it. By tracking the conditions under which the hypothesis is defeated for different collections, we aim to highlight significant differences between datasets.

Our starting point is Kilgariff (1997), who describes a basic method for gauging homogeneity in a corpus, by comparing similarity of term distributions between two halves of a document collection. His basic method involves five steps:

(1) Divide the corpus into two halves by randomly placing text in one of two sub-corpora;
(2) Produce a word frequency list for each sub-corpus;
(3) Calculate the χ^2 statistic for the difference in term frequency distributions between the two sub-corpora;
(4) Normalise for corpus length;
(5) Iterate over successive random halves.

This technique for measuring homogeneity has been linked to gauging the distance between corpora and to genre or sub-language detection (eg Rose and Haddock 1997). The basic technique has been used with different similarity measures. Kilgariff (1997) adopts the χ^2 statistic (by N degrees of freedom). Rose and Haddock (1997) use G2. Alternatives include correlation on term rank frequency data, such as Mann-Whitney (Kilgariff 1996) or Spearman's S (Rose and Haddock 1997). Kilgariff and Rose (1998) compare Spearman's S with χ^2. Cavaglia (2002) uses relative entropy, χ^2 and G2.

We base our measure on χ^2 because it performs well in comparative experiments (Cavaglia 2002; Rose and Haddock 1997), as long as each of the individual frequency values is greater than or equal to 5 and the sample size is large enough (Dunning 1993). On the other hand, our aim of testing the homogeneity hypothesis requires a more fine-grained tool than reporting the χ^2 statistic as a homogeneity measure. We are interested in conditions under which non-homogeneity is detected, and in factors that affect the degree of non-homogeneity in different datasets. (De Roeck et al 2004b) describes in detail how we adapted Kilgariff's methodology. Briefly, we differentiate results by reporting the p-value as well as the CBDF statistic. Given a null hypothesis (in our case, homogeneity), the p-value allows us to estimate the strength of the evidence offered by the data. As usual, a p-value < 0.05 will indicate that evidence of non-homogeneity is statistically significant. The CBDF measure relates to the text and indicates the level of heterogeneity.

In the experiment, a corpus is split in two, by randomly placing text in one of two sub-corpora. Kilgariff (1997) and Rose and Haddock (1997) remove document boundaries and place consecutive chunks of 5000 words in each partition. The method of partitioning a document is important because it affects the outcome of experiments. This is easy to show: a chunk size of 1, for example, would remove all evidence of term dependence in the data, and the experiment would fail to defeat the homogeneity assumption. On the other hand, we know that a chunk size of 5000 shows high levels of heterogeneity. Because we are interested in investigating at which point heterogeneity registers in different datasets, we experimented with alternative ways of partitioning a corpus, and with different ways of handling document boundaries. We conducted three experiments:

docDiv: Assign each document at random to either of two partitions.

halfdocDiv: Randomly assign half of each document to a partition, and the other half to the other partition.

chunkDiv: Remove document boundaries and repeat the same experiments of Kilgariff (1997) with various chunk sizes, from 5 to 5000, and observe the homogeneity measure.

The data

We choose the seven TIPSTER sub-collections for our experiments. Not only are they readily available, they

[3] We are not aware of any current work identifying or investigating cumulative probability distributions of this kind in document collections.

also have been used as an evaluation standard, and so profiling measures would become useful to past evaluations. The datasets are artificially compiled, with some drawn from a narrow base of similar text types.

Data Set	No of Docs	Corpus Length (000 words)	Avge Doc Length in words	No of Distinct Terms
AP	242,918	114,438	471	347,966
DOE	226,086	26,883	119	179,310
FR	45,820	62,805	1,371	157,313
PAT	6,711	32,152	4,791	146,943
SJM	90,257	39,546	438	178,571
WSJ	98,732	41,560	421	159,726
ZF	293,121	115,957	396	295,326
OU	53,681	39,807	744	304,468

Table 3: Basic Rough Profiles of TIPSTER and OU datasets.

To contrast, we also experimented on Open University Intranet (OU) data, a more diverse, but naturally occurring collection. Table 3 lists basic profiles. DOE, for example, appears relatively uniform regarding text length, whereas FR shows the largest range. Comparing the ratio of new to old words gives a rough indication of domain diversity. For instance, there is a significant difference between the rate of new terms occurring, between the OU dataset (1 in 131 words) and the SJM dataset (1 in 260 words), in spite of their similar size. The WSJ and SJM sets are quite close in size and characteristics as well as in genre type, so we would expect them to behave in similar ways.

Experimental Results

We track the distribution of the N most frequent terms, so we examine mainly stylistic homogeneity because the behaviour of function words will dominate the experimental outcomes. We add detail by reporting results at different values for N. Experimental results are shown in Tables 4 to 6, with each cell listing CBDF and p-value (averaged over iterations). Values in bold indicate cases where the homogeneity assumption has not been defeated ($p>0.05$).

The docDiv experiment (Table 4) investigates homogeneity across documents in a collection. The experiment finds heterogeneity ($p<0.05$) in almost all cases. The exceptions are the AP and the DOE datasets for the 20 most frequent terms, and the WSJ and SJM datasets for the 10 most frequent terms. CBDF values provide further insight with high values indicating high levels of non-homogeneity.

The halfdocDiv experiment (Table 5) is sensitive to document boundaries, and shows that very frequent terms distribute more homogeneously within, than across documents. Note that the DOE set appears to be very uniform, and PAT extremely heterogeneous with very low p-value and very high CBDF. Also, our measure appears capable of highlighting similar behaviours in related collections (WSJ and SJM, both newspaper text). FR and PAT appear very heterogeneous, perhaps related to comparatively stylized document formats (particularly PAT). Note extremely high levels of within-document heterogeneity for intranet data.

Data Set	N Most Frequent Terms			
	10	**20**	**50**	**100**
AP	**2.107 0.1216**	**1.576 0.2139**	2.583 0.0003	2.290 0
DOE	**1.172 0.463**	**1.450 0.160**	1.755 0.0259	1.983 0
FR	54.524 0	41.715 0	72.093 0	66.787 0
PAT	21.074 0	29.315 0	62.494 0	55.353 0
SJM	**3.595 0.1193**	2.768 0.0077	3.231 0	2.976 0
WSJ	**2.358 0.178**	2.663 0.0019	2.364 0	2.335 0
ZF	11.947 0	8.133 0	6.907 0	6.576 0
OU	232.9 0	158.52 0	94.75 0	67.29 0

Table 4. docDiv results. Average CBDF and p-values

Data set	N Most Frequent Terms			
	10	**20**	**100**	**500**
AP	**1.774 0.087**	**1.473 0.117**	**1.271 0.066**	1.171 0.021
DOE	**0.728 0.655**	**0.931 0.533**	**1.043 0.372**	**1.061 0.195**
FR	7.905 0.001	9.549 0	11.642 0	8.847 0
PAT	20.360 0	15.568 0	11.886 0	7.694 0
SJM	**1.323 0.3860**	**1.569 0.3919**	**1.469 0.1069**	1.332 0
WSJ	**1.563 0.279**	**1.618 0.248**	**1.298 0.260**	1.236 0.017
ZF	**1.948 0.1288**	**1.858 0.116**	1.609 0.0240	1.559 0
OU	7.721 0.033	6.103 0.0025	8.216 0	6.366 0

Table 5. halfdocDiv results. Average CBDF and p-values

Regarding chunkDiv (Table 6), for space reasons, we only show one table at chunk size 100 (see De Roeck et al (2004b) for more detailed results). Chunkdiv experiments show a clear relationship between increasing chunk size and increased evidence and levels of heterogeneity for all collections.

There appears to be a relationship between registering heterogeneity, document length and domain issues. DOE (narrow domain, short documents) showed high homogeneity in the halfDocDiv experiment, but registers heterogeneity here, possibly because chunk size interferes with document boundaries The

experiment confirms high heterogeneity of PAT, and shows, again, similar behaviour for WSJ and SJM. Again, the intranet data show strong evidence for high levels of heterogeneity

Data set	N Most Frequent Terms			
	10	**50**	**100**	**500**
AP	**0.824** **0.6023**	**1.412** **0.0735**	1.607 0.0019	1.471 0
DOE	**1.102** **0.3937**	1.646 0.0231	1.511 0.0317	1.354 0.0299
FR	**1.006** **0.5071**	**1.608** **0.076**	1.803 0.025	1.924 0
PAT	4.181 0.0232	2.682 0.0007	2.420 0	2.252 0
SJM	**0.995** **0.4720**	**1.146** **0.3203**	**1.180** **0.2463**	1.410 0
WSJ	**1.112** **0.3741**	**1.198** **0.2426**	**1.230** **0.0937**	1.196 0.0383
ZF	**1.576** **0.4152**	1.709 0.011	2.190 0	1.41 0
OU	6.231 0.0004	4.870 0	4.278 0	3.310 0

Table 6: chunkdiv at chunk size 100. Average CBDF and p-values

How good are our measures?

Our experiments seem reliable as they confirm Kilgariff (1996) and Katz (1996) who anticipate that more frequent function words have more similar distributions among documents than less frequent terms. Taken as a measure of stylistic homogeneity, experiments also confirmed that very frequent terms (and function words) distribute more homogeneously within the same document than across document boundaries, and appear to have similar distribution patters in related genres.

At first inspection, our measures appear pretty good profiling tools. They performed quite well against the criteria set out earlier. They are cheap and fast to implement over large and diverse datasets. The combination of an indication of statistically relevant evidence with a similarity measure and a sequence of partitioning methods which introduce varying degrees of randomness yields a more fine grained profiling tool than mere reporting of CBDF. Finally, the measures seem capable of bringing out significant differences between live and artificial collections, with intranet textual data showing markedly higher degrees of heterogeneity. Importantly, by reporting both the p-value and the CBDF, even a small departure from homogeneity can be detected if a sample's size is large enough. As the sample size increases, the p-value will get closer and closer to 0. CBDF provides a measure of homogeneity that is not affected greatly by sample size, so that corpora of different lengths can be compared. However, the similarity measure should be compatible with the test of homogeneity, so that if two corpora are of similar size, the one with the larger value on the similarity scale should also have the smaller p-value for the test of homogeneity. Unlike bare CBDF, this is the case here.

On the other hand, the relationship between document boundary, structure and length, and evidence of heterogeneity is not clear (eg the behaviour of the PAT set), and we intend to investigate this further. Current work is looking at the impact of domain structure and coverage on collection profile.

References

Barbu, C. and R. Mitkov (2001) Evaluation Tool for rule-based anaphora resolution methods. *Proceedings ACL*, pp34-41, Budapest.

Cavaglia, G. (2002) Measuring corpus homogeneity using a range of measures for inter-document distance. ITRI Report Series, ITRI-02-08, University of Brighton, UK.

Donaway, R.L., Drummey, K.W. and L.A. Mather (2000) A Comparison of Rankings produced by summarisation evaluation measures. *Proceedings NAACL-ANLP Workshop on Text Summerisation*.

De Roeck, A., Sarkar, A and Garthwaite, P. (2004a) Defeating the Homogeneity Assumption. Proceedings of JADT04.

De Roeck, A., Sarkar, A. and Garthwaite P. (2004b) Profiling Datasets with Very Frequent Term Distribution Measures. Computing Technical Report TR 2004/05. The Open University.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Co-incidence. *Computational Linguistics*, 19(1):61-74.

Donna Harman: (1991) How effective is suffixing? *JASIS* 42(1): 7-15

Jurafsky, D. and J.H. Martin (2000) *Speech and Language Processing*. Prentice Hall. New Jersey.

Katz, S. (1996) Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15-59.

Kilgariff, A. (1996) Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 33--40.

Kilgariff, A. (1997) Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings ACL-SIGDAT Workshop on very large corpora*, Hong Kong.

Kilgariff, A. and T. Rose (1998) Measures for Corpus Similarity and Homogeneity. *Proc. 3rd Int. Conf. Empirical Methods in NLP*. pp 46-52. Granada.

Krovetz, R. (1993) Viewing Morphology as an Inference Process. *SIGIR-93*, pp 191-202. ACM.

Manning, Christopher and Hinrich Schuetze. (1999) *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.

Rose, T. and N. Haddock (1997) The effects of corpus size and homogeneity on language model quality. *Proceedings ACL-SIGDAT workshop on very large corpora*, pp178-191, Hong Kong.

Wilbur, J. and K. Sirotkin. 1992. The Automatic identification of Stop Words. *Journal of Information* Science. 18:45-55. Elsevier.

Yang, Yiming and J. Wilbur. 1996. Using Corpus Ststistics to Remove Redundant Words in Text Categorisation. *JASIS*. 47(5):357-369.