Usenet Newsgroups' Profile Analysis Utilising Standard and Non-standard Statistical Methods

P.J. Sallis and D.A. Kassabova
Information Science Department, University of Otago,
PO Box 56, Dunedin, New Zealand

Abstract

This paper explores building profiles of Newsgroups from a corpus of Usenet email messages employing some standard statistical techniques as well as fuzzy clustering methods. A large set of data from a number of Newsgroups has been analysed to elicit some text attributes, such as number of words, length of sentences and other stylistic characteristics. Readability scores have also been obtained by using recognised assessment methods. These text attributes were used for building Newsgroup profiles. Three newsgroups, each with similar number of messages were selected from the processed sample for the analysis of two types of one-dimensional profiles, one by length of texts and the second by readability scores. Those profiles are compared with corresponding profiles of the whole sample and also with those of a group of frequent participants in the newsgroups. Fuzzy clustering is used for creating two-dimensional profiles of the same groups. An attempt is made to identify the newsgroups by defining centres of data clusters.

It is contended that this approach to Newsgroup profile analysis could facilitate a better understanding of computer- mediated communication (CMC) on the Usenet, which is a growing medium of informal business and personal correspondence.

I. Introduction

In recent years, computer-mediated human communication carried out over computer networks, has increasingly attracted the attention of researchers from different areas of science and humanities, especially computer and information science, communication,

psychology, languages and linguistics. Thimbleby[9] argues that the Internet has introduced the first truly *new* form of human discourse for five thousand years. Exaggerated as this claim may sound, it raises research questions about the nature of this undoubtedly new phenomenon of human communication.

An enormous amount of human communication is carried out over the Usenet, one of the largest networks in the world. It is even very difficult to define what the Usenet is, because it is not one simple entity. According to Moraes[7] "*Usenet is not the Internet. The Internet is a wide-ranging network[...]. It carries many kinds of traffic, of which Usenet is only one. And the Internet is only one of the various networks carrying Usenet traffic.*"

The basic building block of Usenet is the *newsgroup*, which is a collection of messages with a related topic. There are thousands of active newsgroups on the Usenet, some of them disappearing and many more newsgroups appearing all the time. The variety of topics covered by the newsgroups is enormous, let alone the number of individual messages. Huge also, is the variety of participants who differ widely in cultural and educational background, age, language and communication skills.

II. Data set description

The source set of data used for this research contains a set of messages extracted from a large number of Newsgroups on the Usenet. It was first complied in the U.K. for a text retrieval and indexing research project funded by The British Library[8]. This was part of an international text retrieval project TREC, originated by The National Institute of Standards and Technology (NIST), in the USA[3]. The aims of that work are concerned with corpus indexing and information retrieval issues.

The set of data used for the experiments described in this paper consists of 46621 messages that have been posted by 21006 senders to 2240 newsgroups on the Usenet. Given its comparatively large size, and the large number of newsgroups from which the messages have been extracted, this data set could be considered an indicative sample of the population of messages posted to newsgroups on the Usenet.

Each message belongs to one or more newsgroups and consists of a header and text. To begin with, all texts were computationally processed to remove unnecessary lines (lengthy signatures, lines containing predominantly numbers, etc.). Messages longer than 1000 lines were not processed and were excluded from the data set, as they are usually either 'Frequently asked questions' or just articles from other sources posted to newsgroup and therefore, are not 'genuine' postings that would characterise a particular newsgroup. The pre-processed texts along with relevant information about each message (sender, subject line, the posted newsgroup(s)) have been further structured into a relational database. The pre-processed texts in the database contain 5681386 words altogether.

III. Obtaining stylometric characteristics for the texts

For the purpose of conducting statistical analysis of the texts, a set of stylometric characteristics has been obtained for each text. These include the number of running words, the number of common words, the number of unique word forms, readability scores for each message, passive voice usage in each message, the number and length of sentences and paragraphs in each message. The combination of these attributes characterises the individual texts in a unique way. Together they can be used for compiling text profiles.

It is interesting to note that many messages contain some number of words that are not part of completed sentences. This is due to the fact that senders often do not care much about punctuation, capitalisation, etc. There are also some cases when the automatic pre-

processing of messages has removed lines that contain not only a large proportion of numbers, but also some words which may be part of sentences. This peculiarity of the texts under investigation has to be taken in account when obtaining stylometric statistics. For instance, readability scores are only obtained for those parts of texts that consist of completed sentences. This is the only sensible approach, as one cannot (and should not) measure readability for a string of words that do not form a sentence.

IV. Statistical profiles of newsgroups

It is reasonable to assume that the number of words in messages is a quantitative measure for text traffic on the Usenet. On the other hand, the readability of a text could be considered a measure for its quality in the sense that the more readable a text is, the more likely it is to be comprehended by its readers. In other words, complete text is required in individual instances in order for originators to communicate information effectively to the recipients.

Five groups of messages are considered for building profiles. The first three are newsgroups with similar number of messages as they appear in the database. Group 1 (alt.politics.equality) and group 3 (alt.journalism.criticism) belong to the set of 'alternative' groups on the Usenet, while group 2 (comp.os.ms-windows.apps.utilities) belongs to the most popular class of newsgroups - comp.

The set of messages belonging to senders with more than 20 messages is referred to as 'Frequent senders'(or 'Senders' for short) and is considered a fourth group. The set of all messages in the database is the fifth group and is referred to as 'All'. Table 1 depicts the names and sizes of groups.

A. One-dimensional profile

First the above five groups are investigated by number of words in messages. The relative frequency of messages per a hundred-word interval is found for each group. There are 16 intervals that cover number of words from 0 to 1499. Note that the first interval covers only messages with 0 words. Both in 'All' and 'Frequent senders' there are very few messages with more than 1499 words, while there are not such messages in groups one to three. That is why messages longer than 1499 words are not considered for the purposes of this investigation.

Results displayed in Table 2 are graphically represented in Figure 1. The graphs show convincingly that the group of 'Frequent senders' match very closely to the group 'All', while the other three groups differ more or less. For instance, it can be seen that senders in group 1 are very 'talkative', i.e. there are more long messages and fewer short messages there than in the other groups. On the other hand, the length of messages in group 2 oscillates around 100 words and there is only one message longer than 299 words. Group 3 matches very closely the overall character of the sample ('All' and 'Senders'), but again the percentage of longer messages (between 200 words and 499 words) is higher than for the whole sample.

Next the relative frequency of messages per readability score interval is found for the same five groups. These intervals are defined according to one of the best-known readability measures, namely Flesch Reading Ease method[4]. The formula used by this method produces a difficulty index which relates to comprehension score on a scale of 0 - 100.

Reading ease score=206.835 - (0.846 x SYLLS/100W) - (1.015 x WDS/SEN)

where SYLLS/100W = syllables per 100 words and WDS/SEN = average number of words per sentence.

Note that the higher the score, the more readable the text is. For the purposes of this research the readability scores are interpreted as it is shown in Table 3

The frequency of messages per readability score interval is depicted in Table 4 and Figure 2. As it can be seen in Figure 2, the graph for 'Frequent senders' has very close similarity to that for 'All'. On the other hand, from the rest of the groups only the graph for group 3 is vaguely similar to 'All' and 'Senders'. Group 1 demonstrates surprisingly low readability for a comparatively high percentage of messages. Although most messages in group 2 belong to 'standard' and 'fairly easy' intervals, there is a sudden increase of number of messages in the area of 'very difficult'.

From the above results it could be argued that the group of 'Frequent senders' pre-determines (and therefore represents) the characteristics of the whole sample both in terms of readability and number of words. Next a two-dimensional profile for each of groups 1, 2 and 3 is built and compared to the profile of group 'Senders' assuming that it closely represents the whole sample (group 'All').

B. Two-dimensional profile

Building a two-dimensional profile of a data set should allow for simultaneously exploring how that set behaves in relation to two different parameters. For the purpose of this research it would give a comprehensive picture of how different groups of messages are characterised in terms of quantity (number of words) and quality (readability scores).

One way of building a two-dimensional profile is to use clustering techniques. Clusters characterise the distribution of data in the problem space. Finding the cluster centres

and the membership of each element to clusters could be used for identifying the pattern of data behaviour in the problem space. Comparing cluster centres of one data set to those of another data set could give indication about similarities and differences between sets.

Plotting the data sets from groups 1, 2, 3 and 'Senders' in the problem space 'Readability/Number of words' (Figure 3) shows no obvious clusters of data.

A clustering technique that could account for the ambiguity that characterises data from the real world is *fuzzy clustering*. Fuzzy clustering[1] finds cluster centres by optimising the distance of elements to these centres. Each element in a set can belong to more than one cluster to a degree that is in the interval [0,1]. The sum of its membership degrees to all clusters is 1. Fuzzy clustering allows for fuzzy borders between clusters.

Here a two-dimensional fuzzy clustering with three cluster centres was performed in the problem space 'Readability/Number of words' using MATLAB® environment[2]. Fuzzy clustering within MATLAB® aims at working out cluster centres that mark the mean location of each cluster. Initially these cluster centres are inaccurately placed and every data point has a membership grade for each cluster. Then these cluster centres and membership degrees are iteratively updated until the centres move to the 'right' location. In consecutive steps four data sets (group1, group 2, group 3 and 'Senders') were plotted in the same space and fuzzy clustering was performed for each of them. Figure 4 depicts all data sets and their respective cluster centres. The exact values for readability and number of words for all cluster centres are shown in Table 4.

Two of the cluster centers for group 1 (Centre 1 and 3) show clustering of messages with higher number of words, while the readability for the same group is mostly 'standard" and 'fairly difficult'.

The second group - comp.os.ms-windows.apps.utilities - shows tendencies for much shorter messages. Two of its cluster centres have the lowest 'Number of words' coordinates and the coordinate for the third one is not very high either – only 122.46. Although the participants in this group could be commended for their laconic communication, the readability characteristics for part of the messages in the group is alarmingly low (Centre 2).

Apparently, group 3 is the most consistent in terms of readability as all its clusters oscillate around the area of standard readability. On the other hand, the tendency for verboseness in this group is even more obvious than in group 1.

The last group is 'Senders' and it is assumed to represent the whole sample in terms of readability and number of words. The clustering there strikes with the high values for number of words which suggests that in many other newsgroups there are much longer messages than those in group 1 and group 3. In terms of readability, the messages in 'Senders' (and therefore in the whole sample) are mainly clustered in the area of standard readability.

V. Conclusions and further work

This analysis illustrates how stylometric statistical data could be used for assessing groups of messages from the Usenet in terms of quantity and quality. Different sets of messages (in a particular newsgroup or messages belonging to some groups of senders) could be evaluated and compared. This could be used by researchers in the area of CMC (or by so-called 'newbies' when contemplating joining a newsgroup) as an indication of what the profile for a particular newsgroup or set of messages could be.

Another approach to text analysing and classifying could be using Kohonen self-organising maps[6], particularly if the input vectors represent multiple text characteristics.

Self-organising maps could learn to group or recognise sets of similar input vectors in such a way that neurons physically close to each other in the output neuron layer correspond to similar input vectors.

Self-organising maps can also be used for text classification based on contextual similarities between the individual texts. Such an approach is discussed in[5] and illustrated on the Internet at the address http://websom.hut.fi/websom/.

Another direction for further research could be quality assessment of CMC on the Usenet by defining a criterion as a fuzzy relation of selected fuzzified attributes. Such a criterion could be applied to the database containing statistics for fuzzy information retrieval of messages that fulfil the criterion to a given degree. This could be useful in subject-based information retrieval from such unstructured texts as this corpus of electronic correspondence provides. Inevitably some form of profile building and matching is required and this paper has described by observation of a large extant sample, how such profile could be defined.

References:

[1] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York,1981).

[2] N. Gulley and J.-S. Roger Jang, *Fuzzy Logic Toolbox User's Guide,* (The MathWorks, Inc.,1996).

[3] D. Harman, Overview of the Third Text REtrieval Conference (TREC-3), In: D.K. Harman (ed.), *The Third Text REtrieval Conference, April 1995* (NIST, Gaithersburg, MD, 1995)

[4] C. Harrison, *Readability in the Class Room* (Cambridge University Press, 1980).

[5] T. Kohonen, Exploration of Very Large Databases by Self-Organising Maps, *In: Proceedings of the 1997 International Conference on Neural Networks (ICNN'97), Houston, June 1997*, (IEEE, June, 1997).

[6] T. Kohonen, *Self-Organization and Associative Memory, 2nd Edition*, (Springer-Verlag, Berlin, 1987)

[7] M. Moraes, What is Usenet, Frequently posted article to news:news.announce.newusers, (accessed on 1 July 1997).

[8] S. E. Robertson et al, OKAPI at TREC-3. In: D.K. Harman (ed.), *The Third Text REtrieval Conference, April 1995* (NIST, Gaithersburg, MD, 1995).

[9] H. Thimbleby, Internet, Discourse and Interaction Potential, In: L.K. Yong, L.Herman, Y.K. Leung, and J. Moyes (eds.), *APCHI'96, Proceedings of The First Asia Pacific Conference on Computer Human Interaction, Singapore, 25-28 June 1996 (Information Technology Institute, Singapore, 1996).*

Table 1
Names and sizes of groups of messages

GROUP	Number of messages
Group 1 (alt.politics.equality)	237
Group 2 (comp.os.ms-windows.apps.utilities)	239
Group 3 (alt.journalism.criticis)	212
Frequent senders	5951
All	46621

Table 2
Relative frequency of messages per a hundred-word interval

	Frequency in absolute numbers					*Relative frequency*				
Interval	Group 1	Group 2	Group 3	Senders	All	Group 1	Group 2	Group 3	Senders	All
0	13	3	9	88	756	5.49%	1.26%	4.25%	1.00%	1.62%
001-099	103	195	137	3994	31425	43.46%	81.59%	64.62%	67.11%	67.41%
100-199	62	39	32	1058	8677	26.16%	16.32%	15.09%	17.78%	18.61%
200-299	31	1	15	399	2750	13.08%	0.42%	7.08%	6.70%	5.90%
300-399	15	1	11	142	1092	6.33%	0.42%	5.19%	2.39%	2.34%
400-499	3	0	3	70	574	1.27%	0.00%	1.42%	1.18%	1.23%
500-599	4	0	2	53	322	1.69%	0.00%	0.94%	0.89%	0.69%
600-699	2	0	0	25	191	0.84%	0.00%	0.00%	0.42%	0.41%
700-799	2	0	1	18	154	0.84%	0.00%	0.47%	0.30%	0.33%
800-899	0	0	0	17	125	0.00%	0.00%	0.00%	0.29%	0.27%
900-999	0	0	0	8	74	0.00%	0.00%	0.00%	0.13%	0.16%
1000-1099	0	0	0	6	49	0.00%	0.00%	0.00%	0.10%	0.11%
1100-1199	2	0	2	3	47	0.84%	0.00%	0.94%	0.05%	0.10%
1200-1299	0	0	0	9	33	0.00%	0.00%	0.00%	0.15%	0.07%
1300-1399	0	0	0	4	21	0.00%	0.00%	0.00%	0.07%	0.05%
1400-1499	0	0	0	3	18	0.00%	0.00%	0.00%	0.05%	0.04%
Total	237	239	212	5897	46308	100.00%	100.00%	100.00%	98.61%	99.33%

Table 3
Readability score interpretation

Score	Reading Difficulty
90 -100	Very easy
80 - 90	Easy
70 - 80	Fairly easy
60 - 70	Standard
50 - 60	Fairly difficult
30 - 50	Difficult
0 - 30	Very Difficult

Table 4
Message frequency per readability score interval

		Frequency in absolute numbers					Relative frequency				
Score Interval	Difficulty	Group 1	Group 2	Group 3	Senders	All	Group 1	Group 2	Group 3	Senders	All
90-100	very easy	12	23	23	586	4092	5.06%	9.62%	10.85%	9.85%	8.78%
80-89	easy	9	46	19	848	6668	3.80%	19.25%	8.96%	14.25%	14.30%
70-79	fairly easy	54	53	49	1427	11086	22.78%	22.18%	23.11%	23.98%	23.78%
60-69	standard	54	53	43	1330	10543	22.78%	22.18%	20.28%	22.35%	22.61%
50-59	fairly difficult	52	24	34	814	6073	21.94%	10.04%	16.04%	13.68%	13.03%
30-49	difficult	42	12	19	526	4425	17.72%	5.02%	8.96%	8.84%	9.49%
0-29	very difficult	14	28	25	420	3734	5.91%	11.72%	11.79%	7.06%	8.01%
	Total	237	239	212	5951	46621	100.00%	100.00%	100.00%	100.00%	100.00%

Table 5
Cluster centres for the individual groups

Group	Centre 1		Centre 2		Centre 3	
	Readability	Num of words	Readability	Num of words	Readability	Num of words
Group 1 **x**	62.04	681.07	60.71	62.1334	57.8667	253.8825
Group 2 **o**	72.99	122.46	10.60	17.9038	73.7662	37.8749
Group 3 +	62 6	299.3	60.8	1067.5	61.4	46.5
Senders *	61.2	595.1	67.4	5035	65.8	72.3

Figure 1
Graphical representation of relative frequency of messages per a hundred-word interval for the individual groups.

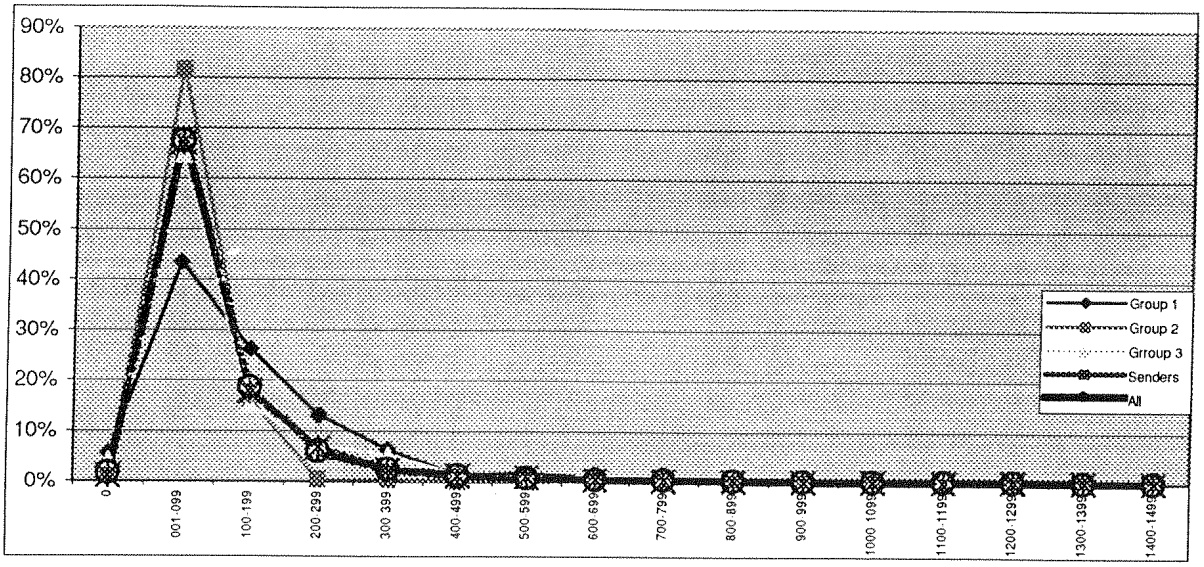


Figure 2
Graphical representation of relative frequency of messages per readability score interval for the individual groups

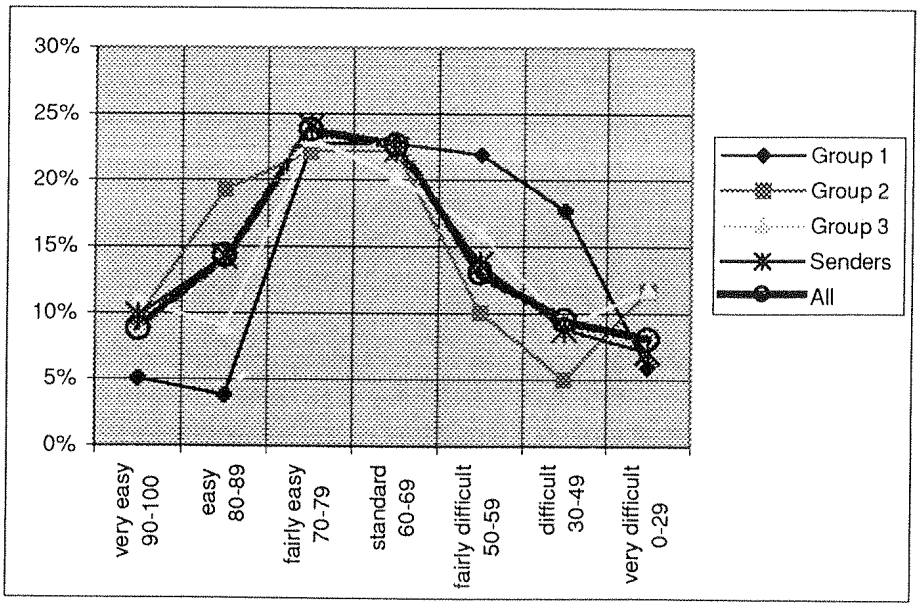
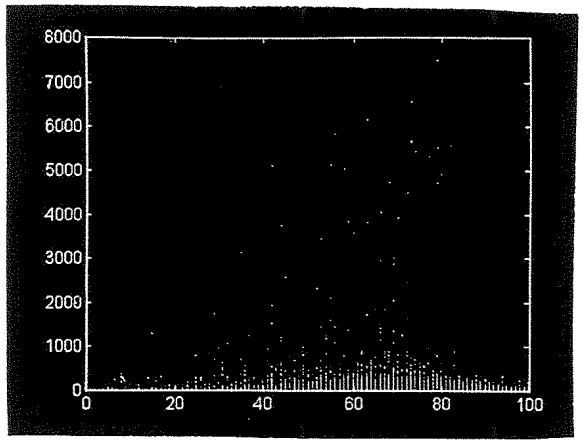
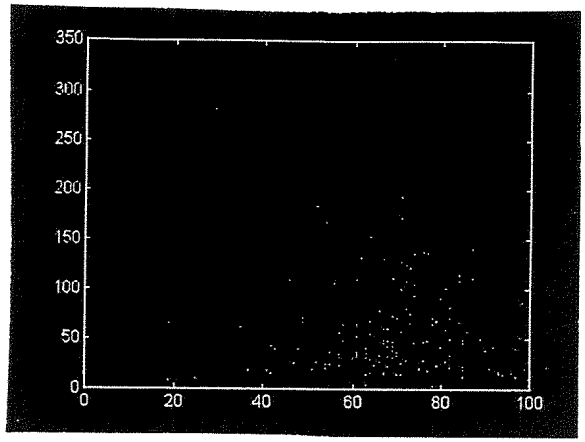


Figure 3
Plots for data in groups Senders, 1, 2 and 3 in the problem space 'Readability/Number of words'

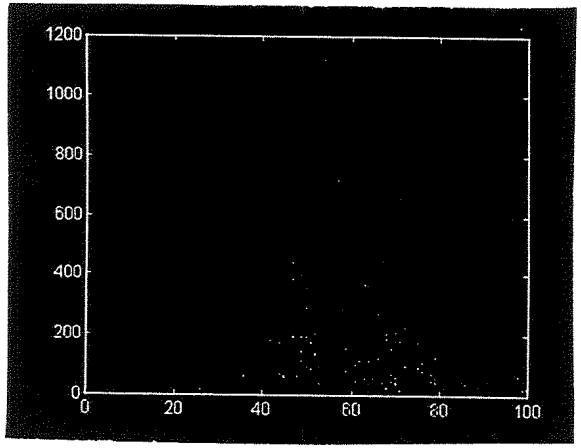
Senders



Group 2



Group 1



Group 3

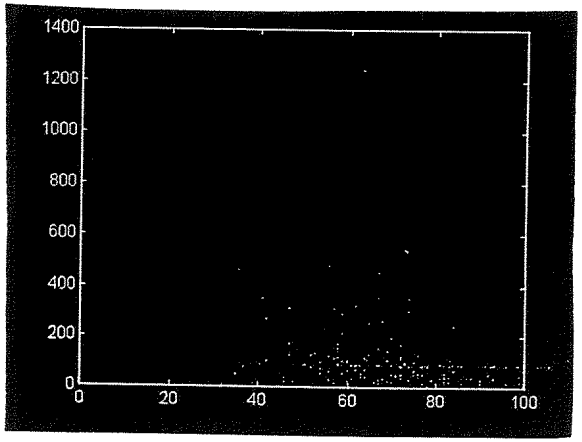


Figure 4
Cluster centres for Group 1, 2, 3 and Senders