https://doi.org/10.48009/2_iis_2007_473-477

BE BRIEF THE KEY IN USING TECHNICAL INTERNET GROUPS

David L. Munro, University of Wisconsin at Whitewater, munrod@uww.edu
George F. Sargent, University of Wisconsin at Whitewater, sargentg@uww.edu

ABSTRACT

This paper examines using threaded discussion technical groups as an aid in solving technical problems. The study focuses on factors that may affect whether a question receives a response. The terms used and the length of subject lines were examined for 66,650 threads from 13 different technical groups. Statistics from threads that received replies were compared to threads that did not receive replies. The average number of terms used in threads that received replies was lower than threads that did not receive replies across all 13 groups. T-Tests were significant for 11 of the 13 groups tested. Some support was found for the use of specific terms in subject lines.

Keywords: Computer Mediated Communication (CMC); Online Groups; Social Networks; Information Seeking;

INTRODUCTION

Problems and questions often arise in the use of computer technology. In many cases they involve the use of software that the user has not written. Some examples would be a user of an enterprise resource package (ERP) or a developer using a specific development language or platform developed by a software company such as Microsoft. Getting the answer to one's questions is often a matter of communicating with the right person. The problem often is the identity of that person is unknown. While many other resources exist such as manuals or support services from the software company, a source many people are turning to is posting the question to a technical group on the internet in hopes that someone will provide an answer.

Response rates vary greatly for a variety of factors for example: a review of the response rate of 17 groups from the Microsoft NetScan Project for Feb. 2007 (*http://netscan.research.microsoft.com/*) of newsgroups that included the term 'vista' in the newsgroup name had a mean reply rate of 76.5% with a range from 62.2% to 96.1%. While a review of the same system and criteria as before using the term 'basic' found 20 groups with the mean reply rate of only 16.4% with a range from 0% to 64.8%. Clearly the topic has an impact. However the user with a question they need answered is not able to choose the topic.

The question of interest to us as both developers and researchers is: What factors within your control influence the probability of getting a reply? Some such factors of interest follow: Is the probability of a response to an information request a function of the group you post to? Is it a function of how the request is made? Does it matter if you participated in the group before your request? Does it matter if you said 'thank you' to a prior helpful response? This study focused primarily on a narrow portion of this domain in that it looked at just the subject lines of the requests.

Prior Research

Text based asynchronous computer mediated communication (CMC) such as Usenet has been in existence for decades and dates back to the early years of the Internet. As a result they have been studied for quite some time. Care should be taken in interpreting the older studies since the population of users has dramatically broadened and changed with the Web. Hence we primarily focused on research from the last five years.

It has been noted that online groups may be viewed as social communities. Efforts have been made to understand these communities by visualizing these online social communities. [2, 4, 5, 9] and the development of "social accounting meda-data" [3, 9].

A response rate to initiated messages has been studied. In evaluating initial posts of newcomers to six newsgroups it was noted that response rates improved among those posts that were longer (16.4%) and asked a question (14.4%) [6]. Arguello, et al found that response rate was impacted by a multitude of issues: if the creator is a newcomer, if the message is cross listed, topic coherence, ask a question, sentence length, the use of 3rd person pronoun or the word 'I' in the message. [1].

RESEARCH METHODOLOGY

Population Studied

Thirteen groups were selected from Yahoo Groups for evaluation. Yahoo groups function similar in manner to other forms of asynchronous text-base computer mediate communication forums such as Usenet. The number of such groups is staggering with over a million Yahoo groups. For the average user looking for which group to join the volume can be overwhelming. For example the section of Computers and the Internet, the subgroup for programming languages alone has 18,989 different groups, and there are 4,701 groups under the Katrina section. Yahoo allows for groups to be public or private. To participate in a private group or access the archive of messages in a private group you must join the group.

Table 1: Groups Studied

General Domain		
Group ID: Yahoo Group Name	**# in Group**	**Group Description**
Visual Basic Domain		
B1:VBBeginners	6439	Visual Basic Beginners
B2: vbhelp	2435	Visual Basic
B3:VisualBasic_Official	4557	Visual Basic
B4:visualbasic6programming	6457	Visual Basic 6
Java Domain		
J1:beginnersclub	2735	Java Beginners
J2:gelide	7251	A freeware Java IDE called Gel
J3:Java_Official	14545	Java Related
J4:Junit	7365	Java Testing Framework
Web Domain		
W1:1stJavaScript	1392	HTML, JavaScript or CSS
W2:JavaScript_Official	5498	Java Script & other web related
W3:JS-Jive	1965	Java Script
W4:overlib	1623	overLIB is a JavaScript library
ERP Domain		
E1:max_users	498	ERP package for MRP

The selection of the specific groups was made from groups where people were primarily looking for answers to technical computer questions and the Yahoo group had to be public. In addition four groups in the same general domain were chosen for three different domains as a way of gaining some qualitative understanding if the domain topic or the group size was significant to response rate. These 12 groups were groups that primarily developers would be interested in. One additional group pertaining to an ERP package for small and mid size manufacturing was included that also included to gain some insight into groups that would also include user technical questions. These groups are listed in Table 1. The two letter groups ID is followed by the Yahoo group name. The #in group was the number reported by yahoo groups. One of the groups 'gelide' ceased functioning in May of 2006.

Variables of Study

The dependent variable for this study was the binary condition of whether or not a message received a response. The independent variables were the number of terms in the subject line and the presence or absence of certain words in the subject line. Since groups can be viewed as online communities with their own social norms, the individual groups were evaluated separately.

While not able to be statistically analyzed due to the small number of groups (13) some of the aggregated group level data was gathered and informally evaluated to gain insight for future study. This included group response rate and mean number of messages per day.

Data Collection

Software was developed to retrieve basic index information from the entire, available Yahoo message archive. In order to not flood Yahoo with requests, the program only retrieved about four pages a minute. This equated to about 120 index records a minute. A total of 210,120 index records were collected for the thirteen groups. Index information consisted of (1) the date time of the message (2) the subject line of the message (3) about the first 200 characters of the message (4) the message number within the group (5) a partial e-mail, (6) and profile identification. The last two fields were used to identify the poster.

Processing Algorithms and Limitations

The "identification" of the poster person was at times problematic in that a poster could use either the partial e-mail or profile identification or both.

Table 2: Group Totals and Averages

	Starting Date	Ending Date	Num of days of data	Number of messages	Number of threads	Messages / day	Threads / day	Ave Thread length	Reply Rate
B1	12/16/98	12/01/06	2,907	13,855	6350	4.77	2.18	2.18	50.1%
B2	03/05/99	01/21/07	2,879	14,959	5968	5.20	2.07	2.51	58.1%
B3	08/28/98	01/23/07	3,070	8,289	3849	2.70	1.25	2.15	48.2%
B4	03/14/99	01/20/07	2,870	27,539	10971	9.60	3.82	2.51	54.9%
J1	08/21/98	01/19/07	3,073	11,764	5970	3.83	1.94	1.97	27.6%
J2	02/18/02	05/08/06	1,540	3,175	1272	2.06	0.83	2.50	61.2%
J3	01/30/02	01/18/07	1,814	20,793	9201	11.46	5.07	2.26	52.0%
J4	11/06/00	01/19/07	2,265	18,681	4635	8.25	2.05	4.03	68.8%
W1	11/28/00	01/18/07	2,242	2,628	1090	1.17	0.49	2.41	56.9%
W2	01/22/02	01/17/07	1,822	9,548	3521	5.24	1.93	2.71	62.2%
W3	01/30/00	01/18/07	2,545	20,877	6206	8.20	2.44	3.36	66.1%
W4	11/20/00	01/15/07	2,247	11,111	2744	4.95	1.22	4.05	74.6%
E1	10/14/98	01/13/07	3,013	16,862	4873	5.60	1.62	3.46	66.4%

The profile identification option did not exist at the start of many groups and some users shifted to using profiles over the time period, which was often five years. In reviewing the data, we observed some cases where a user used two different e-mails. For the purpose of this study the main unit of study was the thread, hence the impact of the identification would be minimal. However future work would need to address this problem if the role of the person in the group was being studied.

The identification of threads was done by removing the "RE:" from a subject line and looking for a match of the remaining text in previous messages. This caused a small number of messages to be treated as separate threads when they were in reality part of the same thread. For example in cases where the e-mail was forwarded to another person and that person replied. The identification of 'terms' was done simply by looking for a space between characters. All the characters between the spaces were considered to be a term.

RESULTS AND DISCUSSION

As can be seen in Table 2 the response rate of the groups ranged from a low of 27.6% of the threads to a high of 74.6% with the mean being 57.5%. As this study used an automated process it included in the statistics message where it was clear a reply was not intended. For example occasionally a job posting might be listed. The group that had the lowest response rate (27.6%) was a group for Java Beginners and is clearly an outlier from the other groups' response rates. One explanation may be that this may be a function of a group where the majority of the group may not know the answers.

Table 2 also give us the average number of messages and threads on a per day basis for each group.

Figure 1: Response Rate by # of Threads per day

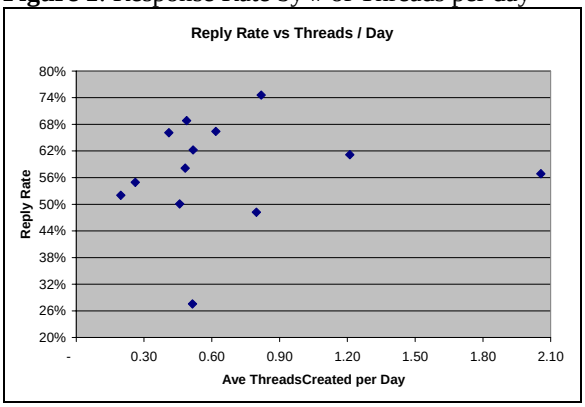

Figure 2: Response Rate by # of Messages per day

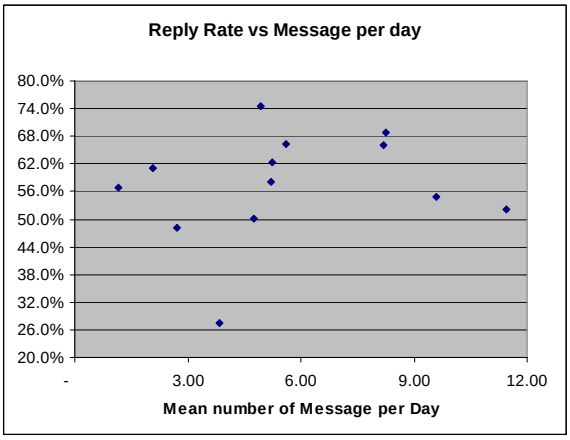

Table 3. Number of occurrences of specific terms and total terms by group and response

Group	Threads without replies					Threads with replies				
	Help	How	Please	Problem	Total # Terms	Help	How	Please	Problem	Total # Terms
B1	242	137	39	86	13580	282	125	39	70	11970
B2	177	138	25	69	10872	240	208	40	97	14487
B3	179	103	37	91	9460	201	93	37	76	7758
B4	435	290	75	148	22974	650	415	108	214	25400
J1	140	114	32	35	28178	169	76	35	67	6796
J2	8	8	3	11	1958	21	16	7	28	2964
J3	558	282	87	220	21403	690	381	114	227	21017
J4	40	70	12	41	7717	116	283	33	67	16323
W1	40	28	9	13	2081	52	48	12	14	2623
W2	99	73	11	52	6481	201	146	38	90	10141
W3	89	43	27	51	9525	155	127	40	133	17700
W4	10	13	6	31	3599	54	39	16	118	9866
Z1	18	19	12	21	7410	47	22	5	49	12788
Alll	2035	1318	375	869	145238	2878	1979	524	1250	159833

Since there were not enough groups to perform statistical tests on the impact of the rate of message or threads per day on the reply rate, simple scatter plots were looked at with no definitive pattern observed.

Four keywords were evaluated as to whether their presence made a difference in response rate. The four words evaluated were 'Help', 'How', 'Please' and 'Problem'. Table 3 shows for each group and terms the number of occurrences in the subject line of the messages that did and did not receive a reply. The keywords were compared using a log-likelihood statistic, as it is preferred over the chi-squared statistic when comparing frequencies of linguistic texts. (8) The method has been used to identify key words in two corpus as well as studying the social differential use of vocabulary (7).

Table 4: Significant words by group

Group	Help	How	Please	Problem
B1	p<.01			
B2				
B3	p<.01			
B4	p<.0001	p<.001		p<.05
J1	p<.0001	p<.0001	p<.0001	p<.0001
J2				
J3	p<.0001	p<.0001	p<.05	
J4		p<.0001		
W1				
W2	p<.05		p<.05	
W3		p<.01		p<.05
W4	p<.05			
Z1			p<.01	
Total	p<.0001	p<.0001	p<.001	p<.0001

The results that are statistically significant are given in Table 4. It is interesting to note that when all thirteen groups are combined all four terms were highly significant however this is not true for individual groups. In fact only one group had all four terms significant and two groups had none of the terms significant. This variety by group provides further evidence that individual groups have unique social norms.

We also evaluated the impact of the number of words in a subject line on response rate. These data along with the T-Test results are shown in Table 5. In all 13 groups the mean of the number of terms in the subject line was smaller for the group that received replies. Eleven of the thirteen groups were significant at .001. A 2-tailed test was used because there was no a priori collection hypothesis on the direction of the difference. In some ways these findings might at first appear to be at odds with [Joyce 06] who found that longer messages increase reply rate. However, this study explored answering technical questions, and we evaluated the subject line not the message. Indeed it may still hold that longer messages may get a greater response rate than shorter messages.

Table 5. Total number of terms used by group and response

	No Reply				Reply Made					
	Term count	**# Threads**	**Mean # terms**	**std dev**	**Term count**	**# Threads**	**Mean # terms**	**std dev**	**t-value**	**2-tail p**
B1	13,580	3170	4.284	2.572	11,970	3,180	3.764	2.235	8.600	0.0001
B2	10,872	2498	4.352	2.539	14,487	3,470	4.175	2.462	2.704	0.0069
B3	9,460	1993	4.747	2.650	7,758	1,856	4.180	2.499	6.817	0.0001
B4	22,974	4943	4.648	2.735	25,400	6,028	4.214	2.575	8.541	0.0001
J1	28,178	4324	6.517	3.398	6,796	1,646	4.129	2.590	25.801	0.0001
J2	1,958	494	3.964	2.348	2,964	778	3.810	2.160	1.198	0.2312
J3	21,403	4413	4.850	2.862	21,017	4,788	4.390	2.661	7.989	0.0001
J4	7,717	1444	5.344	2.809	16,323	3,191	5.115	2.720	2.628	0.0086
W1	2,081	470	4.428	2.561	2,623	620	4.231	2.649	1.233	0.2177
W2	6,481	1330	4.873	2.700	10,141	2,191	4.628	2.682	2.643	0.0083
W3	9,525	2103	4.529	2.562	17,700	4,103	4.314	2.358	3.300	0.0010
W4	3,599	697	5.164	2.709	9,866	2,047	4.820	2.567	3.013	0.0026
E1	7,410	1637	4.527	2.522	12,788	3,236	3.952	2.100	8.420	0.0001
ALL	145,238	29516	4.921	2.883	159,833	37,134	4.304	2.521	29.443	0.0001

CONCLUSIONS

In this paper we have shown that the content of the subject line can indeed have an impact on response rate to a request made to the technical groups that we studied. Specifically shorter subject lines are generally better. We have also shown that for some groups the presence of certain keywords can affect the response rate. The high variability of this impact of keywords provides more evidence that individual online communities develop their own social norms.

REFERENCES

1. Arguello, J., Butler, B., Joyce, E., Kraut, R., Ling, K. S., & Wang, X., Talk to me: Foundations for successful individual-group interactions in online communities. In CHI2006: Proceedings of the ACM Conference on Human Factors in Computer Systems. pp 959-968, New York: ACM Press
2. Boyd, D., Lee, H., Ramage, D., and Donath, J. 2002. Developing Legible Visualizations for Online Social Spaces. In Proceedings of the 35th Annual Hawaii international Conference on System Sciences (Hicss'02)-Volume 4 - Volume 4 (January 07 - 10, 2002). HICSS. IEEE Computer Society, Washington, DC, 115.
3. Brush, A. J. B., Wang, X., Turner, T. C., & Smith, M. A. (2005). Assessing differential usage of Usenet social accounting meta-data. In *CHI2005: Proceedings of the ACM Conference on Human Factors in Computer Systems*, pp. 889-898. New York: ACM Press.
4. Donath, J., Lee,H., Boyd, D., Goler,J., Loom2: Intuitively visualizing usenet - 2001 - CSCW (workshop)
5. Heer, J. and Boyd, D. "Vizster: Visualizing Online Social Networks," *Proc. Information Visualization Conf.*, pp. 33-40, 2005
6. Joyce, E., and Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, *11*(3), article 3. http://jcmc.indiana.edu/vol11/issue3/joyce.html
7. Rayson, P. and Garside, R**.**. Comparing corpora using frequency profiling. In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.
8. Rayson P., Berridge D. and Francis B**.** (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004, Presses universitaires de Louvain, pp. 926 - 936. ISBN 2-930344-50-4.
9. Smith, M. A. (2004). Netscan: A social accounting search engine (http://netscan.Resarch.Microsoft.com). Redmond, WA: Community Technologies Groups, Microsoft Corporation.